The Returns and Limits to Relative-Value ETF Arbitrage

Doering, Philipp*

This Draft: June 5, 2018

Abstract

This paper studies relative price gaps between pairs of nearly identical Exchange-Traded Funds (ETFs) listed on US exchanges. Prices usually move in lockstep, but sometimes diverge from parity by larger amounts. Over the period 2010-2016, a simple pairs trading strategy produced abnormal returns of up to 1.8 percent per year net of fees. Price gaps cannot be justified by fundamental differences in liquidity, replication methodologies, or security lending activities, but are related to both proxies for cross-sectional and time-varying limits to arbitrage. Prices typically diverge following a sequence of days with abnormally low liquidity in both the relatively over- and underpriced fund.

Keywords: law of one price, arbitrage, limits to arbitrage, market efficiency, Exchange-Traded-Funds, ETFs, pairs trading

* Ruhr-University Bochum, Department of Finance and Banking, Universitätsstrasse 150, 44801 Bochum, Germany; telephone: +49(0)234-32-21739, e-mail: philipp.doering@rub.de.

1 Introduction

With the ongoing shift from active to passive investing, Exchange-Traded Funds (ETFs) experience an increasing attention among both investors and academics. In 2015, assets under management surpassed hedge fund assets for the first time[1], and ETF shares are now accounting for about 30 percent of the overall US trading volume.[2] Besides providing low-cost access to diversification across all common asset classes, an often-cited key feature of ETFs is that they combine the benefits of closed- and open-ended funds. Like closed-end funds, ETF shares can be traded intraday. Like open-ended funds, additional ETF shares can be created (or existing shares redeemed) through the "creation/redemption mechanism". The combination of these two characteristics provides a natural arbitrage channel: once the ETF share price diverges from the underlying net asset value (NAV) by some extent, arbitrageurs buy the less expensive of both assets, convert it into the more expensive one and sell it, generating an (almost) immediate arbitrage profit (Ben-David, Franzoni and Moussawi 2017). This practice is also referred to as "primary market arbitrage", as it involves a change in the number of outstanding ETF shares.

Given this mechanism that by design intends to eliminate any mispricing within a short time, academics and practitioners long time paid little attention to the efficiency of ETF prices. However, recent evidence suggests that assuming ETFs to always trade at their NAV may be a quite expensive mistake. Angel, Broms, and Gastineau (2016) argue that NAV deviations can be much greater than the bid-ask spread and thus, ETF transaction costs are often higher than investors realize. In fact, the aggregate of these hidden transaction costs is remarkable: in the US, investors pay approximately $40 billion each year for trading at premiums or discounts (Petajisto, 2017).

While focusing on price-NAV deviations as a measure of mispricing is certainly the most intuitive way to test the law of one price among ETFs, another view is that funds tracking the same benchmark (and thus holding security baskets with similar payoffs) should sell for the same amount. In other words, if the market for ETF shares is truly efficient, then it should neither be possible to profitably arbitrage ETFs against their underlying basket, nor against each other. In a perfect market without any impediments to arbitrage, there must be a portfolio combination in which the price spread between two competing funds is always zero, as otherwise, risk-free arbitrage profits would be possible. This is the rationale underlying this paper. The high product homogeneity in the ETF market is somewhat unique and lends itself to view the law of one price in relative terms. For example, as of April 4, 2017, NYSE Arca alone lists 45 ETFs tracking US technology stocks. Some funds are virtually duplicates: iShares, RBS, State Street and Vanguard all offer their own funds tracking the S&P MidCap 400 index. As illustrated in Figure 1, about 20 percent of all US-listed ETFs track indices that are also covered by at least one

[1] *"Stock Market Milestone: ETFs One-Up Hedge Funds As Investor Assets Hit $3 Trillion"*, Forbes Online, May 8, 2015.
[2] *„ETFs are eating the US stock market"*, Financial Times Online, January 1, 2017.

more fund. In value terms, approximately half of the total US assets under management are invested in funds that have at least one competitor tracking the same benchmark.

Despite the ETF market providing a fertile ground for relative-value arbitrage, there is surprisingly little evidence yet. Marshall, Nguyen, and Visaltanachoti (2013) study the profitability and determinants of intraday mispricings between two large S&P 500 ETFs and report returns of up to 6 percent per year, net of bid-ask spreads. More recently, Petajisto (2017) and Fulkerson, Jordan, and Riley (2014) show that even when using daily data, funds from similar investment categories sometimes trade at quite different prices. Yet, comprehensive evidence, covering a large number of different ETFs and accounting for trading costs, is missing so far. More importantly, previous research provides little insights into the determinants of mispricing. The primary goal of this paper is thus to provide in-depth evidence on the profitability of relative-value ETF arbitrage, and to explore why similar ETFs can and occasionally do trade at different prices. I employ a pairs trading approach (e.g., Gatev, Goetzmann, and Rouwenhorst, 2006) to trade on price gaps large enough to be indicative of mispricing, and then use regressions to explain arbitrage profits both in the cross-section and time-series.

The overall contribution is twofold. First, it is a quite challenging observation that pairs of similar ETFs occasionally exhibit different prices. The literature has a long-standing interest in empirical asset pricing puzzles, where pairs of similar securities trade at different prices. For example, Schultz and Shive (2010) analyze price differentials between pairs of dual-class shares. Gagnon and Karolyi (2010) focus on price-parity among cross-listed shares. De Jong, Rosenthal, and Van Dijk (2009) study arbitrage returns in the context of dual-listed companies ("Siamese twins"). These papers have in common that though examining *close* substitutes, the paired-up securities are still exposed to some fundamental differences. Dual-class shares usually have different voting rights. Cross-listed shares in the US and their corresponding home-market shares as well as shares of dual-listed companies often trade in markets with different institutional features, such as disparately binding short-selling constraints, taxes, currency controls, or ownership limits (e.g., Gagnon and Karolyi, 2010; De Jong, Rosenthal, and Van Dijk, 2009; Froot and Dabora, 1999). The bottom line is that arbitrage profits can at least to some extent be interpreted as a premium for bearing fundamental risk, i.e. the risk that prices remain disconnected for an extended period of time. On the contrary, as also addressed in Marshall, Nguyen, and Visaltanachoti (2013) and discussed in more detail later in this article, fundamental risk is minimized among pairs of similar ETFs. At the same time, as I focus on pairs trading in the same market, cross-market differences in institutional features can neither play a role. Thus, I contribute to the more general empirical asset pricing literature by studying mispricings among assets that can be considered near-perfect substitutes, making ETF pairs an interesting new laboratory to evaluate the profitability and limits to relative-value arbitrage.

Second, I also contribute to the growing body of ETF literature. On the one hand, despite both practitioners and academics long time assumed primary market arbitrage to be the major price-correction

mechanism, evidence suggests that arbitrage activities between ETFs and their underlying baskets are rather scarce. Share creations and redemptions for a randomly selected ETF can only be observed on 6 to 13 percent of all trading days, and changes in ETF premiums or discounts are largely unrelated to prior share creations and redemptions (Fulkerson, Jordan, and Travis, 2017; Petajisto, 2017). Additionally, there is evidence that primary market activity in a given ETF declines after new competitors enter the market (Box, Davis, and Fuller, 2017). These findings suggest that a substantial part of price correction happens in the secondary market alone, potentially even more when there are competing funds. Yet, little is known about the mechanics of arbitrage in the secondary market for ETF shares. This paper provides empirical insights into secondary market ETF arbitrage by evaluating the profitability and limits to such a strategy. Though secondary market arbitrage may be implemented in a variety of ways, the pairs trading framework is both well-known among practitioners and been applied to test the law of one price in a variety of different contexts (e.g., Gatev, Goetzmann, and Rouwenhorst, 2006; Marshall, Nguyen, and Visaltanachoti, 2013).

On the other hand, as introduced by Petajisto (2017), considering an ETF's price distance to similar funds, rather than using the distance to its NAV as a measure of mispricing, prevents the results being biased by NAV staleness. To provide an intuitive example, consider the SPDR S&P Russia ETF (ticker RBL). On April 17, 2014, RBL traded at a remarkable premium of roughly 350 basis points on its NAV. Was RBL actually mispriced? As the last NAV was recorded on Russian market close at 3:45 pm and the ETF closing price at 9:00 pm (UTC), the NAV lagged the ETF share price by approximately 5 hours. Within this timeframe, the Russian government agreed on a pact to defuse the Ukraine Crisis. This agreement is priced in the ETF share, but not in the last available NAV. Thus, the observed premium most likely reflected an informational gap, and focusing on the premium alone would have falsely suggested a mispricing. On the contrary, RBL was not mispriced in relative terms: the share prices of competing funds were all up by nearly the same amount.[3] Hence, the ETF pricing efficiency tests conducted in this paper do not suffer from stale pricing biases (see Petajisto, 2017).

The major results can be summarized as follows. First, prices of similar ETFs typically move in close lockstep, but it is surprisingly common that the bid price of one ETF exceeds the ask price of a competing fund. Between 2010 and 2016, and across a total number of 4,118 selected pairs, there were 4,983 price gaps, with prices differing by 1.23 percent on average.

These price gaps must not necessarily be the result of mispricing, but can also arise for fundamental reasons. ETF pairs typically exhibit pronounced differences in liquidity and often employ different methodologies to replicate their benchmark. However, if prices deviate for fundamental reasons, one would expect gaps to be permanent and not to converge again. Yet, a simple pairs trading strategy historically generated excess returns in the order of 1.2 to 1.8 percent per year, net of bid-ask spreads and

[3] Two competing ETFs are the iShares MSCI Russia and VanEck Vectors Russia ETF, tickers ERUS and RSX.

common estimates for commissions, short-selling fees, and margin interest. Out of 4,983 identified price gaps, 71 percent and thus a vast majority converges, which should not be the case if prices disconnect for fundamental reasons. Besides, returns are positive and significant regardless whether the more liquid or more illiquid of both ETFs represents the overpriced leg, and are also significantly positive for positions in pairs that hold explicitly different security baskets to track their benchmark. These results provide additional evidence that at least on average, price gaps stem from mispricing rather than fundamental differences.

The persistence of price gaps can neither be explained by fundamental risk or short-selling constraints impeding arbitrage. However, my findings suggest that a number of other limits to arbitrage are in effect. First, arbitrage profits increase with the holding and transaction costs involved with positions in a particular ETF pair: returns are positively related to idiosyncratic risk and decrease with both secondary market liquidity and liquidity of the underlying assets. Besides, returns are higher in times of strong market-wide limits to arbitrage, most importantly when overall funding liquidity is low. Price gaps typically arise following a sequence of days with abnormally low liquidity in both the relatively over- and underpriced ETF. On the day of divergence, there is an abnormally high trading activity in both legs, which is indicative for an increase in liquidity risk. Both liquidity and trading activity return to normal levels over the days following the price gap. The is no convincing evidence that arbitrage using the creation/redemption mechanism is behind price correction. Altogether, the above findings indicate considerable inefficiencies in the pricing of ETF shares and suggest investors are well-advised to factor in the current pricing when considering to buy or sell a fund on a certain benchmark index.

The remainder of this article is structured as follows. The next section describes the ETF creation/redemption mechanism and provides a short review of the literature on ETF pricing. Section 3 discusses the sample and methodology employed. Results are presented in section 4 and section 5 concludes the paper.

2 A Brief Overview of ETF Pricing and the Creation/Redemption Mechanism

In a frictionless market, an asset always trades at its fundamental value, as the concept of arbitrage implies that mispricing is corrected immediately. In real markets, however, arbitrage is limited to the extent that (i) both cognitive biases and constraints may impede information diffusion (e.g., Barberis and Thaler, 2003) and (ii) transaction and holding costs make arbitrage costly (e.g., Pontiff, 2006). Transaction costs refer to bid-ask spreads, commissions, and market impact, while holding costs include the opportunity cost of capital, short-selling fees and idiosyncratic risk, with the latter often being considered as “the single largest cost faced by arbitrageurs” (Pontiff, 2006). The existence of holding costs implies that arbitrage in real markets is risky, as it makes the profitability of positions even in obviously mispriced assets contingent upon the time till convergence.

Among the core features of ETFs is that the combination of open-endedness and intraday tradability facilitates arbitrage activities to keep share prices in line with underlying NAVs. Specifically, primary market arbitrage is implemented as follows. Arbitrageurs monitor the price spread between the ETF share and the underlying security basket. Once the spread gets too large, the arbitrageur buys the less expensive of both assets and short sells the more expensive one. At market close (09:00 P.M. UTC), the arbitrageur submits a creation/redemption order to deliver the less expensive asset to the ETF sponsor in exchange for the more expensive one, allowing him to cover the short sale and realize an arbitrage profit at market close. Thus, the creation/redemption mechanism allows to exploit mispricing at minimum holding costs and by design aims to eliminate any observable price-NAV deviation in a short time. In a word, by creating or redeeming ETF shares in response to premiums, arbitrageurs can adjust the supply of ETF shares in a way that the fund trades close to the value of its underlying basket.[4] In order to engage in the primary market arbitrage mechanism, i.e. to trade directly with the ETF's capital market desk, it is necessary to become an "Authorized Participant" (AP) by entering into an agreement with the fund sponsor first. APs are typically institutional market participants (e.g. investment banks or brokers).

In recent years, academic literature concerned with ETFs and the aforementioned creation/redemption mechanism grew considerably. It can be broadly split into two different categories. First, there is controversy whether the increasing number of assets managed by ETFs may increase or reduce the efficiency of underlying security prices (e.g., Madhavan and Sobczyk, 2016 and Glosten, Nallareddy, and Zou, 2017). In particular, there are concerns whether the creation/redemption mechanism may serve as a shock propagator (e.g., Ben-David, Franzoni, and Moussawi, 2017).

The second strand of research is concerned with the pricing efficiency of ETFs themselves. Among the major findings is that ETFs tracking domestic equity benchmarks, if any, exhibit premiums or discounts that are typically small and short-lived (e.g., Engle and Sarkar, 2006). On the other hand, premiums or discounts (henceforth just "premiums" for brevity) are usually larger and often last for several days among ETFs tracking more exotic or more illiquid benchmarks, such as corporate and municipal bonds (Petajisto, 2017; Fulkerson, Jordan, and Riley, 2014). Besides, premiums also tend to be larger and more persistent for funds holding non-domestic underlyings (Delcoure and Zhong, 2007; Ackert and Tian, 2008; Levy and Lieberman, 2012, Angel, Broms, and Gastineau, 2016). This observation also holds when NAV staleness is accounted for (Petajisto, 2017).

[4] An intuitive explanation on the creation/redemption mechanism is provided by the Investment Company Institute (ICI), see https://www.ici.org/viewpoints/view_12_etfbasics_creation.

3 The Case for Secondary Market ETF Arbitrage

A distinctive feature of ETFs is that similar funds can be easily identified by comparing benchmark indices. Thus, the basic methodology to select ETF pairs underlying this paper is as follows. In the first step, I consider all possible pairs of funds tracking the same benchmark index. This should already result in careful selection of ETF pairs that share similar systematic risk exposures.

While identical benchmark indices are a necessary requirement for ETFs to be considered close substitutes, matching benchmarks alone are not a sufficient criterion. The prices of same-index ETF pairs can still disconnect for a variety of logical (i.e. fundamental) reasons that do not contradict the law of one price. A prominent example are ETFs tracking the same index, but employing different leverages. In principle, one could always replicate the return of an unlevered asset by holding part of the available capital in the leveraged equivalent of that asset and part in cash. To provide some intuition, consider (a) a leveraged position with the initial leverage set to two, and (b) an unleveraged position in the same asset. In this case, a short position in (b) could always be hedged by a long position that is half invested in the leveraged portfolio (a) and half invested in cash (and vice versa). However, when considering to hedge a position in a common ETF with an opposite position in a leveraged or inverse ETF (henceforth "LIN ETFs" for brevity), this logic does not hold. The reason is that LIN ETFs usually aim to deliver a multiple of the *daily* benchmark return and thus reset their leverage at the end of each trading day. The periodical leverage reset results in a volatility drag in the cumulative return of these funds (e.g., Charupat and Miu, 2011; Jiang and Peterburgsky, 2017). In consequence, prices of LIN ETFs and their unleveraged counterparts certainly diverge over time, but for reasons other than mispricing, making positions in pairs of ETFs employing different leverages rather a bet on volatility than on mispricing. Though tracking the same benchmark, they cannot be considered substitutes in price space, and there is no constant linear combination in which the fund values offset. Thus, in the second step, I remove all same-index pairs from the pre-selection, where the two funds maintain different leverage ratios or bet on opposite market directions. The final selection should result in a set of ETF pairs that are nearly identical.[5,6]

There are several implementation issues arising aside from the plain pair selection procedure. Specifically, by what extent do prices have to converge before a price gap is considered a potential "mispricing", i.e. a position is established? When did prices "converge" again, and how long are arbitrage positions maintained in case prices do not converge within expected time? To overcome these issues and

[5] There may still be fundamental differences among the paired-up funds, exposing arbitrageurs to the risk of betting non-converging, fundamental price differences. This will be discussed in more detail later.

[6] Leverage mismatches can also arise implicitly, when a fund pair handles dividends received or other cash income differently. For example, the SPDR S&P 500 ETF holds dividends in cash until it pays them out to investors, while the iShares S&P 500 ETF reinvests dividends, implying a small leverage, as the fund is investing money „borrowed" from investors. Thus, it slightly outperforms the SPY in case the S&P 500 rises in the meantime and vice versa.

most appropriately capture the profits that could potentially be earned by an average arbitrageur trading on relative ETF price gaps, I adopt the widely used pairs trading framework, which has been employed to study mispricing in a variety of different contexts (e.g., Gatev, Goetzmann, and Rouwenhorst, 2006; Schultz and Shive, 2010; Marshall, Nguyen, and Visaltanachoti, 2013). Specifically, in order to avoid a look-ahead bias and increase the power of my results, the algorithm is implemented rolling in two stages. The 12-month formation periods serve to select appropriate ETF pairs and to record the corresponding thresholds by which prices have to diverge in order to be viewed as an arbitrage opportunity. The selected pairs are then eligible for trading during the subsequent 6-month trading period.

3.1 Methodology

3.1.1 Formation Period

In each formation period, I first follow Petajisto (2017) and screen out the most illiquid ETFs, defined as having a daily average trading volume below $100,000 over the recent 12 months. Of the remaining funds, I pre-select pairs as outlined above. For each fund considered in a pair, I then compute daily cum-dividend bid and ask prices, normalized to start at $1 at the beginning of the formation period. Normalized prices are then used to compute daily spread series for each fund pair $p = \{y, x\}$:

$$spread_{yx,t} = \left(Bid_{y,t} - Ask_{x,t}\right)/Ask_{x,t}, \tag{1}$$

$$spread_{xy,t} = \left(Bid_{x,t} - Ask_{y,t}\right)/Ask_{y,t}, \tag{2}$$

where $Bid_{y,t}, Ask_{y,t}$ and $Bid_{x,t}, Ask_{x,t}$ are the cum-dividend bid and ask prices on day t of the two funds y and x forming pair p, respectively. The standard deviations $\hat{\sigma}_{yx}$ and $\hat{\sigma}_{xy}$ of these spread series are recorded and serve as a trigger for opening and closing positions in the subsequent trading period.

3.1.2 Trading Period

All pairs selected over the formation period are then eligible for trading over the following 6-month trading period. With the beginning of this period, prices are again set to unity and price spreads are monitored. I then open an equal-weighted long-short position in a pair, once the bid price of one leg exceeds the ask price of the other leg by some extent. In principle, one could use a constant threshold as a trigger for establishing arbitrage positions. However, to allow for a wider variation in arbitrage profits, I require the spread to exceed both a constant of 20 bps and two historical standard deviations as measured over the formation period. Specifically, I sell short ETF y and buy ETF x, when

$$\left(Bid_{y,\tau} - Ask_{x,\tau}\right)/Ask_{x,\tau} > \min\left(2\hat{\sigma}_{yx}, 0.0020\right). \tag{3}$$

If, on the other hand,

$$\left(Bid_{x,\tau} - Ask_{y,\tau}\right)/Ask_{y,\tau} > \min\left(2\hat{\sigma}_{xy}, 0.0020\right), \tag{4}$$

I sell short ETF x and buy ETF y, again with equal dollar amounts. Positions are closed out when (i) the spread returns to zero, (ii) a fund is delisted, or (iii) at the latest by the end of the current six-month trading period. Short sales are then covered at the current ask price and long positions sold at the current bid. Using bid and ask rather than closing prices prevents my results being biased by the bid-ask bounce (e.g, Gatev, Goetzmann, and Rouwenhorst, 2006; Jegadeesh and Titman, 1995; Jegadeesh, 1990). After a pair completed a whole roundtrip within the trading period, it is eligible for another trade and subjected to the same methodology again. Pairs earn zero interest if they do not actually trade, i.e. capital not allocated to a pair is not invested at the risk-free rate.

Two things should be noted here. First, a pair of similar ETFs is not necessarily mispriced when there is a price gap large enough for a position to be triggered, even though this is the basic premise underlying the strategy outlined above. Price deviations could be indicative of mispricing, but they can also exist for fundamental reasons, for example due to pronounced differences in liquidity among the paired-up ETFs (a topic that will be discussed further later in this article). For this reason, I follow the convention of Schultz and Shive (2010) and often refer to more general terms instead, such as "price gaps", "price deviations", or "price differences". However, if price gaps are typically for reasons other than mispricing, we would expect deviations to be permanent rather than to converge again. Thus, I follow Schultz and Shive (2010) and interpret converging price gaps as being indicative of mispricing.

Second, the choice of trading parameters is of course arbitrary. In order to avoid data snooping, I chose to follow simple instead of optimal trading rules. However, as will be shown in later parts of this article, the results reported are qualitatively very similar when varying the length of both the formation and trading cycles as well as the position triggers.

3.1.3 Return Calculation

In a frictionless market, arbitrage positions involving equal dollar amounts were self-financing, as the long leg could be financed by short sale proceeds. However, real markets require arbitrageurs to post collateral. Thus, I follow the vast literature concerned with arbitrage in other settings (e.g., Schultz and Shive, 2010; De Jong Rosenthal, and Van Dijk, 2009; Mitchell, Pulvino, and Stafford, 2002) and calculate the return of a pair based on the capital that is required to bring up the position. Specifically, I assume that arbitrageurs have to meet Regulation T (Reg T) margin requirements. According to Reg T, investors are required to bring up 50 percent of the long and 50 percent of the short market value as initial margin. Additionally, I assume a maintenance margin requirement of 30 percent for both the long and short position. For the sake of simplicity, positions are liquidated once the equity in a leg drops below the maintenance margin requirement.

The returns on a *pair* are then obtained by dividing the sum of payoffs from the long and short positions by the required equity. For the largest part of my analysis, I use net-of-fee returns, i.e. the payoffs considered in the nominator are net of bid-ask spreads, commissions, short rebates and any interest paid on margin borrowing. Specifically, positions are marked-to-market daily by dividing the daily net-of-fee payoffs by previous day's equity. As I assume trading at bid and ask prices, spreads are already accounted for. Commissions have been fairly low in recent years and are thus typically ignored in the related literature (e.g., Marshall, Nguyen, and Visaltanachoti, 2013). For example, investors with direct access to NYSE Arca are charged between 0.1 and 0.3 cents per traded share.[7] However, not all institutional investors are provided a direct access. Thus, I follow the more general estimates provided by the Investment Technology Group (ITG). According to ITG,[8] commissions average to approximately 5 bps in recent years. Thus, as a whole roundtrip involves four transactions, commissions amount to 20 bps in total. Finally, as short rebate data are difficult to obtain for ETFs, I rely on the estimates provided by Stratmann and Welborn (2013), who report an average rebate rate of -1.13 percent per year for US ETFs. In other words, ETF short sellers pay 1.13 percent per year to the lender on average, reflecting that ETFs are typically hard to borrow. As a proxy for the margin borrow interest on the long position, I use the Fed Open Rate on a daily basis plus 50bps (see, e.g., De Jong, Rosenthal, and Van Dijk, 2009).

I then use daily pair returns to compute daily returns for the *portfolio* of all ETF pairs. While these could serve as a rough approximation of the profits generated by an average arbitrageur, it does not aim to test whether abnormal returns can actually be earned by implementing the strategy. Rather, it aims to test whether price gaps among pairs of similar ETFs are indicative of mispricing. To compute portfolio returns, I assume that all pairs have the same weight at the beginning of the trading period. However, weights may change over time, since I assume that proceeds from previous trades within the same trading cycle are reinvested into the pair. I compute two different portfolio return measures: return on committed capital (ROCC) and return on employed capital (ROEC). ROCC adjusts the pair payoffs by the number of pairs that were selected for trading, while ROEC adjusts by the number of pairs that actually traded. To ease interpretation, I follow the convention of the pairs trading literature (e.g., Gatev, Goetzmann, and Rouwenhorst, 2006) and compound daily portfolio returns to monthly returns before reporting.

As in the momentum and pairs trading literature, the above trading cycle is implemented in a manner that a six-month trading period begins every month (except in the first 12 months of the sample, which solely serve as an initial formation period for the first trading period). The result is a set of monthly return series for overlapping six-month trading periods, giving rise to six different return observations

[7] NYSE Arca fees and charges can be found at https://www.nyse.com/publicdocs/nyse/markets/nyse-arca/NYSE_Arca_Marketplace_Fees.pdf.
[8] The preliminary version of the 2017 Global Cost Review is available at https://www.itg.com/assets/ITG_Global-Cost-Review-2017Q2-Prelim-BrokerCostUpdated.pdf.

for each month. This effectively mimicks a hedge fund of six managers whose implementation cycles are staggered by one month (Do and Faff, 2010). The actual monthly return on the pair portfolio is computed as the average across these six returns.

3.2 Data and Sample Characteristics

I combine Morningstar Direct and Thomson Reuters Datastream to construct my sample. First, I use Morningstar Direct to obtain a list of all dead and alive US ETFs ever traded between 2009 and 2016, implying that the first trading period begins in 2010. The choice of the sample is thought as a compromise between the time period covered and the number of possible fund pairs. The initial sample covered a total of 2,531 dead and alive ETFs (558 dead and 1,973 active funds as of December 31, 2016). I then screened out a number of funds to obtain my final sample. As I utilize the funds' benchmark for pair selection, I drop all funds that do not disclose benchmark indices in their prospectus from the sample. Specifically, I remove all funds which Morningstar classifies as being "actively managed" or "enhanced index funds". I do however retain funds grouped as "strategic beta", which weigh constituents according to their factor exposures rather than market capitalization, as these funds always track a benchmark index. Overall, these filters reduced the sample size from an initial 2,531 to 2,262 funds. As of December 31, 2016, and across all share classes, the funds in my sample managed approximately $2.5 trillion, accounting for close to 99% of the total assets under management across all US ETFs. For the remaining funds, I again used Morningstar Direct to obtain other qualitative fund characteristics, such as style categories[9] and replication methodologies, as well as daily shares outstanding, total net assets, and NAVs. For funds that Morningstar classified as "leveraged" or "inverse" (317 funds in total), I hand-collected the corresponding leverage ratios from the fund prospectuses.

Second, I downloaded daily consolidated bid and ask prices, dividends and trading volumes using Datastream. For a total of 143 funds, Datastream has no coverage, further reducing the sample from 2,262 to 2,119 ETFs.[10] In line with Petajisto (2017), I then define daily premiums as percentage difference of the daily mid-quotes from the corresponding NAV. To mitigate the effect of potentially erroneous quotes on my results, I apply a number of data filters closely following the related literature (e.g., Schulz and Shive, 2010; Marshall, Nguyen, and Visaltanachoti, 2013). For each fund, I discard all trading days where at least one of the following applies:

1. the bid quote, the ask quote, or both are missing,
2. the bid quote is equal to or greater than the ask quote,
3. the ask quote is exceeding the bid quote by more than 10 percent,

[9] Morningstar offers both Global and US categories. For my study, I use the US category classifications.
[10] Of the 143 removed funds, 53 were Exchange-Traded Notes. Besides, 95 of these 143 funds were active and 48 were dead as of December 31, 2016. Note that since my results are based on long-short returns, survivorship bias is not a concern.

4. the ask or bid quote is below $5, as many brokers prohibit shorting “penny stocks”.

Finally, I also remove all observations where premiums are larger than 20 percent in absolute terms, as these are likely to be erroneous (see Broman, 2016 and Petajisto, 2017). Trading is only allowed on the remaining days. Table 1 provides some sample characteristics.

[*Insert Table 1 here.*]

At the end of 2016, the median (average) ETF has $78 million ($1.4 billion) assets under management. However, there is a large disparity: while the smallest fund has only $0.2 million in assets, the largest fund (the SPDR S&P 500 ETF, ticker SPY) manages $225 billion. Thus, there are few market-leading ETFs with exceptionally high net assets. Likewise, we observe a right-skewed distribution for trading activity measures. For example, while the median daily trading volume is a mere $0.5 million, 5 percent of the funds in our sample have shares traded worth $63 million on an average day. At the same time, we observe bid-ask spreads and Amihud (2002) ratios being right-skewed as well. The average ETF trades at a bid-ask spread of only 1 bp, for instance. On the other hand, there are also funds trading at spreads in the order of several percent. In other words, most ETFs are quite liquid, but there are few funds with exceptionally low liquidity.

Both the mean and median premium are close to zero, indicating that the typical ETF trades quite close to its NAV on an average day. However, premiums vary substantially over time: the median ETF exhibits a premium volatility of 46 bps, implying that 95 percent of the time, the typical ETF fluctuates in a range of -90 to +90 bps around its NAV. Thus, ETFs occasionally trade at an economically quite significant premium. These above observations are both quantitatively and qualitatively very similar to those by Petajisto (2017).

4 Empirical Results

4.1 Descriptive Pair and Pair Portfolio Characteristics

Table 2 reports the pair frequency, size and liquidity characteristics of paired-up ETFs, as well as the pair portfolio composition. As can be seen from Panel A, the total number of selected ETF pairs amounts to around 4,100, pooled across all overlapping trading cycles, and the average number of pairs considered per trading period is close to 50.

[insert Table 2 here]

Panel B shows descriptive statistics for the funds considered in a pair. The key insights can be summarized as follows. First, the average ETF matched into a pair has about $7 billion in net assets and is thus from the highest size quartile, compared to the full sample of funds (see Table 1). Likewise, with an average spread of 11 bps and an average daily turnover of 2.8 percent, the typical pair constituent is

from the highest liquidity quartile. However, medians are in all cases substantially lower, suggesting a pronounced skewness in the size and liquidity characteristics of pair constituent funds.

A similar pattern can be observed when looking at the pair portfolio composition as reported in Panel C. The typical pair portfolio consists to about 85 percent of ETFs with above-median net assets and a below-median bid-ask spread. The weight of funds with a below-median Amihud (2002) illiquidity typically amounts to around 70 percent. Thus, on average, portfolios primarily consist of relatively large and liquid funds. However, at the same time, there are large within-pair differences with respect to size and liquidity. Approximately 80 percent of all pairs are from different size and liquidity deciles, and on average, the paired-up funds differ by 2 to 3 deciles. These observations reflect that for many benchmark indices, there is one market-leading ETF and a number of smaller and less liquid competing funds. Thus, despite the paired-up funds track the same index and thus share the same underlying liquidity, they typically exhibit pronounced differences in secondary market (i.e. fund-level) liquidity.

The pronounced differences in liquidity imply that potential price gaps between otherwise identical ETFs must not necessarily imply mispricing. Investors may be willing to pay higher prices for the more liquid fund. In this case, one would expect a clientele of long-term investors to buy the cheaper and less liquid of both funds, while short-term investors would be willing to buy the more expensive, more liquid ETF. However, unless the difference in liquidity varies over time, the price gap should be fairly stable. If price gaps are, on the other hand, indicative of mispricing, prices should converge over time, and arbitrage strategies should allow to produce abnormal returns. The extent to which liquidity differences can explain price gaps will be analyzed in more detail later.

The remainder of Panel C shows that 71 percent and thus the vast majority of all pairs selected for trading consist of funds tracking stock indices, whereas commodity and bond indices make up 17 and 12 percent, respectively. 40 percent of all pairs consist of funds tracking factor-weighted instead of traditional cap-weighted indices, reflecting the substantial growth in "smart beta" products in recent years. Besides, the paired-up funds are typically competing, i.e. ETFs issued by different sponsors, and the majority of pairs (61 percent) employs the same methodology to replicate their underlying benchmark index. On the other hand, the latter implies that 39 percent of all pairs certainly hold different security baskets to track their benchmark, providing another non-mispricing reason aside from differences in liquidity for why a pair of similar ETFs may trade at different prices. Though most ETFs track their benchmark well most of the time (e.g., Meinhard, Mueller, and Schoene, 2015), there is the risk that both funds sometimes differ in their tracking ability, implying a permanent price gap. Aside from that, some investors may simply be willing to pay higher prices for physically replicating ETFs, compared to synthetically replicating funds that involve some counterparty risk. In either case, we would expect the price gap to be permanent.

Potential price deviations between similar ETFs could also be fundamentally justified by differences in management fees or security lending activities. Participation in securities lending allows the ETF sponsor to partially offset the costs for running the fund, typically resulting in a smaller management fee. As these are subtracted from the fund value pro rata on a daily basis, diverging management fees will certainly lead to prices deviating over time, with funds charging smaller fees having higher prices in the long run. At the same time, investors may be only willing to pay lower prices for funds participating in securities lending due to the counterparty risk involved. Which of these effects dominates would ultimately be an empirical question. However, as reported in the last two rows of Table 2, the average pair portfolio is only to 11 percent made up of pairs where one ETF participates in securities lending, while the other fund does not. Likewise, per end of 2016, a mere 3 percent of the typical pair portfolio consists of pairs with considerable differences (20 bps or larger) in management fees.[11]

[insert Table 3 here]

Table 3 reports premium and price spread characteristics for paired-up funds. As can be seen from Panel A, the funds selected in a pair typically have close to zero premiums and a below-median premium volatility of 32 bps on average. Besides, premiums of paired-up funds typically exhibit a correlation of 0.4 and are significantly correlated in 71 percent of all cases. These results suggest a co-movement in premiums among similar funds as also documented by Broman (2016).

Panel B shows that the bid price of a paired-up fund is on average 24 bps below the ask price of the other fund in the pair, indicating that ETF pairs are not mispriced on average. However, the standard deviation averages to 53 bps across all pairs, implying that 95 percent of the time, this spread fluctuates in a range between -128 and +80 bps. Thus, for the typical ETF pair, it is not an extremely unlikely event that the bid price of one leg exceeds the ask price of the other leg. Again, due to the fundamental differences discussed above, such a price discrepancy must not stem from mispricing. However, if prices typically converge subsequently, we would interpret the initial price gaps as mispricing.

Indeed, Panel B also reports that close to 80 percent of all ETF pairs considered exhibit co-integrated price series. This implies that price differences are typically mean-reverting and provides some initial evidence that the observed deviations from price parity may indeed be the result of mispricing. Finally, 64 percent of all pairs remain co-integrated during the subsequent 6-month trading periods, suggesting that a long-short strategy as outlined in section 3 should generate positive returns, at least gross of fees.

[11] I consider 20 bps to be considerable, since the strategy requires prices to diverge by a minimum of 20 bps. A 20 bps difference in management fees per year implies that ceteris paribus, prices of an ETF pair should diverge by 0.05 basis points per day.

4.2 Do Price Gaps Indicate Mispricing? Evidence from Pairs Trading

If price gaps are indicative of mispricing, they should disappear over time. Furthermore, it should be possible to profitably trade on them, at least before trading costs and frictions. Table 4 summarizes the monthly portfolio returns for the arbitrage strategy outlined in section 3. Panel A shows that regardless of the portfolio return measure used and whether fees are accounted for or not, the strategy exhibits both statistically and economically significant positive returns. When adjusting with employed capital, net-of-fee returns average to 16 bps per month, which is an annualized 2 percent. Using the more conservative approach to compute portfolio returns, i.e. adjusting with committed capital, leads to an average net-of-fee return of 12 bps per month (annualized: 1.5 percent), which is still significant in both economic and statistical terms. Unsurprisingly, gross-of-fee portfolio returns are almost twice as large.[12] In all cases, returns remain significantly positive after subtracting the risk-free rate.

[insert Table 4 here]

The overall market return (gross of fees) averages to 110 bps per month, which is almost ten times as large as the most conservative return measure of the ETF arbitrage strategy, the net-of-fee return on committed capital. However, the strategy exhibits a substantially lower risk, regardless of the risk measure used. The realized market return volatility is 380 bps per month, whereas the strategy's return volatility amounts to a mere 25-40 bps. This disparity largely persists when adjusting with downside risk measures. Consequently, as reported in Panel B, the strategy outperformed the market, regardless of the performance measure used.

[insert Table 5 here]

Table 5 shows that only a small portion of the excess returns reported in Table 4 can be attributed to common systematic risk factors. As a market-neutral strategy, arbitrage profits are not significantly exposed to the market risk premium. Returns do neither load on any of the other factors considered. Thus, regardless of the return measure and factor model used, net-of-fee alphas are statistically and economically significant, ranging from 10 bps to 15 bps per month (annualized: between 1.2 and 1.8 percent).

The above findings have two important implications. First, price deviations among pairs of similar ETFs at least in part represent mispricing, implying noticeable inefficiencies in the pricing of ETF shares. If differences in liquidity, replication methodologies, or security lending activities make one fund more valuable than the other, we would expect the price gap to be permanent, which does not appear to be the case on average. Second, while we cannot ultimately rule out whether arbitrageurs can actually realize the documented returns, at least some investors could benefit from trading on these price gaps. An investor who considers gaining exposure to a certain index by buying an ETF and thus has to bear trading

[12] Note that since the strategy assumes buying and selling at ask and bid quotes, „gross-of-fee" means gross of margin interest, short selling fees and commissions, but still net of bid-ask spreads (see section 3.1.3).

costs anyway would be well advised to buy the fund that is underpriced relative to its competitors. Investors who are going to sell funds from their portfolio would benefit from selling those ETF shares that are overpriced relative to other ETF holdings.

[insert Table 6 here]

Table 6 reports the frequency and distribution of returns across all individual long-short positions. Pooled across all trading periods, there is a total number of 4,983 price discrepancies, and on average, there is a single price gap per ETF pair in a given six-month trading period. Of all identified price gaps, 3,556, or 71 percent, converge within the six-month trading periods and thus provide profitable trading opportunities. On average, converging trades produce a significant return of 65 bps net of bid-ask spreads and common estimates for commissions, margin interest and short selling costs. For half of these price gaps, it takes 3 days or less to converge. However, the average time till convergence amounts to a substantially larger 10 days, implying that there are a few price gaps that remain open for an extended time.

Taking into account the 1,427 (or 29 percent) non-converging price gaps, the average return across all trades reduces to a still significant 37 bps net of fees. Losses can occur for three reasons. First, if trades did not converge within the six-month trading period, positions are closed out on the last day of the period. Second, in case one of the paired-up ETFs is closed and delisted, existing positions involving that fund are liquidated on the delisting day, whether converged or not. Third, in case one leg of the position drops below the maintenance margin of 30 percent, positions are also closed out. This appears to happen in only 0.54 percent of all trades though. Across all three cases, the return of non-converging trades averages to a significant -32 bps.

When looking at the entire return distribution, it can be seen that large losses are quite seldom. In only 10 percent of all cases, returns are equal to or below -80 bps. Only 1 percent of all price gaps result in a loss of 418 bps or larger. Overall, the return distribution is left-skewed and leptokurtic, implying that there are many small and converging price gaps, coming along with few but possibly large losses. Altogether, the analysis of individual position returns indicates that the vast majority of price gaps between pairs of similar ETFs represent mispricing that provide an opportunity to produce excess returns, which should not be the case if prices diverge for fundamental reasons.

[insert Table 7 here]

Table 7 shows that the results discussed above are not the result of data snooping, but rather remain qualitatively very similar when varying the length of the formation or trading period, the standard deviation threshold serving as a trigger for initiating positions, or the required minimum price distance. Though the number and profitability of single positions varies when altering trading parameters, net-of-fee alphas remain significantly positive in all cases.

4.3 Why do Price Gaps Persist? Cross-Sectional Determinants of Arbitrage Profits

The remaining analysis will focus on exploring the reasons why pairs of seemingly identical ETFs occasionally trade at different prices. Aside from fundamental differences, which at least on average seem not to explain price gaps, behavioral finance provides some rationales why price deviations persist in real markets. In fully efficient and integrated capital markets, arbitrageurs would immediately bet against any price discrepancy, correcting prices within a matter of seconds. The behavioral finance literature states that in real markets, due to the existence of trading costs and other frictions that limit the implementation of arbitrage, price anomalies can persist for an extended time, even in the presence of fully rational market participants (e.g., Barberis and Thaler, 2003). A testable implication of this framework is that arbitrage returns should ceteris paribus be larger in settings where impediments to arbitrage are particularly strong.

Thus, as fundamental differences appear to play (if any) a little role in explaining why price deviations persist, the major part of the remaining analysis focusses on assessing the extent to which price gaps relate to a battery of different limits to arbitrage proxies. These investigations serve two purposes. First, they provide some cross-check on whether the price gaps identified are actually the result of mispricing and not an artifact of remaining data errors. Second, they allow to dig deeper into the ETF price discovery process, which might help academics and practitioners likewise. For the latter, the results might provide helpful insights for practitioners to optimize the way in which they buy or sell ETFs.

The following sections will not be based on portfolio returns as computed in section 4.2. The reason is that portfolios can be thin in some trading cycles, and in results unreported for brevity I found that portfolio size varies strongly over time. Thus, following the pairs trading literature (Jacobs and Weber, 2015; Engelberg, Gao, and Jagannathan, 2009), I use 20-day event-time returns instead, which are simply the cumulative returns of single trades over the 20 days following the price gap, net of fees and the risk-free rate.

4.3.1 Short Selling Constraints

While the baseline analysis already assumed a short rebate rate of -1.13 percent per year, considering that ETFs are typically hard to borrow (see Stratmann and Welborn, 2013), there is anecdotal evidence that short selling may be more difficult or sometimes even impossible for some types of funds. In particular, Avellanada and Dobi (2013) report that the costs for borrowing leveraged and inverse ETFs are rather ranging from 250 to 850 bps. Besides, in 2009, the Financial Industry Regulatory Authority

(FINRA) implemented stricter margin requirements for leveraged and inverse ETFs.[13] Some brokers even entirely exclude a subset of these funds from margin trading.[14]

[insert Table 8 here]

Thus, the reported abnormal returns could simply be an artifact of short-selling restrictions by assuming long-short positions in funds that are difficult to borrow. If short-sale constraints can explain why relative ETF price gaps persist, deviations should be driven by an overpricing in hard-to-borrow ETFs. Thus, compared to non-leveraged ETF pairs tracking the same benchmark, arbitrage profits among pairs of leveraged or inverse ETFs (LIN ETFs) should rather stem from the short leg than the long leg of the position. Panel A of Table 8 provides support for this notion. While positions in both LIN ETF pairs and pairs of non-leveraged funds tracking identical benchmarks achieve statistically similar returns in total, they exhibit quite different return profiles. For pairs of LIN ETFs, returns are driven by the short leg, while the average return of the long leg is insignificant. The opposite is true for positions in similar non-leveraged fund pairs. Thus, for pairs of LIN ETFs, the data does not allow to reject the hypothesis that short-sale constraints contribute to the persistence of price gaps. However, positions in LIN ETF pairs only account for 2 percent of all trades. It is thus not surprising that average returns do not significantly change when excluding LIN ETF pairs from the sample, as presented in Panel B. For the remaining pairs, returns predominantly stem from the long leg.

4.3.2 Within-Pair Differences in Liquidity

As discussed in section 4.1, the paired-up funds are typically from quite different liquidity deciles. Liquidity differences may provide a plausible explanation for two otherwise identical ETFs to trade at different prices. Given that some investors are willing to pay a premium for more liquid or command a discount for more illiquid (but otherwise equal) funds, we would expect prices not to converge, but rather to hold permanently.[15] Thus, if liquidity differences can explain why two otherwise identical ETFs exhibit different prices, we should not observe positive (or at least smaller) returns for price gaps where the more liquid ETF trades at a higher and the less liquid fund at a lower price. The results presented in Table 9 show that this does not appear to be the case though.

[insert Table 9 here]

Table 9 splits all trades into two different groups, depending on whether the more liquid or more illiquid of both ETFs is held long, and reports returns separately for each category. The key result is that regardless of whether the more liquid or more illiquid of both funds is held long, returns are on average positive

[13] See Regulatory Notice 09-53, effective as of December 1, 2009. As a consequence, margin requirements for leveraged and inverse funds increased by a factor commensurate with their leverage.
[14] For example, Merrill Edge prohibits trading ETFs that are leveraged three times or larger on margin.
[15] For the moment, I just assume that there is no considerable change in the level of liquidity around the day of divergence. The extent to which liquidity changes around the event-day will be adressed later in this article.

and significant. Besides, when comparing the magnitude of returns, positions that short the more liquid and buy the more illiquid fund even exhibit larger arbitrage profits, at least when liquidity is measured using bid-ask spreads or Amihud (2002) illiquidity. This is the opposite from what one would expect if liquidity differences could explain why price gaps exist. Thus, the major implication from Table 9 is that within-pair differences in the level of liquidity cannot explain price gaps.

Aside from that, the results reported in Table 9 provide further evidence that short-selling constraints are unlikely to contribute to the persistence of price gaps. In case the more liquid fund is held long (i.e. the more illiquid ETF held short), returns from the short leg are significantly negative. If, on the other hand, the more illiquid ETF is held long, short position returns are insignificant. Thus, for both types of positions, arbitrage profits are driven by the long leg.

4.3.3 Differences in Replication Methodologies and Fundamental Risk

Textbook arbitrage assumes that both legs of the arbitrage position are substitutable, i.e. that for every mispriced asset, there is a perfect hedge available. In real capital markets, however, perfect substitutes are difficult to find, exposing arbitrageurs to the risk that prices deviate for fundamental reasons and thus do not converge again. This kind of arbitrage risk, also referred to as *fundamental risk*, has been argued to impede arbitrageurs from correcting mispricing in a number of different settings (e.g., Mitchell, Pulvino, and Stafford, 2002).

The results in section 4.2 already suggested that fundamental risk should be quite limited among ETF pairs, as more than 70 percent of price gaps converge within a few days after divergence, and losses are typically small for the remaining discrepancies that do not converge. Table 10 provides further evidence by reporting both price and NAV distances for the day of divergence and the day positions are closed out. The NAV reflects the fundamental value of a fund. Thus, if prices of similar ETFs typically deviate for fundamental reasons, we should observe that price differences come along with a similar distance in NAVs.

[insert Table 10 here]

This does not appear to be the case though. On the day of divergence, the bid price of the short leg diverges from the ask price of the long leg by 120 bps on average. The median is a considerably smaller 74 bps, pointing towards some outliers. Subsequently, this price gap drops to an insignificant average of 10 bps on the day positions are closed out. The average change in price distance across all individual positions amounts to a significant -110 bps, while the median is a substantially smaller -66 bps. At the same time, the fundamental difference as measured by percentage NAV distances averages to only 58 bps (median: 30 bps) at position opening and decreases to 40 bps (median: 2 bps) over time. Though statistically significant, the percentage distance in NAVs at position opening is not even half as large as the share price distance. Besides, the average change in NAV distances from position opening to closing

is statistically not different from zero. Thus, price gaps are largely driven by the non-synchronicity of premiums and discounts among nearly identical ETFs: the short leg typically trades at a significantly larger premium than the long leg on the day prices diverge, and this disparity inverts until positions are closed. This is in line with the findings for intraday arbitrage between two S&P 500 ETFs as reported in Marshall, Nguyen, and Visaltanachoti (2013) and again suggests that price gaps are mostly non-fundamental, i.e. the result of mispricing.

Nevertheless, the observation that NAVs also gap by a significant (though not the same) amount on the day positions are initiated points towards some fundamental risk. If all fund pairs considered would be fundamentally identical, they should rather have the same NAV across the entire state space, which is not supported by the data. Given that around 40 percent of all pairs in the sample (see sction 4.1) employ different replication methodologies, implying that the paired-up funds certainly hold different underlying security baskets, it should be a particular concern that prices diverge due to different tracking errors. Most ETFs track their benchmark well most of the time, regardless whether they fully replicate their benchmark, use a sampling approach or do so synthetically by entering total return swap agreements with investment banks (Meinhard, Mueller, and Schoene, 2015). Yet, tracking abilities can differ occasionally, especially in times of market stress such as the May 2010 Flash Crash (e.g., Marshall, Nguyen, and Visaltanachoti, 2013), most probably resulting in permanent price gaps.

If arbitrageurs are concerned about that kind of risk, we should observe higher returns especially among pairs employing explicitly different replication methods. At the same time, we should observe a higher frequency and magnitude of loss-making, non-converging trades among these kind of pairs, compared to pairs with matching replication methodologies. In that case, the documented excess returns could be partially construed as a premium for bearing fundamental risk. Thus, I split the sample of all price gaps into two groups: those occurring among pairs with matching and non-matching replication methodologies. The results are reported in Table 11 and provide an at best mixed evidence.

[insert Table 11 here]

Though there is evidence that pairs employing different replication methodologies suffer greater losses on non-converging positions than same-replication pairs, the return difference is statistically not different from zero. Mixed-replication pairs also exhibit a 1-percentage-point lower convergence probability, but this difference is again not significant. There is solely evidence for a significantly larger standard deviation in losses from non-converging trades among different-replication pairs, implying a wider dispersion of losses compared to same-replication pairs. As a whole, I infer from these results that even for pairs certainly holding different security baskets to track their underlying benchmark index, fundamental risk cannot be a major concern discouraging arbitrageurs from eliminating price gaps.

4.3.4 Why Are Some Pairs More Profitable Than Others? The Role of Cross-Sectional Differences in Arbitrage Costs

The previous sections suggested that neither short-selling constraints, nor fundamental risk seem to be major frictions impeding arbitrageurs from correcting mispricing among pairs of similar ETFs. On the other hand, there is a large cross-sectional variation in the profitability of price gaps (see Table 6), which naturally raises the question of why some pairs are more profitable than others. Aside from short-selling constraints or fundamental risk, the literature points towards transaction and holding costs contributing to the persistence of mispricing in real markets (e.g., Pontiff, 2006; Barberis and Thaler, 2003). If arbitrage costs impede arbitrageurs from correcting mispricing immediately, we should see larger arbitrage returns among ETF pairs with higher transaction and holding costs. The baseline analysis already uses net-of-fee returns. However, though representative, the underlying fee estimates do not consider fund-specific characteristics, but rather assume the same level of trading costs across all ETF pairs considered. Thus, in the following I analyze the extent to which arbitrage profits relate to a number of different transaction and holding cost proxies inspired by the literature.

First, while fundamental risk is minimized, trading on relative ETF price deviations does still involve convergence risk. Even if the selected ETFs are fundamentally the same and price gaps certainly converge, it is ex ante unclear how long it will take (*synchronization risk*, see Abreu and Brunnermeier, 2002). In the meantime, noise traders may cause prices to diverge even further, potentially forcing arbitrageurs to provide additional equity to their margin account or unwind the position (*noise-trader risk*, see De Long et al., 1990). While these convergence risks play a negligible role in primary market arbitrage (see section 2), they are at least theoretically a concern when attempting to arbitrage ETFs against each other. Indeed, as discussed in section 4.2, about 30 percent of all positions do not converge within the given time frame, and there is some dispersion in the time till convergence. When arbitrage positions are subjected to convergence risk, arbitrageurs have to bear periodical costs for holding the position until convergence. Thus, price gaps between ETF pairs should increase with the holding costs involved with a potential arbitrage position. The most important and most frequently used holding cost proxy in the literature is idiosyncratic volatility (e.g., Pontiff, 2006), reflecting the risk of an arbitrage position that is unrelated to systematic factors and any other available hedge portfolio.

Despite the unambiguous definition of idiosyncratic volatility, it is unclear how to precisely measure this risk in the context of arbitrage positions (see, e.g., Gagnon and Karolyi, 2010; Jacobs and Weber, 2015). The approach used to proxy for idiosyncratic risk in this paper is as follows. For both funds in a pair, I first compute daily mid-price returns net of the underlying NAV returns. I then regress the pairwise difference of the resulting series on the three Fama and French (1992) factors. The residual vola-

tility from these regressions is then used to proxy for the idiosyncratic risk involved with arbitrage positions in an ETF pair. Factor regressions are performed separately for each of the 12-month formation periods.[16]

Second, if transaction costs contribute to the persistence of mispricing, returns should be related to liquidity. Compared to other settings in which similar assets trade at different prices (e.g., dual-class shares), ETFs are somewhat special in that they have a two-tier liquidity structure. First, just like stocks, ETF shares can be traded in the secondary market, and the more actively the funds forming a pair are traded in the secondary market, the cheaper their shares can be purchased or sold throughout the trading day. Thus, arbitrage profits should be larger among ETFs pairs with more inactive secondary markets. I use a number of different variables to proxy for secondary market liquidity following the literature (e.g., Jacobs and Weber, 2015; Broman and Shum, 2018). I compute (1) the natural logarithm of pair-average total net assets, recorded on the last day of the preceding formation period, (2) the time-series median of daily bid-ask spreads, serving as a proxy for transaction costs involved with normal-sized quantities, (3) Amihud (2002) illiquidity ratios to measure the price impact costs associated with larger transactions, as well as two proxies to measure overall trading activity, namely (4) the median of daily turnover ratios and (5) the natural logarithm of average daily trading volumes.

A distinctive feature of ETFs is that even if "on-screen" liquidity is zero, fund shares may be traded through the creation/redemption mechanism (see section 2). The traditional liquidity proxies discussed above do not necessarily capture primary market transactions. For example, if newly created shares are not loaded off in the secondary market, share creations lead to a decrease in turnover ratios, as the denominator (shares outstanding) grows, but the nominator is unaffected (see Broman and Shum, 2018). There is vast evidence that premiums and discounts (i.e. absolute mispricing) are related to underlying liquidity (e.g., Petajisto, 2017; Ackert and Tian, 2008; Engle and Sarkar, 2006). The reason is that the costs for arbitraging an ETF against its benchmark securities increase with underlying illiquidity, implying higher transaction costs for buying or selling the underlying security basket.

At least theoretically, price differentials between pairs of similar ETFs (i.e. relative mispricing) should also be related to underlying liquidity: pairs holding relatively illiquid securities should be more prone to trade at different prices than pairs with relatively liquid underlyings. To provide some intuition why this should be the case, consider the theoretical example of an ETF pair with zero underlying liquidity. In this case, it would be impossible to arbitrage the ETF against its (non-tradable) underlying through

[16] The median idiosyncratic volatility for all pairs, pooled across all trading cycles, amounts to a mere 25 bps per year, which is quite low compared to other settings in which similar assets trade at different prices (e.g., Jacobs and Weber report an average of 112 bps for pairs of statistically similar stocks). I have repeated the same analysis using other approaches to estimate idiosyncratic risk instead, including running Carhart (1997) four-factor or Fama and French (2015) five-factor regressions, or using equal-weighted pair-averages of residual volatilities from individiual factor regressions following Jacobs and Weber (2015). The results remain qualitatively and quantitatively very similar and are available from the author upon request.

the share creation/redemption mechanism. Price correction would entirely depend on secondary market trading. However, as opposed to primary market arbitrage, secondary market arbitrage involves convergence risk. Thus, in order to compensate arbitrageurs for bearing additional risk, relative price gaps should be larger among ETF pairs where primary market arbitrage is impeded. The above is of course an extreme example, as non-tradable benchmark indices contradict the fundamental idea of ETFs. Nevertheless, the basic logic should also apply to more general cases, and we should observe larger arbitrage profits among pairs with less liquid underlyings.

Unfortunately, underlying liquidity data is not readily available in common databases, especially for more exotic benchmark indices. Thus, I follow Broman (2016) and proxy primary market liquidity by the share creation/redemption activity, computed as

$$PrimActivity_{i,T} = \log\left(1 + \frac{1}{T}\sum_{t=1}^{T}\frac{|SHR_{i,t} - SHR_{i,t-1}|}{SHR_{i,t-1}}\right), \quad (5)$$

where $SHR_{i,t}$ is the number of shares of ETF i outstanding on day t. The underlying notion is that if arbitrage using the share/creation redemption mechanism is impeded, this will likely result in infrequent and small changes in the number of shares outstanding, compared to funds where primary market arbitrage is less limited.[17] All of the aforementioned liquidity variables are first measured individually for the paired-up funds over the preceding 12-month formation periods, and the individual measures are then used to compute equally-weighted pair-level averages. By using several different liquidity proxies simultaneously, I aim to capture all different facets of ETF liquidity.

[insert Table 12 here]

Table 12 reports how event-time returns relate to the arbitrage cost proxies outlined above. Panel A presents evidence for univariate analysis, where I determine the top and bottom quartile with respect to each variable and report returns separately for both groups. The results reveal that arbitrage profits are related to all limits to arbitrage proxies considered, except for fund size and dollar trading volume. Returns are significantly larger for pairs carrying relatively high idiosyncratic risk and exhibiting relatively low liquidity when measured by bid-ask spreads, Amihud (2002) illiquidity, or turnover. Besides, arbitrage profits are significantly larger among pairs with less active primary markets, indicating that price gaps are indeed more pronounced among ETF pairs tracking less liquid underlyings. In results unreported for brevity, I found that low-activity pairs often track indices that may at least occasionally be difficult to trade, including non-domestic equities, small caps with factor tilts, aggregate bond indices,

[17] It should be noted that since the share/creation mechanism is often considered the major price-correcting mechanism among ETFs, primary market activity in a given ETF could either be low because the fund is always priced efficiently, or because share creations/redemptions are impeded. A better measure would be the actual bid-ask spread or Amihud (2002) ratio of the underlying basket. Unfortunately, underlying liquidity data is difficult to obtain from common databases.

or physically-backed precious metal funds. In the latter case, share creations literally involve a physical delivery of bullions into the vault of the fund's custodian, which especially for large transactions may be more difficult than accumulating and settling exchange-traded underlyings.[18] With respect to total net assets and trading volume, the differences in average returns have the predicted signs, but are statistically insignificant.

Panel B reports multivariate evidence, where event-time returns are regressed on multiple arbitrage cost proxies and a number of control variables simultaneously. The results suggest that the findings from Panel A largely hold in a multivariate setting. In all specifications considered, arbitrage profits still increase with holding costs as proxied by idiosyncratic risk and Amihud (2002) illiquidity. Besides, returns remain negatively related to turnover and primary market activity. Evidence remains rather mixed with respect to fund size and trading volume. The coefficient of bid-ask spread has the expected sign, but is not consistently significant across all specifications. Thus, transaction costs for normal-sized transactions are, if any, a weak impediment to arbitrage. Nevertheless, the overall picture presented in Table 12 indicates that cross-sectional limits to arbitrage are often binding. Both holding and transaction costs appear to deter market participants from immediately correcting price gaps between nearly identical ETFs.

4.4 Conditions Prevailing on the Day of Divergence and Time-Varying Limits to Arbitrage

The analysis so far focused on cross-sectional differences in the level of arbitrage frictions and costs to explain why price gaps among ETF pairs persist. This section instead focuses on the conditions prevailing on the day of price divergence, which is among other things motivated by Figure 2. Following Jacobs and Weber (2015), Figure 2 plots the distribution of 20-day event-time returns for all days where at least one pair diverges, averaged across all positions initiated on that day. It can be seen that returns are anything else than uniformly distributed across event days, pointing towards time-varying factors influencing arbitrage profitability.

4.4.1 Liquidity and Trading Activity Around the Event-Day

It is widely recognized that the *level* of liquidity, which is inherently tied to the transaction costs involved with an arbitrage trade, may contribute to persistence of mispricing. Aside from that, asset prices may also be affected by liquidity shocks, i.e. sudden *changes* in the level of liquidity. Most importantly, according to the model by Campbell, Grossman and Wang (1993), noise trader's non-informational liquidity demand may lead to temporary price pressure, conditional on the level of liquidity. A number of empirical studies support the view that deviations from price parity among similar assets coincide

[18] A closer look at the data also revealed that there is no substantial difference in the frequency of primary market transactions between precious metal and other ETFs, but quantities (measured in percent of the shares outstanding) are typically smaller for these funds. Besides, significantly higher average returns for low-activity funds can still be observed when only focusing on equity ETFs. The results are available upon request.

with abnormal liquidity. For example, Engelberg, Gao, and Jagannathan (2009) argue that the profits to trading pairs of similar stocks can to some part be construed as a reward for providing immediate liquidity. Schultz and Shive (2010) find evidence that mispricing between dual-class shares tends to occur in times of abnormal volume in both share classes. In particular, Marshall, Nguyen, and Visaltanachoti (2013) report that there is both a negative liquidity shock and an increase in trading volume in the minutes surrounding intraday arbitrage opportunities between two S&P 500 ETFs.

To explore the microstructure conditions prevailing when price gaps arise, I compute *abnormal* bid-ask spreads, trading volumes, turnover ratios, and primary market activity for the day of divergence and the six days surrounding. Specifically, for all seven days and all ETFs separately, I compute the logarithm of daily (i) bid-ask spreads, (ii) trading volumes, (iii) turnover, and (iv) the daily absolute change in the number of shares outstanding. For each fund separately, these measures are then normalized by their respective time-series averages over the 180 trading days surrounding the day where the price gap emerges. Table 13 presents the results, averaged across all identified price gaps and reported separately for the long and short leg of the position.

[insert Table 13 here]

It can be seen that price gaps typically arise following a sequence of days with abnormal low liquidity. In both the relatively over- and underpriced leg of the position, we observe logarithmized bid-ask spreads being up to 4.8 percent higher than normal over the days preceding price divergence. Then, consistent with the notion that arbitrageurs bet on the correction of mispricing, liquidity is restored to levels that are not significantly different from normal. With respect to trading volume and turnover, we do not observe either abnormally low or high levels preceding the day of divergence (except for the short leg on the previous day). At first glance, this finding may be contradicting the observation of abnormally high spreads reported for the same days. However, it can be can reconciled with the results by Johnson (2008), who shows that volume and liquidity are largely unrelated. His model implies that liquidity reflects the average risk-bearing capacity or willingness of the market to accommodate trades at the prevailing prices, while volume reflects the changing contribution (i.e. second moment) of individuals to that average. Thus, volume is not related to liquidity, but rather to the variance of liquidity, or liquidity risk. Hence, as there is abnormally high trading volume and turnover in both position legs on the day of divergence, price gaps coincide with an increase in liquidity risk in both the relatively underpriced and overpriced ETF. Liquidity risk is then restored to normal levels over the subsequent days. These findings are very similar to those reported by Marshall, Nguyen, and Visaltanachoti (2013) for intraday ETF mispricing.

The above observations once again suggest that the day of divergence is not a random day, but they also have another important implication. First, the analysis so far assumed the initial price gap as being indicative for mispricing, given that it subsequently converges. However, another possible consideration

is that there is in fact a fundamental and sustainable change in liquidity on the event day. In that case, we would view the subsequent convergence rather than the initial divergence as a mispricing. It should be noted here that the selected time frame of price divergence ± 3 days in Table 13 is not chosen arbitrarily, but rather thought of as a representative choice reconciled with the median time till convergence of 3 days after price gaps arise (see Table 6).[19] In this light, the findings discussed above hardly provide any convincing evidence for the aforementioned rationale. If prices disconnect for fundamental shifts in the level of liquidity, we should see a permanent change in liquidity in one of the two ETFs, but we rather observe temporary drops in the liquidity of *both* legs preceding the day of divergence. Similarly, we observe a temporary abnormal volume in both legs of the position on the day of divergence. On the third day after divergence, we observe the liquidity of *both* legs returning to normal levels. These results provide further evidence that at least on average, the initial price divergence (and not the subsequent convergence) is indicative of mispricing.

The remainder of Table 13 provides insights into the share creation/redemption activity around the event day. The creation/redemption mechanism is typically assumed to be the major price-correcting mechanism to keep prices of ETF shares in lockstep with their fundamental value. If it is primary rather than secondary market arbitrage that causes relative price gaps between similar ETFs to converge, we should observe abnormal creation activity in the overpriced (i.e. short) leg, abnormal redemptions in the underpriced (i.e. long) leg, or both. The results presented in Table 13 provide hardly any evidence for this rationale. First, in absolute terms, we indeed observe an abnormal change in the number of shares outstanding on the event day.[20] Thus, there is an evidently overall abnormal trading activity in the primary market for both funds.

The last two rows focus on the magnitude of share creations (redemptions), conditional on that there is a positive (negative) change in shares outstanding, i.e. given that there is a net creation (redemption) in the fund on that day. Now, we do not observe evidence for the notion that primary market arbitrage alone causes prices to correct anymore. First, for those ETFs that exhibit share creations, we indeed observe an abnormal magnitude of creations on the day of divergence (and the subsequent days), but only among funds forming the long leg. For funds that exhibit creations and are held short, the magnitude is statistically not different from their normal creation size. This is opposite from what one would expect if arbitrageurs trade against the overpriced ETF through the primary market.

For those ETFs that experience share redemptions, there is evidence for abnormal redemption sizes among *both* funds forming the long and the short leg. Again, if primary market arbitrage were the driver

[19] The overall picture remains qualitatively very similar when taking more days into account.

[20] It should be noted that share creation and redemption orders are settled two days after the trade date (T+2). However, in a personal consultation, Morningstar (from which we sourced the shares outstanding) confimed that in their data, changes in the shares outstanding of ETFs are accounted for on the day of transaction and not the settlement date. Thus, the day where prices diverge is the central day of interest in this context.

for subsequent price correction, we should observe abnormal redemption activity only in the underpriced long leg. Overall, these findings indicate that for that kind of ETF mispricing, i.e. when similar ETFs exhibit different prices, there is no convincing evidence that primary market arbitrage is behind price convergence, suggesting that prices are mostly corrected through trading in the secondary market.

4.4.2 Time-Varying Limits to Arbitrage

There is broad evidence that a variety of price anomalies are related to the market-wide availability of arbitrage capital. For example, Asness, Moskowitz, and Pedersen (2013) find that value and momentum returns are strongly linked to overall funding liquidity constraints. Jacobs (2015) shows that widely discussed violations of the law of one price (i.e. price gaps among dual-class shares, cross-listed stocks and "Siamese twins") are strongly related to both funding and market illiquidity and tend to increase with overall market volatility.

Motivated by these findings, I henceforth analyze the sensitivity of ETF pairs trading profits with respect to a number of commonly used proxies that aim to capture time-variation in market-wide limits to arbitrage. To proxy funding liquidity, I use (i) the TED spread (e.g., Brunnermeier and Pedersen, 2009), defined as 3-month LIBOR minus T-Bill rate, (ii) the Moody's spread, defined as Moody's BAA minus AAA corporate bond rate (e.g., Engelberg, Gao, Jagannathan, 2009), (iii) the LIBOR itself (Jacobs and Weber, 2015), and (iv) the first principal component across the aforementioned variables. All measures attempt to capture the ease with which arbitrageurs can fund their positions to bet against mispricings. The underlying notion is that in real markets, where brokers demand traders to set up margin accounts as collateral, exploiting mispricings always requires arbitrageurs to bring up some capital, which becomes increasingly costly when the overall credit spread rises (Brunnermeier and Pedersen, 2009).

To capture market liquidity, I rely on (i) the aggregate level of liquidity as in Pástor and Stambaugh (2003), signed negatively to reflect illiquidity, (ii) the time-series of average bid-ask spreads, computed across the entire ETF universe and following the methodology proposed by Corwin and Schultz (2012), and again (iii) the first principal component. In times of low overall market liquidity, arbitrage returns should be larger, as market liquidity is positively related to transaction costs. With increasing transaction costs, executing arbitrage trades becomes more costly, widening the magnitude of price deviation that is necessary for arbitrageurs to profitably exploit mispricing.

Finally, I also consider the VIX, whose use as a proxy for market-level limits to arbitrage can be motivated in various ways (see Jacobs, 2015). Previous theoretical and empirical work suggests that the VIX is negatively related to funding liquidity (e.g., Brunnermeier and Pedersen, 2009) and positively related to aggregate idiosyncratic risk (Jacobs, 2015). Besides, the VIX is often interpreted as a "fear gauge", as evidence suggests that investors become more risk averse with an increasing VIX (Vayanos, 2004).

In Table 14, I regress both the percentage share of ETF pairs opening on a given day as well as event-time arbitrage returns individually on the aforementioned proxies. From Panel A it can be seen that the share of ETF pairs opening on a given day is significantly higher on days with low funding liquidity, but is unrelated to overall market liquidity. The share of pairs opening is also positively related to the VIX and the first principal component across all market-level proxies considered. Panel B reports coefficients for the regression of event-time returns on the respective proxies. Arbitrage profitability tends to be larger for positions opened on days with both low funding and market liquidity, as well as a higher VIX. The results are not only statistically, but also economically significant. For example, two-standard deviation increase in the first principal component across all proxies considered implies a 30 bps higher net-of-fee return for positions opened on that day. Overall, these results indicate that price gaps between nearly-identical ETFs are both more likely to occur and tend to be larger on days where the overall availability of arbitrage capital is limited.

[insert Table 14 here]

5 Conclusion

This paper examines price gaps between pairs of nearly identical Exchange-Traded Funds (ETFs). The rationale is that if ETFs are priced efficiently, prices should not only be aligned with NAVs, but also with the prices of competing ETFs. I utilize the benchmark index and leverage as stated in the fund prospectuses to identify pairs of close substitutes.

In line with expectations, prices normally move in close lockstep. However, it is surprisingly common that the bid price of an ETF exceeds the ask price of a competing fund. Over the period 2010 to 2016, I identified 4,983 cases where prices diverged by 20 bps or more, corresponding to a single price gap in a given ETF pair every 6 months. The median price distance amounts to 74 bps, but deviations can reach large extremes of up to several percent.

In principle, these price gaps must not necessarily stem from mispricing. Prices of seemingly identical ETFs can theoretically diverge for a number of fundamental reasons. Aside from differences in replication methodologies, I find that otherwise identical ETFs often exhibit pronounced differences in liquidity. Around 40 percent of the pairs in my sample employ different replication methodologies, and close to 80 percent are from different size and liquidity deciles.

However, the results from backtesting a simple pairs trading strategy that involves short selling the more expensive and buying the cheaper of both funds suggest that at least on average, price gaps cannot be justified by fundamental differences. In a portfolio context, this strategy historically yielded excess returns in the order of 1.2 to 1.8 percent per year, net of bid-ask spreads and common estimates for commissions, short-selling fees, and margin borrow costs. Out of a total of 4,983 identified price deviations, 71 percent and thus the vast majority converges, which should not be the case if prices typically gap for

fundamental reasons. We observe positive and significant returns for both, pairs where the more liquid fund is held short while the more illiquid is held long, and for pairs holding opposite positions. Similarly, there is no significant difference in the returns and convergence probabilities of ETF pairs holding identical or explicitly different security baskets to track their index. These results cast further doubt on the idea that the observed price deviations have its roots in fundamental differences, or that fundamental risk is a major concern that impedes arbitrageurs from betting against price gaps. Altogether, my findings suggest that the identified price distortions are typically due to mispricing.

Given that there is no convincing evidence for fundamental risk, and at the same time considering that the selected pairs trade in the same market, ETF pairs are a somewhat unique setting to examine the profitability and limits of relative-value arbitrage (see also Marshall, Nguyen, and Visaltanachoti, 2013). In previously studied contexts, asset pairs were either fundamentally different (e.g. dual-class shares, Schultz and Shive, 2010) or traded in different markets, i.e. were subjected to different institutional features (e.g., cross-listings as in Gagnon and Karolyi, 2010). This naturally raises the question what factors otherwise drive pairs of assets that can be considered *near-perfect* (rather than just close) substitutes to trade at different prices.

I find that despite fundamental risk being negligible, arbitraging ETFs against each other involves a number of other limitations. Price gaps are positively related to cross-sectional differences in transaction and holding cost proxies, chief among them being idiosyncratic risk and both secondary and primary market liquidity. Besides, I find that arbitrage profits are largely time-varying, and both more likely to occur and more pronounced in times of high market-wide limits to arbitrage. With respect to the microstructure conditions surrounding the day of price divergence, I find that price gaps typically arise following a sequence of days with abnormally low liquidity in both the relatively over- and underpriced ETF. On the day of divergence itself, we typically observe an abnormally high trading volume and turnover, which is indicative of a surge in liquidity risk. Both liquidity and liquidity risk revert to normal levels on the days succeeding the mispricing. Altogether, the findings discussed above imply that pairs of nearly identical ETFs can occasionally trade at different prices, because information diffusion is impeded by both cross-sectional and time-varying limits to arbitrage.

Consider a cautionary note at the end. It is impossible to assess whether the abnormal returns reported in this paper can actually be earned in real markets by implementing the outlined trading strategy. The backtests conducted throughout the article rather aimed on evaluating whether the observed price gaps are indicative of mispricing. We cannot say with certainty whether the documented arbitrage profits stand the test of an actual implementation.

Notwithstanding, there is an important implication for practitioners that consider buying or selling ETF shares and thus bear trading costs anyway. ETFs are becoming an increasingly popular investment vehicle in recent years. Aside from tight bid-ask spreads, investors benefit by growing competition pushing

down management fees to a few basis points per year. On the downside, however, the growing number of nearly identical ETFs implies an increase in market segmentation (see also Box, Davis, and Fuller, 2017) and comes at the risk that investors no longer see the forest for the trees when it comes to actual trading costs. My results suggest that practitioners are well advised to also compare share prices of competing funds before trading in order to avoid paying "shadow costs" (Petajisto, 2017) for trading at inefficient prices, in particular when intending to transact in less liquid ETFs or on days with high overall market frictions. To illustrate the economic magnitude of mispricing, consider that as of December 31, 2016, the management fee across all funds in my sample averages to 50 bps per year. The median price gap across all positions initiated between 2010 and 2016 amounts to 74 bps. Thus, when catching an unfavorable day and trading at inefficient prices, investors on average suffer an opportunity or effective loss worth more than an entire year of management fees.

References

Abreu, Dilip and Markus Brunnermeier (2002): Synchronization Risk and Delayed Arbitrage, Journal of Financial Economics 66, 341-360.

Ackert, Lucy, and Yisong Tian (2000): Arbitrage and Valuation in the Market for Standard & Poor's Depositary Receipts, Financial Management 29(3), 71-87.

Amihud, Yakov (2002): Illiquidity and Stock Returns: Cross-Section and Time-Series Effects, Journal of Financial Markets 5(1), 31-56.

Angel, James, Todd Broms, and Gary Gastineau (2016): ETF Transaction Costs Are Often Higher Than Investors Realize, Journal of Portfolio Management, 42(3), 65-75.

Asness, Clifford, Tobias Moskowitz, and Lasse Pedersen (2013): Value and Momentum Everywhere, Journal of Finance 68(3), 929-985.

Avellaneda, Marco, and Doris Dobi (2013): Price Inefficiency and Stock-Loan Rates of Leveraged ETFs, RISK Magazine (July 16, 2013), 37-40.

Barberis, Nicholas, and Richard Thaler (2003): A Survey of Behavioral Finance, in: Handbook of the Economics and Finance, Vol. 1, pp. 1053-1128.

Ben-David, Itzhak, Francesco Franzoni, and Rabih Moussawi (2017): Exchange Traded Funds (ETFs), Annual Review of Financial Economics 9, forthcoming.

Ben-David, Itzhak, Francesco Franzoni, and Rabih Moussawi (2017): Do ETFs Increase Volatility? NBER Working Paper.

Box, Travis, Ryan Davis, and Kathleen Fuller (2017): ETF Competition and Market Quality. Working Paper.

Broman, Markus (2016): Liquidity, Style Investing, and Excess Comovement of Exchange-Traded Fund Returns, Journal of Financial Markets 30, 27-53.

Broman, Markus, and Pauline Shum (2016): Does Liquidity Encourage Short-Term Trading? Evidence from Exchange-Traded Funds. Working Paper.

Brunnermeier, Markus, and Lasse Pedersen (2009): Market Liquidity and Funding Liquidity, Review of Financial Studies 22(6), 2201-2238.

Campbell, John, Sanford Grossman, and Jiang Wang (1993): Trading Volume and Serial Correlation in Stock Returns, Quarterly Journal of Economics 108, 905-939.

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

Charupat, Narat, and Peter Miu (2011): The Pricing and Performance of Leveraged Exchange-Traded Funds, Journal of Banking & Finance 35(4), 966-977.

Corwin, Shane, and Paul Schultz (2012): A Simple Way to Estimate Bid-Ask Spreads from Daily High and Low Prices, Journal of Finance 67(2), 719-760.

De Jong, Abe, Leonard Rosenthal and Mathijs van Dijk (2009): The Risk and Return of Arbitrage in Dual-Listed Companies, Review of Finance 13(3), 495-520.

De Long, Bradford, Andrei Shleifer, Lawrence Summers, and Robert Waldmann (1990): Noise Trader Risk in Financial Markets, Journal of Political Economy 98(4), 793-738.

Delcoure, Natalya, and Maosen Zhong (2007): On the Premiums of iShares, Journal of Empirical Finance 14(2), 168-195.

Do, Binh, and Robert Faff (2010): Does Simple Pairs Trading Still Work?, Financial Analysts Journal 66(4), 83-95.

Engelberg, Joseph, Pengjie Gao, and Ravi Jagannathan (2009): An Anatomy of Pairs Trading: The Role of Idiosyncratic News, Common Information and Liquidity. Working Paper.

Engle, Robert, and Clive Granger (1987): Co-integration and Error Correction: Representation, Estimation, and Testing, Econometrica 55(2), 251-276.

Engle, Robert, and Debojyoti Sarkar (2006): Premiums-Discounts and Exchange Traded Funds, Journal of Derivatives, 27-45.

Fama, Eugene, and Kenneth French (1992): The Cross-Section of Expected Stock Returns, Journal of Finance 47(2), 427-465.

Fama, Eugene, and Kenneth French (2015): A Five-Factor Asset Pricing Model, Journal of Financial Economics 116(1), 1-22.

Froot, Kenneth, and Emil Dabora (1999): How Are Stock Prices Affected by the Location of Trade?, Journal of Financial Economics 53, 189-216.

Fulkerson, Jon, Susan Jordan, and Denver Travis (2017): Bond ETF Arbitrage Strategies and Daily Cash Flow, Journal of Fixed Income 27(1), 49-65.

Fulkerson, Jon, Susan Jordan, and Timothy Riley (2014): Predictability in Bond ETF Returns, Journal of Fixed Income 23(3), 50-63.

Gagnon, Louis, and George Karolyi (2010): Multi-Market Trading and Arbitrage, Journal of Financial Economics 97(1), 53-80.

Gatev, Evan, William Goetzmann, and Geert Rouwenhorst (2006): Pairs Trading: Performance of a Relative Value Arbitrage Rule, Review of Financial Studies 19(3), 797-827.

Glosten, Lawrence, Suresh Nallareddy, and Yuan Zo (2017): ETF Activity and Informational Efficiency of Underlying Securities. Columbia Business School Working Paper.

Jacobs, Heiko and Martin Weber (2015): On the Determinants of Pairs Trading Profitability, Journal of Financial Markets 23, 75-97.

Jacobs, Heiko (2015): What Explains the Dynamics of 100 Anomalies?, Journal of Banking & Finance 57, 65-85.

Jegadeesh, Narasimhan (1990): Evidence of Predictable Behavior of Security Returns, Journal of Finance 45, 881-898.

Jegadeesh, Narasimhan, and Sheridan Titman (1990): Short-Horizon Return Reversals and the Bid-Ask Spread, Journal of Financial Intermediation 4(2), 116-132.

Jiang, Xinxin, and Stanley Peterburgsky (2017): Investment Performance of Shorted Leveraged ETF Pairs, Applied Economics 49, 4410-4427.

Johnson, Timothy (2008): Volume, Liquidity, and Liquidity Risk, Journal of Financial Economics 87(2), 388-417.

Levy, Ariel, and Offer Lieberman (2012): Overreaction of Country ETFs to US Market Returns: Intraday vs. Daily Horizons, Journal of Banking & Finance 37(5), 1412-1421.

Madhavan, Ananth, and Aleksander Sobczyk (2016): Price Dynamics and Liquidity of Exchange-Traded Funds, Journal of Investment Management 14(2), 1-17.

Marshall, Ben, Nhut Nguyen, and Nuttawat Visaltanachoti (2013): ETF Arbitrage: Intraday Evidence, Journal of Banking & Finance, 37, p. 3486-3498.

Meinhard, Christian, Sigrid Mueller, and Stefan Schoene (2015): Physical and Synthetic Exchange-Traded Funds: The Good, the Bad, or the Ugly?, Journal of Investing 24(2), 35-44.

Mitchell, Mark, Pulvino, Todd and Stafford, Erik (2002): Limited Arbitrage in Equity Markets, Journal of Finance, 57(2), p. 551-584.

Pástor, Lubos, and Robert Stambaugh (2003): Liquidity Risk and Expected Stock Returns, Journal of Political Economy 111(3), 642-685.

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

Pontiff, Jeffrey (2006): Costly Arbitrage and the Myth of Idiosyncratic Risk, Journal of Accounting and Economics 42(1), 35-52.

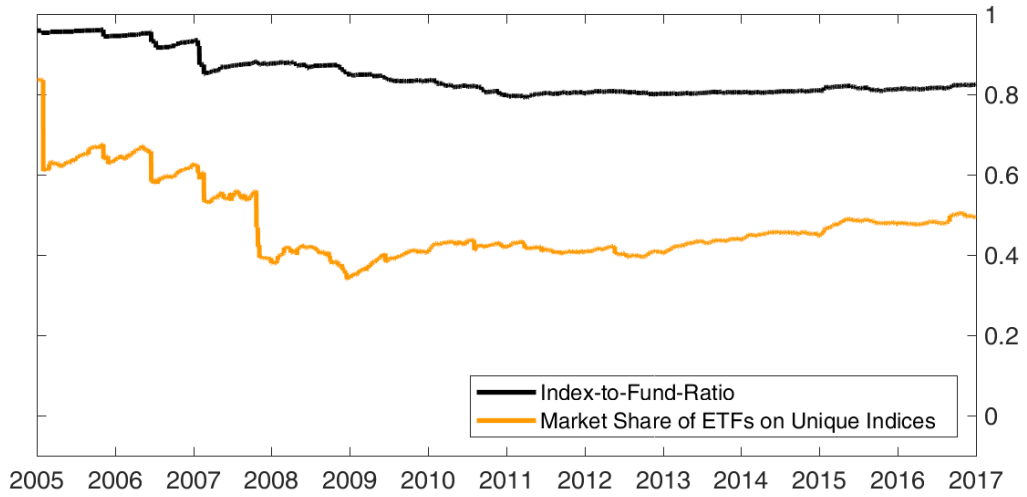
Schultz, Paul and Shive, Sophie (2010): Mispricing of Dual-Class Shares: Profit Opportunities, Arbitrage, and Trading, Journal of Financial Economics, 98, p. 524-549.

Stratmann, Thomas, and John Welborn (2013): Exchange-Traded Funds, Fails-to-Deliver, and Market Volatility. Working Paper.

Vayanos, Dimitri (2010): Flight to Quality, Flight to Liquidity, and the Pricing of Risk, NBER Working Paper.

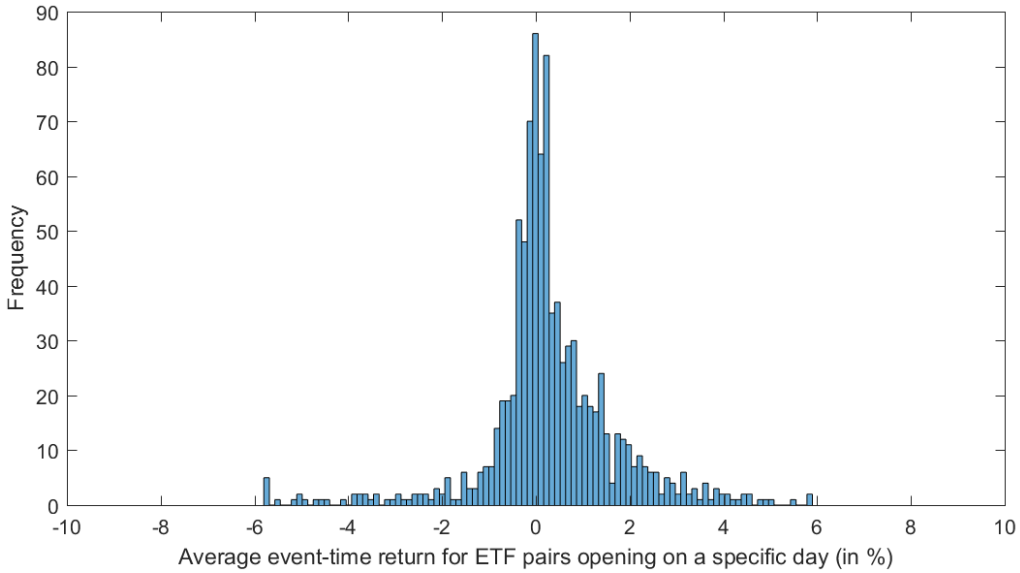
Appendix 1 – Figures

Figure 1: Index Homogeneity in the US ETF Market (2005-2016)



Homogeneity of the US ETF market over time. Specifically, this figure plots the (i) ratio of unique investable indices to investable ETFs, as well as (ii) the market share of unique ETFs, i.e. funds for which there is not a single competing fund tracking the same index.

Figure 2: Distribution of Average Returns on ETF Pairs Opening on a Given Day (2010-2016)



Following Jacobs and Weber (2015), this figure plots the distribution of one-month event-time returns for all ETF pairs opening on a given day. For each trading day in my sample where at least one ETF pair opens, I compute the average event-time return over the 20 days following the price gap. The figure is based on a total number of 4,983 trades.

Appendix 2 – Tables

Table 1: Sample Characteristics (2009-2016)

			Percentile						
	Mean	StdDev	Min	5	25	50	75	95	Max
Net assets ($ millions)	1,405.8	7,365.3	0.2	1.8	11.8	77.5	489.9	6,250.6	224,820.2
Daily volume ($ millions)	32.7	510.0	0.0	0.0	0.1	0.5	2.9	62.6	22,688.3
Daily turnover (%)	59.5	1,837.2	0.0	0.2	0.7	1.3	2.2	12.8	64,745.9
Bid-ask spread (bps)	43	78	1	3	10	21	49	131	879
Amihud ratio	5.98	21.78	0.00	0.00	0.01	0.14	2.92	28.56	586.81
Premium (bps)	2	74	-1,982	-25	-2	2	11	42	552
Premium volatility (bps)	57	61	0	5	18	42	75	158	788

This table shows cross-sectional characteristics of the funds considered in the sample, based on the entire sample period 2009-2016. Daily volume, turnover and premiums are computed as means throughout the period, whereas bid-ask spread is the time-series median and premium volatility is the time-series volatility. Amihud (2002) ratio is computed as time-series average across the daily ratios between absolute mid-quote returns and dollar trading volume.

Table 2: Descriptive Pair and Pair Portfolio Statistics

Panel A: Pair Frequency		
Number of selected pairs (total)	4,118	
Number of selected pairs (period average)	49	
Panel B: Size and Liquidity Characteristics of Paired-Up Funds		
Net assets ($ mil)	7,224	(590)
Bid-ask spread (bps)	11	(5)
Turnover ratio (%)	2.8	(0.8)
Amihud ratio	0.066	(0.002)
Panel C: Pair Portfolio Composition		
Weights by size and liquidity:		
Funds in top five size deciles	0.84	
Funds in lowest five spread deciles	0.85	
Funds in lowest five Amihud deciles	0.68	
Pairs from different size deciles	0.78	
Pairs from different spread deciles	0.76	
Pairs from different Amihud deciles	0.78	
Avg. size decile difference	2.66	
Avg. spread decile difference	2.35	
Avg. Amihud decile difference	3.09	
Weights by asset class:		
Equity fund pairs	0.71	
Commodity fund pairs	0.16	
Fixed income fund pairs	0.12	
Other fund pairs	0.01	
"Smart beta" fund pairs	0.39	
Weights by ETF sponsor and replication:		
Pairs with matching sponsors	0.05	
Pairs with same replication	0.61	
Pairs with mixed security lending activities	0.11	
Pairs with different management fees	0.03	

This table reports the frequency of pairs (Panel A), size and liquidity characteristics for the funds selected in at least a single pair (Panel B), as well as the average composition of pairs portfolios (Panel C), pooled across all trading cycles. In Panel B, net assets are recorded at the end of the preceding formation periods, whereas the other measures reported are computed using daily data throughout the 12-month formation periods. “Amihud ratio” is computed as in Amihud (2002). Bid-ask spreads are computed as time-series medians, whereas turnover ratios are time-series averages. In Panel C, pairs classified as “other” include funds seeking allocation to more than one asset class and funds tracking alternative benchmarks (such as volatility indices). “Pairs with different management fees” refers to the average weight of ETF pairs whose management differ by more than 20 basis points per year. Numbers reported in Panel B and C are averages, except for numbers in parentheses, which are medians.

Table 3: Premium and Price Spread Characteristics of ETF Pairs

Panel A: Premium Characteristics		
Premium (bps)	2	(1)
Premium volatility (bps)	33	(17)
Premium correlation	0.40	(0.33)
Pairs with significant correlation (%)	71	
Panel B: Price Spread Characteristics		
Spread mean	-0.0024	(-0.0001)
Spread standard deviation	0.0053	(0.0027)
Co-integrated pairs (%)	78	
Co-integrated in trading period (%)	64	

This table reports premium and price spread characteristics for the selected ETF pairs, pooled across all trading cycles. Both the numbers in Panel A and B are based on price and premium series over the preceding formation periods. In Panel A, premium volatility is the time-series standard deviation of daily premiums. Premium correlation is the pairwise Pearson correlation coefficient for the premium/discount series of the paired-up funds. "Pairs with significant correlation" refers to the share of pairs, whose premium/discount series are significantly correlated at the 5-percent-level. In Panel B, "spread" refers to the percentage distance between the bid price of one pair leg and the ask price of the other leg. "Co-integrated pairs (%)" refers to the percentage share of pairs where the null hypothesis of co-integration cannot be rejected at the 1-percent-level, based on the Engle and Granger (1987) two-step method. "Co-Integrated in trading period (%)" refers to the share of pairs that remain statistically co-integrated during the trading period. All measures reported are cross-sectional means across, except for percentage shares and numbers in parentheses, which are cross-sectional medians.

Table 4: Monthly Return Distribution and Performance Measures (2010-2016)

	Gross-of-fee Returns		Net-of-fee Returns		Market
	ROCC	ROEC	ROCC	ROEC	
Panel A: Return Distribution					
Average return	0.0020	0.0025	0.0012	0.0016	0.0110
t-statistic	3.6157 ***	3.5646 ***	2.4647 ***	2.4769 ***	2.6401 ***
Average excess return	0.0019	0.0024	0.0011	0.0015	0.0109
t-statistic	3.4685 ***	3.4748 ***	2.3152 **	2.3853 ***	2.6280 ***
Return distribution					
Standard deviation	0.0023	0.0036	0.0021	0.0033	0.0380
Median	0.0013	0.0019	0.0009	0.0010	0.0124
Skewness	1.1307	1.2451	1.1227	1.2543	-0.1073
Kurtosis	3.9570	4.3679	4.1286	4.6616	3.0867
Min	-0.0026	-0.0039	-0.0030	-0.0051	-0.0788
Max	0.0081	0.0137	0.0067	0.0124	0.1135
Observations < 0	0.2143	0.1905	0.3810	0.3810	0.3571
Value-at-risk	-0.0015	-0.0022	-0.0021	-0.0031	-0.0604
Conditional value-at-risk	-0.0020	-0.0031	-0.0025	-0.0038	-0.0692
Panel B: Risk-Adjusted Performance					
Sharpe ratio	0.6559	0.6705	0.4429	0.4626	0.2885
Excess return on VaR	0.9903	1.0754	0.4339	0.4916	0.1806
Excess return on CVaR	0.7330	0.7870	0.3639	0.3977	0.1576

Summary statistics for the monthly pair portfolio and market returns. Market returns are obtained from the Kenneth French Data Library. "ROCC" is return on committed capital, whereas "ROEC" is return on employed capital. ROCC divides by the number of selected pairs, while ROEC adjusts with the number of pairs that actually traded. Both value-at-risk and conditional value-at-risk are computed using a 0.95 confidence level. ***, **, * means statistically significant on a 1 percent, 5 percent and 10 percent level, respectively. *t*-statistics are computed using Newey-West standard errors with six-lag correction.

Table 5: Net-of-fee Alphas and Systematic Risk (2010-2016)

	Return on Committed Capital (ROCC)			Return on Employed Capital (ROEC)		
	(1)	(2)	(3)	(1)	(2)	(3)
Factors						
Intercept	0.0011 **	0.0010 **	0.0010 **	0.0015 **	0.0015 **	0.0014 ***
Market	0.0002	0.0007	0.0003	-0.0017	-0.0013	-0.0009
SMB	-0.0101	-0.0107 *	-0.0091	-0.0166 *	-0.0171 *	-0.0137
HML	-0.0076	-0.0050	-0.0116	-0.0149	-0.0128	-0.0203
MOM		0.0057			0.0046	
RMW			0.0016			0.0088
CMA			0.0131			0.0194
R-squared	0.02	0.03	0.02	0.03	0.03	0.03

Alphas and systematic risk exposures for the ETF arbitrage strategy, based on monthly net-of-fee portfolio returns. Specification (1) corresponds to Fama-French three-factor model (Fama and French, 1992), whereas (2) refers to the Carhart four-factor model (Carhart, 1997). Specification (3) is the Fama-French five-factor model (Fama and French, 2015). All factor premiums are sourced from the Kenneth French Data Library. ***, **, * means statistically significant on a 1 percent, 5 percent and 10 percent level according to the *t*-statistic, respectively. *t*-statistics are computed using the Newey-West standard errors with six-lags correction.

Table 6: Trading Statistics and Return Distribution for Individual Positions

Total number of trades		4,983		Net-of-fee returns		
Total number of converged trades		3,556		Mean	0.0037	***
Convergence probability (%)		71.36			*(2.95)*	
Margin calls (% of all trades)		0.54		Mean (converged positions)	0.0065	***
					(4.54)	***
Duration of converged trades (days)	Mean	10		Mean (unconvgd positions)	-0.0034	***
	Median	3			*(-3.28)*	
Number of pairs traded per period	Mean	32		Standard deviation	0.0132	
	Median	35		Min	-0.0653	
Share of pairs traded per period	Mean	0.64		P01	-0.0418	
	Median	0.66		P10	-0.0080	
Number of trades per pair	Mean	1.2		P25	-0.0011	
	Median	1.0		P50	0.0023	
Number of roundtrips per traded pair	Mean	1.3		P75	0.0101	
	Median	1.0		P90	0.0190	
				P99	0.0397	
				Max	0.0445	
Mean return (gross of fees)		0.0064	***	Skewness	-0.79	
		(5.38)		Kurtosis	5.39	

This table shows the frequency of price and the distribution of event-time returns, pooled across all trading cycles. Numbers in parentheses are *t*-statistics, computed using standard errors that are two-way clustered for the underlying benchmark index and the day prices diverge. ***, **, * means statistically significant on a 1 percent, 5 percent and 10 percent level, respectively. All return measures reported are winsorized at the 1-percent-level.

Table 7: Robustness of Arbitrage Profitability for Varying Trading Parameters

		#Positions	%Converged	Single Position Return	3F-Alpha (ROEC)		3F-Alpha (ROCC)	
Formation period length	6 months	6,544	73	0.0029	0.0015	**	0.0011	**
	18 months	3,945	70	0.0040	0.0019	***	0.0013	***
Trading leriod length	3 months	2,745	64	0.0037	0.0027	***	0.0013	***
	12 months	8,471	81	0.0033	0.0016	**	0.0011	**
Std. deviation trigger	k = 1	8,063	77	0.0017	0.0013	*	0.0010	*
	k = 3	3,190	67	0.0036	0.0019	**	0.0010	**
Minimum price difference	10 bps	5,175	71	0.0033	0.0019	**	0.0012	**
	80 bps	2,639	68	0.0079	0.0028	***	0.0013	***

Sensitivity of the pairs trading strategy towards a variation of trading parameters. In each row, one of the trading parameters is varied, while the others are held constant as defined in the baseline strategy (see section 3). All return measures reported are net of fees and winsorized at the 1-percent-level. Alphas are computed using the Fama-French three-factor specification (Fama and French, 1992). ***, **, * means statistically significant on a 1 percent, 5 percent and 10 percent level, respectively. *t*-statistics are computed using Newey-West standard errors with six-lags correction.

Table 8: Arbitrage Profitability and Short-Selling Constraints

			Return		
	N	(%)	Total	Long Leg	Short Leg
Panel A: Leveraged and Inverse (LIN) ETFs					
Pairs of LIN ETFs	90	2	0.0082 *	-0.0034	0.0116 **
			(1.91)	*(-0.87)*	*(2.31)*
Pairs of non-leveraged equivalents	1,301	26	0.0048 **	0.0088 ***	-0.0040
			(2.23)	*(3.52)*	*(-1.07)*
Return difference			-0.0034	0.0122 ***	-0.0156 *
			(-1.59)	*(3.38)*	*(-1.70)*
Panel B: All Pairs					
All pairs	4,983	100	0.0037 ***	0.0063 ***	-0.0028
			(2.95)	*(3.98)*	*(-1.58)*
All pairs (without LIN pairs)	4,893	98	0.0038 ***	0.0069 ***	-0.0030 *
			(3.02)	*(4.33)*	*(-1.71)*
Return difference			0.0001	0.0005	-0.0003
			(0.78)	*(1.47)*	*(-0.95)*

This table reports arbitrage returns separately for pairs of leveraged/inverse and non-leveraged ETFs. For both types of pairs, returns are separately reported for both the long and short leg of the position. Returns are event-time net-of-fee returns. Panel A shows returns for leveraged and inverse fund pairs (for brevity grouped under "leveraged funds") versus equivalent non-leveraged pairs, i.e. non-leveraged pairs tracking the same indices as those covered by at least one leveraged/inverse fund pair. Panel B shows how the results change when excluding leveraged/inverse pairs. Numbers reported in parentheses are *t*-statistics with standard errors two-way clustered for the day positions are initiated and the underlying benchmark index. ***, **, * means statistically significant on a 1 percent, 5 percent and 10 percent level. All return measures reported are winsorized at the 1-percent-level.

Table 9: Liquidity Differences and Arbitrage Profitability

		By Bid-Ask Spread		By Amihud Ratio		By Turnover	
Long liquid ETF	N	2,346		2,386		2,427	
	% of all trades	47		48		49	
	Mean return	0.0031	**	0.0023	*	0.0039	***
		(2.16)		*(1.67)*		*(2.92)*	
	Mean return: long leg	0.0102	***	0.0108	***	0.0086	***
		(4.19)		*(4.72)*		*(3.86)*	
	Mean return: short leg	-0.0072	***	-0.0085	***	-0.0049	**
		(-3.63)		*(-4.40)*		*(-2.40)*	
Long illiquid ETF	N	2,637		2,588		2,556	
	% of all trades	56		54		53	
	Mean return	0.0041	***	0.0050	***	0.0035	**
		(3.29)		*(3.91)*		*(2.57)*	
	Mean return: long leg	0.0028	**	0.0021	*	0.0041	***
		(2.21)		*(1.79)*		*(3.12)*	
	Mean return: short leg	0.0012		0.0029		-0.0007	
		(0.45)		*(1.12)*		*(-0.23)*	
Illiquid minus liquid	Total	0.0010		0.0028	***	-0.0004	
		(0.90)		*(2.78)*		*(-0.41)*	
	Long leg	-0.0073	**	-0.0088	***	-0.0045	
		(-2.39)		*(-3.02)*		*(-1.33)*	
	Short leg	0.0084	**	0.0117	***	0.0043	
		(2.58)		*(3.75)*		*(1.14)*	

This table reports event-time returns separately for (i) positions where the more liquid of both ETFs in pair is held long, and (ii) for positions where the more illiquid fund is held long. Liquidity is either measured by bid-ask spreads, Amihud (2002) illiquidity, or turnover. Liquidity measures are computed as in Table 2. "Illiquid minus liquid" refers to differences in average returns between both types of price gap. Numbers in parentheses are *t*-statistics with standard errors two-clustered for both the underlying benchmark index and the day positions are initiated. ***, **, * means statistically significant on a 1 percent, 5 percent and 10 percent level. All return measures reported are winsorized at the 1-percent-level.

Table 10: Price and NAV Deviations on Event Days (2010-2016)

		Open	Close	Change
Relative price distance	Mean	0.0120 ***	0.0010	-0.0110 ***
	t-stat	8.19	1.11	-6.62
	Median	0.0074	-0.0006	-0.0066
Relative NAV distance	Mean	0.0058 ***	0.0040 ***	-0.0018
	t-stat	5.42	3.58	-1.42
	Median	0.0030	0.0002	-0.0001
Premium difference	Mean	0.0061 ***	-0.0013 ***	-0.0073 ***
	t-stat	4.38	-3.67	-4.28
	Median	0.0010	-0.0002	-0.0018

This table reports percentage price and NAV deviations as well as premium differences for both the days arbitrage positions are established and the days positions are closed, measured across all positions initiated between 2010 and 2016. Relative price distances are computed as percentage distance between the bid price of the short leg and the ask price of the long leg. NAV differences are the percentage difference between the NAV of the short and long leg. Premium differences are defined as premium of the short leg minus the premium of the long leg, whereas the premium of a fund is computed as percentage distance between the bid-ask mid-quote price from the underlying NAV. The reported *t*-statistics refer to *t*-tests with standard errors two-way clustered for the day of divergence and underlying benchmark index. ***, **, * means statistically significant on a 1-percent-, 5-percent- and 10-percent level. All variables are winsorized on the 1-percent-level.

Table 11: Differences in Replication Methods and Fundamental Risk

	Same Replication	Different Replication	Difference
Number of trades	2,368	2,615	247
Mean return	0.0054 ***	0.0021 *	-0.0033
	(2.81)	*(1.68)*	*(-1.59)*
Mean return (converged)	0.0085 ***	0.0047 ***	-0.0038
	(3.94)	*(3.41)*	*(-1.61)*
Mean return (unconverged)	-0.0025 ***	-0.0042 **	-0.0018
	(-3.21)	*(-2.35)*	*(-0.94)*
Return std. dev. of unconverged trades	0.0112	0.0153	0.0041 ***
			[0.54]
Convergence probability	0.7196	0.7082	-0.0114
			{0.89}
Duration of converged trades (days)	15.44	16.29	-0.85

Frequency and profitability of arbitrage positions for both pairs of funds employing the same and pairs employing explicitly different replication methodologies. Returns are 20-day event-time returns, net of fees (see section 4.3). Numbers in parentheses are *t*-statistics with standard errors two-clustered for the day where positions are initiated and the underlying benchmark index, whereas ***, **, * means statistically significant on a 1 percent, 5 percent and 10 percent level. Numbers in square brackets and braces refer to the test statistics for two-sample *F*-tests on equal variances and two-sample *z*-tests on equal proportions, respectively. All return measures reported are winsorized at the 1-percent-level.

Table 12: Profitability and Cross-Sectional Differences in Limits to Arbitrage

Panel A: Univariate analysis

					Difference	
High idiosyncratic risk	0.0068	**	Low idiosyncratic risk	0.0003	0.0064	**
	(2.54)			*(0.39)*	*(2.34)*	
	[0.0060]			[0.0011]	[0.0049]	
Low net assets	0.0043	***	High net assets	0.0021	0.0022	
	(2.61)			*(0.90)*	*(0.95)*	
	[0.0048]			[0.0012]	[0.0037]	
High bid-ask spread	0.0064	**	Low bid-ask spread	0.0002	0.0062	**
	(2.26)			*(0.15)*	*(2.07)*	
	[0.0054]			[0.0011]	[0.0043]	
High Amihud (2002)	0.0088	***	Low Amihud (2002)	-0.0007	0.0095	***
	(4.36)			*(-0.73)*	*(3.93)*	
	[0.0093]			[0.0005]	[0.0088]	
Low turnover ratio	0.0061	***	High turnover ratio	0.0015	0.0046	***
	(3.62)			*(0.89)*	*(3.11)*	
	[0.0030]			[0.0020]	[0.0010]	
Low trading volume	0.0053	***	High trading volume	0.0017	0.0036	
	(4.16)			*(0.65)*	*(1.55)*	
	[0.0038]			[0.0015]	[0.0023]	
Low primary market activity	0.0066	***	High primary market activity	0.0010	0.0056	**
	(3.04)			*(0.70)*	*(2.31)*	
	[0.0037]			[0.0018]	[0.0018]	

Table 12: Profitability and Cross-Sectional Differences in Limits to Arbitrage *(continued)*

Panel B: Multivariate analysis				
Intercept	0.0094 **	0.0109 **	0.0085 **	0.0125 ***
	(2.34)	*(2.06)*	*(2.42)*	*(2.88)*
Idiosyncratic risk	0.0031 ***			0.0017 ***
	(3.95)			*(3.17)*
Net assets	0.0010			
	(1.33)			
Bid-ask spread	0.0022 *		0.0011	0.0009
	(1.85)		*(0.81)*	*(0.81)*
Amihud (2002) ratio		0.0043 ***	0.0039 ***	0.0033 ***
		(6.85)	*(6.34)*	*(5.71)*
Turnover ratio		-0.0026 ***		-0.0012 *
		(-2.88)		*(-1.69)*
Trading volume		0.0016 **	0.0006	0.0015
		(2.20)	*(1.25)*	*(1.05)*
Primary market activity	-0.0026 ***		-0.0021 ***	-0.0021 ***
	(-4.87)		*(-3.86)*	*(-3.60)*
N	4,962	4,974	4,974	4,962
Adjusted R^2	0.24	0.25	0.26	0.27

Cross-sectional determinants of arbitrage profits. Panel A reports average event-time returns conditioned on subsamples formed based on seven different arbitrage cost proxies, whereas "low" refers to the lowest and "high" to the highest quartile with regard to the respective measure. Panel B shows estimated coefficients of multivariate regressions, where event-time returns are regressed on arbitrage cost proxies and a number of control variables, including dummies for the year, month, and day of the week were positions are initiated, dummies for the underlying asset class, as well as the daily premia on the three Fama and French (1992) factors. The explanatory variables considered are the seven arbitrage cost proxies from Panel A, standardized to have zero mean and unit variance in order to ease interpretation. All arbitrage cost proxies are estimated using daily data over the preceding formation period, except for total net assets, which are recorded on the last day of the formation period. Returns are winsorized at the 1-percent-level. Numbers in parentheses are *t*-statistics with standard errors two-way clustered for the day where positions are initiated and the underlying benchmark index. ***, **, * means statistically significant on a 1 percent, 5 percent and 10 percent level. Numbers in square brackets are medians.

Table 13: Returns, Liquidity and Trading Activity Around the Day of Divergence

		Event-Day						
		-3	**-2**	**-1**	**0**	**1**	**2**	**3**
Abnormal spread (%)	Long leg	2.55 ***	4.82 ***	3.14 ***	-0.48	0.12	0.88	-0.79
	Short leg	2.34 ***	3.01 ***	2.86 *	-0.80 *	1.70 **	-0.04	0.07
Abnormal volume (%)	Long leg	-1.47	-1.47	0.92	3.84 ***	-4.27 ***	-2.21 ***	0.13
	Short leg	-0.13	-0.01	1.44 **	2.79 ***	-3.65 ***	-0.53	-0.71
Abnormal turnover (%)	Long leg	-1.65	-1.10	1.54	4.99 ***	-6.14 ***	-3.39 ***	0.42
	Short leg	1.21	-0.02	2.69 ***	3.17 ***	-5.04 ***	-0.55	-1.33
Abnormal primary market activity (%)	Long leg	1.13	0.95	1.20 *	3.39 ***	0.71	-0.44	1.10
	Short leg	-0.28	0.83 **	1.62 **	2.70 ***	-0.23	0.52	-0.07
Abnormal creations (%)	Long leg	1.46	0.92	1.44 ***	1.62 *	3.19 ***	1.30 *	2.62 ***
	Short leg	0.02	0.85	2.23 ***	0.42	-0.28	0.27	0.02
Abnormal redemptions (%)	Long leg	-0.08	-2.94 **	1.72	4.26 ***	-0.30	-1.69	1.74
	Short leg	-0.91	0.59	1.02	5.40 ***	-0.05	0.95	-0.91

This table reports daily abnormal bid-ask spreads, abnormal trading activity, and abnormal primary market activity for the day where prices diverge and the six days surrounding. The figures reported are averages across all detected price gaps, separately computed for the long and short leg. Spread, volume, turnover, and primary market activity are the logarithm of the daily bid-ask spread in basis points, number of shares traded, shares traded in basis points of shares outstanding, and absolute change in shares outstanding, respectively. All measures are then divided by their corresponding time-series average across all 180 days surrounding the day of divergence. "Abnormal creations" and "abnormal redemptions" are conditional on a positive and negative change in shares outstanding, respectively. Specifically, share creations (redemptions) equal the absolute change in shares outstanding on days where the change from the previous day is positive (negative), and are zero else. ***, **, * means statistically significant on a 1 percent, 5 percent and 10 percent level according to *t*-statistics with standard errors two-way clustered for the day where positions are initiated and the underlying benchmark index. All variables are winsorized at the 1-percent-level separately for each event-day. The figures reported control for day-of-the-week effects.

Table 14: Sensitivity Towards Market-Level Limits to Arbitrage

Panel A: Share of Pairs Opened				
Funding illiquidity	TED spread	0.0015	*	*(1.85)*
	Moody's spread	0.0025	***	*(3.01)*
	LIBOR	0.0028	**	*(2.12)*
	Funding liquidity PC	0.0019	***	*(2.78)*
Market illiquidity	Corwin-Schultz	0.0000		*(-0.08)*
	Pástor-Stambaugh	0.0001		*(0.17)*
	Market liquidity PC	0.0000		*(0.06)*
Aggregated	VIX	0.0019	**	*(2.44)*
	All PC	0.0016	***	*(2.83)*
Panel B: Event-Time Returns				
Funding illiquidity	TED spread	0.0012	**	*(2.39)*
	LIBOR	0.0024	***	*(3.48)*
	Moody's spread	0.0008	*	*(1.88)*
	Funding liquidity PC	0.0012	***	*(3.39)*
Market illiquidity	Corwin-Schultz	0.0008		*(1.26)*
	Pástor-Stambaugh	0.0006	*	*(1.77)*
	Market liquidity PC	0.0009	**	*(1.99)*
Aggregated	VIX	0.0016	**	*(2.52)*
	All PC	0.0014	***	*(4.07)*

This table explores how the frequency of pairs opening on a given day as well as arbitrage profitability relates to market-level impediments to arbitrage. Panel A reports estimated coefficients of multivariate regressions, where the percentage share of pairs opening on a given day, averaged across all trading cycles including that day, is regressed separately on a number of time-varying limits-to-arbitrage proxies and a set of control variables. Panel B examines how event-time returns relate to the conditions prevailing on the day prices diverge. Specifically, one-month event-time returns, pooled across all trading cycles, are regressed individually on the limits-to-arbitrage proxies and control variables. Both the regressions in Panel A and Panel B control for calendar effects (year, month, day of the week), daily market returns, size, and value premia. Panel B also includes dummies for the asset class underlying the fund pair and the cross-sectional determinants analyzed in Table 12. "Moody's spread" refers to the daily difference between Moody's BAA and AAA corporate bond rate. "Corwin-Schultz" is the daily bid-ask spread computed as in Corwin and Schultz (2012), averaged across the whole ETF universe on a given day. "Pástor-Stambaugh" refers to the Pástor and Stambaugh (2003) liquidity factor measured over the month preceding the price gap. To ease interpretation, the limits-to-arbitrage proxies considered are standardized with respect to their time-series mean and variance. "Funding liquidity PC" and "Market liquidity PC" are the first principal component of funding liquidity and market liquidity proxies, respectively. "All PC" is the first principal component across all proxies considered. Event-time returns are winsorized at the 1-percent-level. Numbers in parentheses are *t*-statistics. In Panel B, *t*-statistics are computed using standard errors two-way clustered for the day where positions are initiated and the underlying benchmark index. ***, **, * means statistically significant on a 1 percent, 5 percent and 10 percent level.