ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics
Master Thesis Financial Economics

U.S. equity ETFs: tracking efficiency and the COVID-19 crisis

Name student: Mariska Groeneveld
Student ID number: 450653
Supervisor: Dr. R.R.P. Kouwenberg
Second assessor: Dr. J.J.G. Lemmen
Date final version: 09/12/2021

Abstract

This study examines how the tracking efficiency of U.S. equity ETFs evolves over time, and by which factors it can be explained. Tracking efficiency is measured by the tracking error, which is the difference in performance between an ETF and its target index. We analyze 1,149 U.S. equity ETFs from 2001 to 2020, using panel data from Morningstar Direct and by applying regression analysis and t-tests. The results show that U.S. equity ETFs exhibit significant tracking errors from 2001 to 2020. In addition, these tracking errors are higher during the financial crisis of 2008 and the year of the coronavirus stock market crash in 2020. Tracking errors increase with volatility and the bid-ask spread, and when tracking a foreign index. In contrast, having a larger fund size, applying a full replication strategy, following a smart beta strategy and having a sector orientation, all result in lower tracking errors. Lastly, the net expense ratio, dividends and trading volume do not affect tracking errors.

Acknowledgements

I would like to thank my Master Thesis supervisor, Dr. R.R.P. Kouwenberg, for his great guidance. During the process the communication was clear and effective. In addition, the professor motivated me and provided me with valuable feedback. Furthermore, I would like to express my gratitude towards Dr. J.J.G. Lemmen for being my second assessor. Lastly, I would like to thank my family and friends for their support.

Table of contents

1 Introduction

In 2008, U.S. investors had around 530 billion dollars invested in exchange-traded funds (ETFs) (Horch, 2020). In the course of 2020 this amount increased to more than 4 trillion dollars. The popularity of ETFs has skyrocketed after the 2008 financial crisis, because investors began to recognize that their actively managed portfolios involved much higher costs and failed to outperform index funds. In 2020 alone, a flow of 507.4 billion dollars into U.S. listed ETFs was observed, breaking the 2017 record of a 476.1 billion dollar inflow (Saha, 2021). Furthermore, the supply of new ETFs has increased considerably in recent years. In 2020 a total of 318 new ETFs were launched and already 83 ETFs have been introduced in the first quarter of 2021.

As ETFs grow in popularity and size, it is important to keep an eye on whether their objective of providing investors with the same returns as the benchmark index, is being achieved. The extent to which an ETF tracks the returns of its underlying benchmark index is called tracking efficiency (Buetow & Henderson, 2012). The tracking efficiency of ETFs is usually measured by the tracking error, which is the difference in performance between an ETF and its target index (Chu, 2011). Many studies, focusing on different samples of ETFs and several time periods, show that ETFs exhibit significant tracking errors (Blitz et al., 2012; Chu, 2011; W. F. Johnson, 2009; Milonas & Rompotis, 2006; Shin & Soydemir, 2010; Svetina & Wahal, 2008). In addition, it was investigated whether particular factors have an increasing or decreasing effect on tracking errors of ETFs (Blitz et al., 2012; Blitz & Huij, 2012; Buetow & Henderson, 2012; Chu, 2011; Frino & Gallagher, 2002; Rompotis, 2011; Svetina & Wahal, 2008). The main factors examined are the expense ratio, volatility, dividends, trading volume, fund size, bid-ask spread, and whether the fund is applying a full replication strategy. Moreover, various studies have been conducted with a focus on the financial crisis of 2008 and 2009, which showed that tracking errors were significantly higher during this crisis period (Buetow & Henderson, 2012; Mateus & Rahmani, 2017; Qadan & Yagil, 2012). Although some studies have been carried out on the effect of the COVID-19 crisis on bond funds and ETFs, to our knowledge, no research has yet been conducted on the tracking efficiency of equity ETFs during the COVID-19 crisis (Falato et al., 2020; O'Hara & Zhou, 2021). This is the gap this study seeks to fill, by analyzing tracking errors of U.S. equity ETFs over time with a focus on the crisis years 2008 and 2020, and by assessing the effect of the most important factors on tracking errors. Therefore, the main research question of this study is:

"How does the tracking efficiency of U.S. equity ETFs evolve over time, and by which factors can it be explained?"

This research question will be answered by testing several hypotheses, which are mentioned in the theoretical framework. The first hypothesis states that U.S. equity ETFs exhibit significant tracking errors over the period from January 2, 2001 to December 31, 2020. Subsequently, a positive effect is expected from the expense ratio, volatility, dividends and bid-ask spread on tracking errors of U.S. equity ETFs. In contrast, the trading volume, fund size and the application of a full replication strategy are expected to have a negative impact on tracking errors of U.S. equity ETFs. Finally, the latter hypothesis states that tracking errors of U.S. equity ETFs are higher during the financial crisis of 2008 and the COVID-19 crisis of 2020.

To test these hypotheses and answer the research question, this study analyzes 1,149 U.S. equity ETFs from 2001 to 2020. All required panel data is collected from Morningstar Direct. The methodology of this study builds on the methodology of Buetow and Henderson (2012). First of all, the tracking errors per year are calculated for each ETF. Then, one-tailed one-sample t-tests are applied to test whether tracking errors are significantly greater than zero. Subsequently, regression analysis is applied to examine the effect of the different factors on tracking errors. For this, a random-effects GLS regression model with time-fixed effects and robust clustered standard errors is used. Finally, Welch's one-tailed two-sample t-tests are applied to test whether tracking errors are significantly higher in the crisis years 2008 and 2020 compared to the other years in the sample period.

The results of this study show that U.S. equity ETFs exhibit significant tracking errors over the period from January 2, 2001 to December 31, 2020. This means that these ETFs are generally unable to perfectly replicate the total returns of their benchmark indices. In addition, it is demonstrated that tracking errors are significantly higher during the financial crisis of 2008 and the year of the coronavirus stock market crash in 2020. Moreover, it appears that several factors have a significant influence on tracking errors. The volatility of the underlying benchmark index total returns and the bid-ask spread of an ETF have a significant positive effect on tracking errors. Also, tracking a foreign index results in higher tracking errors. In contrast, having a larger fund size, applying a full replication strategy, following a smart beta strategy and having a sector orientation results in significantly lower tracking errors. Lastly, the net expense ratio, dividends and trading volume do not appear to have a significant effect on tracking errors.

This research increases the knowledge about the tracking efficiency of U.S. equity ETFs, by studying tracking errors and the factors that affect them over a 20-year period that includes two different crisis periods. Tracking errors are not only found to be higher during the financial crisis of 2008, as already shown in previous studies, but also in the year of the coronavirus stock market crash in 2020. This result is a significant contribution to the existing literature on the tracking efficiency of ETFs. In addition, the knowledge about tracking errors and the factors that influence them has been extended to a longer and more recent period, which runs from 2001 to 2020.

The remainder of this paper is organized as follows. Section 2 gives a brief explanation of index tracking products and the concept of tracking efficiency, and reviews relevant previous research. Subsequently, section 3 describes the data. Section 4 describes the methodologies applied. The results found are discussed in section 5. Section 6 includes the discussion, which further elaborates on the results, and whether these are in line with expectations and findings of previous studies. The seventh section covers the conclusion, in which the hypotheses and research question are answered, the limitations of this research are discussed and recommendations for future research are given. Section 8 provides a list of references used. Finally, section 9 includes the appendix, with supporting figures and tables.

2 Theoretical Framework

2.1 Index tracking products and the concept of tracking efficiency

Both index funds and ETFs are passive index tracking products, which aim to provide investors with the same risk and returns as the underlying benchmark index (Chu, 2011). According to Kostovetsky (2003) the objective of both types of funds is the same, namely to provide investors with a well-diversified portfolio of investments that track a particular index, by exploiting economies of scale by buying large volumes of stocks at low cost. However, index funds originated in 1972, while ETFs only emerged in 1993 (Agapova, 2011).

Although both products have the same purpose and many similarities, they cannot be considered as perfect substitutes because of their distinctly different structure. Accordingly, both products may satisfy different investor needs. Agapova (2011) provides evidence suggesting that ETFs are preferred by tax-sensitive investors, while index mutual funds are preferred by investors who are excluded from, or insensitive to, paying taxes. The key differences between index funds and ETFs have to do with taxation efficiency, shareholder transaction fees, management fees and several other qualitative differences (Kostovetsky, 2003). Another important difference between ETFs and index funds is that ETFs are traded on an exchange, while index funds are not (Buetow & Henderson, 2012). Next to the liquidity on the secondary market, ETFs even have an additional layer of liquidity on the primary market, where authorized participants create and redeem ETF shares in response to the prevailing market conditions. However, unlike index funds, ETFs are not exchangeable on a daily basis at the reported net asset value (NAV). This is because the price of an ETF is not only dependent on the value of the assets it invests in, but also on the supply and demand of the ETF itself.

Now that it is clear what an ETF is, we can look at the concept of tracking efficiency, which is the ultimate goal of the product. The tracking efficiency of an ETF shows how closely the ETF tracks the returns of its underlying benchmark index (Buetow & Henderson, 2012). Perfect tracking efficiency

is achieved when the ETF returns exactly replicate the returns of the benchmark index. However, perfect tracking is not feasible, because the index cannot be directly invested in, and it is considered as a 'paper' portfolio without frictions while ETFs are not able to duplicate the returns of the index without cost (Buetow & Henderson, 2012; Pope & Yadav, 1994). ETFs are therefore expected to underperform their underlying indices, if only because of transaction costs. As a result, many ETFs apply partial replication rather than full replication of the benchmark index in order to minimize transaction costs. So, there is a trade-off between tracking efficiency and transaction costs (Svetina & Wahal, 2008).

The tracking efficiency of an ETF can be measured by the tracking error, which represents the difference in performance between an ETF and its target index (Chu, 2011). This tracking error can be determined in different ways, of which the most widely used are the three definitions suggested by Pope and Yadav (1994). Firstly, they define tracking error as the average absolute difference between the ETF and benchmark index returns. Subsequently, they measure tracking error as the standard deviation of return differences between the ETF and its benchmark index. The third and last way in which they estimate tracking errors is by the Standard Error of Regression (SER) for the application of the Capital Asset Pricing Model (CAPM), which will be further clarified in the methodology section. Cresson, Cudd and Lipscomb (2002) extend these definitions of tracking error by taking the R-squared of the same CAPM model regression, which can be considered as a more straightforward and naïve measure of tracking performance.

2.2 Overview of findings on tracking errors of equity ETFs

Research on the tracking efficiency of index tracking products started with the investigation of index mutual funds and proceeded with researching the tracking ability of ETFs. Frino and Gallagher (2001) find significant tracking errors for 42 S&P 500 index funds for the five years to February 1999, induced by market frictions. Significant tracking errors are also found in the performance of Australian equity index funds over the period running from July 1989 to March 1999 (Frino & Gallagher, 2002). However, these funds do not systematically under- or outperform their benchmark index on a before cost basis.

Milonas and Rompotis (2006) do find significant underperformance for ETFs traded on the Swiss Stock Exchange from August 2001 to April 2006. The average tracking error is equal to 1.02%, which they find to be significant and sufficiently large to conclude that these Swiss ETFs are unable to fully replicate the returns on their benchmark indices. A similar study on the tracking efficiency of 35 Swiss equity ETFs is performed for a subsequent period from August 2012 to August 2014, which also shows significant tracking errors (Naumenko & Chystiakova, 2015). In addition, Blitz, Huij and Swinkels (2012) find an annual underperformance of 50 to 150 basis points for 40 European index funds and

ETFs tracking multiple broadly diversified equity market indices from 2003 to 2008. Another European study into the tracking efficiency of ETFs is performed by Meinhardt, Mueller and Schoene (2015). They report significant tracking errors for a sample of 286 equity, 117 fixed income and 18 total return/commodity ETFs listed at the Frankfurt Stock Exchange from January 2010 to August 2011. Moreover, significant tracking errors are observed for ETFs listed in New Zealand and Hong Kong (Chen, Chen & Frijns, 2017; Chu, 2011).

Most of the literature on the tracking efficiency of ETFs is based on the European and U.S. market. Therefore, now the main results regarding the tracking errors of ETFs listed at the U.S. market will be discussed. Economically significant premiums are found for the 20 iShares equity country ETFs from inception to October 2002, after controlling for time-zone measurement error and transaction costs (Delcoure & Zhong, 2007). However, this mispricing is not persistent. Johnson (2009) also explores the tracking efficiency of 20 U.S. foreign country ETFs, by analyzing the correlation coefficients between the ETF and its corresponding index. In general, the ETFs exhibit poor short-term tracking ability from 1997 to 2006. Though the ETFs investing in Mexico, Canada and Brazil do consistently track their benchmark index. Furthermore, a large sample of 584 U.S. domestic equity, international equity and fixed income ETFs on average are not immune from tracking error. These ETFs show underperformance with respect to their benchmark indices from their inception to the end of 2007 (Svetina & Wahal, 2008). Similarly, persistent tracking errors are found for 20 MSCI country ETFs and 6 U.S. ETFs invested in broad domestic equity markets for the period from July 2004 to June 2007 (Shin & Soydemir, 2010). In addition, 50 Barclay's iShares ETFs invested in various sector, broad and country indices show return superiority compared to their benchmark from 2002 to 2007, which is strongly persistent in the short term (Rompotis, 2011). This large body of evidence on the tracking efficiency of ETFs suggests that ETF returns do not replicate the returns of their underlying benchmark index. This results in the first hypothesis:

Hypothesis 1: U.S. equity ETFs exhibit significant tracking errors over the period from January 2, 2001 to December 31, 2020.

2.3 Overview of findings on factors impacting tracking errors of equity ETFs

In this section the most important factors that may affect the tracking efficiency of ETFs are discussed. The influence of costs on the tracking efficiency of ETFs is the factor most widely reviewed and accepted in literature. Direct costs, mostly measured by the fund's annual total expense ratio, increase tracking errors and thus decrease the tracking efficiency of ETFs (Blitz et al., 2012; Chu, 2011; Naumenko & Chystiakova, 2015; Osterhoff & Kaserer, 2016; Rompotis, 2011).

But also implicit transaction costs, caused by low liquidity, reduce the tracking efficiency of ETFs (Mateus & Rahmani, 2017; Osterhoff & Kaserer, 2016). High bid-ask spreads and low traded volumes make the creation and redemption process more expensive and risky, thereby negatively impacting the tracking ability of ETFs. Bid-ask spreads therefore have a positive effect on tracking errors (Delcoure & Zhong, 2007; Frino & Gallagher, 2002). For their subsample of equity ETFs, Meinhardt, Mueller and Schoene (2015) also find a positive significant effect of spreads on tracking errors for three out of four tracking error measures. Besides, the log transformed average daily trading volume has a significant negative effect on tracking errors (Buetow & Henderson, 2012; Meinhardt, Mueller & Schoene, 2015). Naumenko and Chystiakova (2015) expected to find a similar result, as a higher degree of liquidity should result in lower tracking errors. However, they find no significant negative effect of the log transformed average daily trading volume on ETF tracking errors. On the other hand, Delcoure and Zhong (2007) use the natural logarithm of trading volume as an approximation for the difference in investor views, and therefore expect a positive effect of this variable on tracking errors. In contrast to the other studies, they find a positive significant effect of trading volume on tracking errors. However, based on the aggregate of these results, a positive effect of liquidity on the tracking efficiency of ETFs is expected. As a result, the bid-ask spread is expected to have a positive effect on tracking errors and the trading volume is expected to negatively influence tracking errors.

Another factor that appears to have a negative effect on tracking errors is fund size (Buetow & Henderson, 2012; Chu, 2011). The rationale behind this factor is that a large fund has lower transaction costs because of economies of scale, resulting in lower tracking errors.

Furthermore, Chiang (1998) identifies that index volatility is one of the main factors contributing to index fund tracking errors. Subsequently, Frino and Gallagher (2002) stress that an index fund mostly does not have the exact same composition and weighting of stocks as the benchmark index, which also applies to ETFs. Therefore, unsystematic changes in stock prices underlying the benchmark index, that are not part of the ETF, result in tracking error. The same applies to price changes of stocks the ETF is overweighted in relative to the benchmark index. Consequently, benchmark index volatility, or what is widely referred to as market volatility, is found and expected to have a positive effect on tracking errors. Rompotis (2011) finds a similar result which, however, is defined and described in a slightly different manner. According to this study, the volatility of ETF returns on an annual basis has a significant positive effect on tracking errors, which is explained by higher risk. Meinhardt, Mueller and Schoene (2015) also find a significant positive effect of the standard deviation of ETF returns on tracking errors for a large sample of 412 physical and synthetic ETFs traded on the Frankfurt Stock Exchange. These studies have in common that they find a positive

effect of market volatility, also defined as risk, on tracking errors. As a result, the volatility of index returns is expected to positively influence tracking errors.

As mentioned earlier, there is a trade-off between transaction costs and tracking efficiency (Frino & Gallagher, 2002; Svetina & Wahal, 2008). With a full replication strategy all securities underlying the benchmark index are held in the exact same proportion as in the index. This suggests low tracking errors, but the high transaction costs involved negatively impact the tracking efficiency. There are also funds that apply partial replication, which is also called stratified sampling. With this strategy only a subset of the securities underlying the benchmark index are invested in, resulting in higher tracking errors, but lower transaction costs. Frino and Gallagher (2002) therefore expect and also find statistically significant lower tracking errors for ETFs with a full replication strategy. This is consistent with the finding of Blitz and Huij (2012) that ETFs with a statistical replication strategy are prone to significantly higher levels of tracking error compared to funds with a full replication strategy. Furthermore, several studies investigate the effect of synthetic replication on ETF tracking errors. Compared to traditional ETFs, synthetic ETFs use derivatives to replicate the benchmark index, instead of owning the physical assets underlying the index (Naumenko & Chystiakova, 2015). Johnson, Bioy, Kellet and Davidson (2013) find that synthetic ETFs have tracking errors that are on average 30 basis points lower than those of physical ETFs. They claim that synthetically replicated ETFs have better tracking performance because they do not pay dividends and tend to be cheaper. Meinhardt, Mueller and Schoene (2015) do not find significant differences in tracking errors for their sample of synthetic equity ETFs compared to the physical equity ETFs examined. However, for their sample of fixed income ETFs, the application of a synthetic replication strategy does result in lower tracking errors compared to physical replication. In contrast, Naumenko and Chystiakova (2015) conclude that synthetically replicated ETFs have significantly higher tracking errors than traditional ETFs. Whereas Mateus and Rahmani (2017) do not find significant differences in the daily tracking performance of synthetically versus physically replicated ETFs. So despite the many studies conducted, there seems to be no consensus on the effect of synthetic replication on the tracking efficiency of ETFs. However, for this study the effect of synthetic replication on the tracking performance of ETFs is irrelevant, because only few U.S. asset managers make use of a synthetic replication strategy. This is because of the specific regulations set by the U.S. Securities and Exchange Commission (SEC) in 2010. These regulations proscribe the launch of new funds with a synthetic replication strategy by asset managers that did not already sponsor a synthetic ETF. The sample of this study does not contain any ETF with a synthetic replication strategy. Therefore, the effect of synthetic replication on tracking errors is not examined.

High tracking errors can also arise from the payment of dividends (Frino & Gallagher, 2002; Kostovetsky, 2003; Meinhardt, Mueller & Schoene, 2015; Osterhoff & Kaserer, 2016; Shin & Soydemir, 2010). Since there is a delay in the receipt of those dividends and there are transaction costs involved

in their reinvestment, while the index assumes them to be immediately reinvested without cost, this results in tracking error. Contrary to expectations, Frino and Gallagher (2002) and Shin and Soydemir (2010) do not find a significant effect of dividends on tracking errors. Meinhardt, Mueller and Schoene (2015) do find a significantly positive effect of dividends on tracking errors, but only for the subsample of synthetic equity ETFs for two out of four tracking error measures. To conclude, the findings of Osterhoff and Kaserer (2016) do confirm that dividends have a significantly positive effect on tracking errors. Although the results of previous studies are not unequivocal about the effect of dividends on tracking errors, based on the rationale and results of Osterhoff and Kasarer (2016), the expectation is that dividends have a positive effect on tracking errors.

Based on the discussed results of the leading studies on the performance and tracking efficiency of ETFs, the following hypotheses are formulated:

Hypothesis 2: The expense ratio has a positive effect on tracking errors of U.S. equity ETFs.
Hypothesis 3: The volatility of benchmark index total returns has a positive effect on tracking errors of U.S. equity ETFs.
Hypothesis 4: Dividends have a positive effect on tracking errors of U.S. equity ETFs.
Hypothesis 5: The trading volume has a negative effect on tracking errors of U.S. equity ETFs.
Hypothesis 6: The fund size has a negative effect on tracking errors of U.S. equity ETFs.
Hypothesis 7: The application of a full replication strategy has a negative effect on tracking errors of U.S. equity ETFs.
Hypothesis 8: The bid-ask spread has a positive effect on tracking errors of U.S. equity ETFs.

2.4 Overview of findings on tracking errors of equity ETFs in crisis times

From September 2008 to March 2009 extraordinary conditions applied to financial markets, with frozen credit markets, high volatility and large price swings (Buetow & Henderson, 2012). During this period of market stress, ETFs were not able to closely replicate the returns on their underlying indices (Buetow & Henderson, 2012; Drenovak, Urošević & Jelic, 2014; Johnson, Bioy, Kellett & Davidson, 2013; Mateus & Rahmani, 2017; Qadan & Yagil, 2012). According to Buetow and Henderson (2012) this particularly applied to less-liquid asset classes. They determine tracking errors by subtracting the annual return of the benchmark index from the annual ETF return. For the year 2009, this results in an average tracking error of -1.7% for equity ETFs and even -2.6% for fixed income ETFs. These results indicate that the extreme market conditions in the years 2008 and 2009 significantly impacted ETF tracking errors across all asset classes, and particularly those tracking less-liquid asset classes. Also Qadan and Yagil (2012) find evidence of a substantial lower tracking ability of 42 iShares ETFs tracking

Dow Jones industrial sector stock indices during the year 2008 in comparison to the years 2006 and 2007. Tracking errors of these U.S. domestically invested ETFs increased significantly in 2008 compared to 2006 and 2007. Mateus and Rahmani (2017) similarly found larger daily tracking errors during this crisis period for ETFs listed at the London Stock Exchange, which they attribute to higher bid-ask spreads, lower trading volumes, and a high volatility of exchange rates. Together, all these influences made the creation and redemption process more risky and costly, which negatively impacted the tracking efficiency of ETFs during the financial crisis. After the financial crisis, the ETFs' tracking performance improved significantly. The aggregate of these findings indicate that tracking errors are higher in times of crisis, which results in the following hypothesis:

Hypothesis 9: Tracking errors of U.S. equity ETFs are higher during the financial crisis of 2008 and the COVID-19 crisis of 2020.

3 Data

3.1 Data collection

This study analyzes all equity ETFs listed in the United States which are classified as an index fund during the period running from January 3, 2000 to May 28, 2021, with complete data. Ultimately, this results in the analysis of 1,149 ETFs from 2001 to 2020. This research looks from the perspective of a U.S. investor who trades in U.S. dollars and does not want to be exposed to exchange rate risk. Therefore, only the ETFs with the U.S. dollar as their base currency are considered, and all data is displayed in U.S. dollars. In addition, both existing ETFs and ETFs that no longer exist are included in the sample, in order to mitigate the survivorship bias problem.

All required data is collected from Morningstar Direct, which is a global investment analysis platform that provides a global multi-asset investment database. To start with, daily ETF total returns are calculated based on the daily market return index. This index is a total return index that represents the value of one ETF unit purchased and owned since the ETF's origination, assuming that all dividends are reinvested on the ex-dividend date. Important to note is that this variable is by Morningstar described as a market return index, but in literature it is referred to as a total return index. Therefore, this variable is referred to as total return index (TRI) from now on. The daily ETF total returns are calculated according to the following formula:

$$TR_{i,t} = \frac{TRI_{i,t} - TRI_{i,t-1}}{TRI_{i,t-1}} \cdot 100\% \qquad (1)$$

In this formula, $TR_{i,t}$ gives the total return of ETF i on day t in percentages; $TRI_{i,t}$ is the total return index of ETF i on day t and $TRI_{i,t-1}$ is the total return index of ETF i on day t-1.

Index total returns are also determined on the basis of this daily market return index, which I refer to as total return index. In this case the total return index represents the value of the benchmark index from inception, assuming that all dividends are reinvested on the ex-dividend date. Daily index total returns are calculated according to the following formula:

$$TR_{b,t} = \frac{TRI_{b,t} - TRI_{b,t-1}}{TRI_{b,t-1}} \cdot 100\% \quad (2)$$

In this formula, $TR_{b,t}$ gives the total return on benchmark index b on day t in percentages; $TRI_{b,t}$ is the total return index of benchmark index b on day t and $TRI_{b,t-1}$ is the total return index of benchmark index b on day t-1.

After all ETF and benchmark index daily total returns are calculated, all weekend days and U.S. stock market holidays are excluded from the dataset. Furthermore, observations where the ETF or benchmark index total return equals minus 100 are removed from the dataset. These observations concern the day the respective ETF ceased to exist or from which no prices are available anymore for the benchmark index. Subsequently, these ETF and benchmark index total returns, which are the determining inputs for all tracking error measures, are winsorized at the 0.1th and 99.9th percentile. This reduces the effect of extreme outliers. The originally unwinsorized extremely high ETF total returns are mostly observed on days when the trading volume of the respective ETF is very low, which results in a sharp increase in the price of the ETF. In many cases a correction follows the day after, resulting in an extremely negative return of the ETF with a corresponding high trading volume, causing a large negative difference in return between the ETF and its benchmark index on that day. High ETF total returns also occur on days when the ETF is heavily traded, while it has not been traded at all in the few preceding years. This results in a high ETF return and therefore a substantial return difference. Winsorizing the original data also removes extreme errors from the dataset.

Expenses are captured in the annual report net expense ratio. This annual ratio reflects the percentage of assets used to pay management fees and operating expenses, including administrative fees, 12b-1 fees, and all other asset-based expenses incurred by the fund, except for brokerage costs. This expense ratio does not include sales charges. The effect of the bid-ask spread is captured by the daily average spread-to-price ratio. This ratio takes the average spread and divides it by the day end mid price, which is the average of the bid and ask price at the end of the day. The average spread is an average of all the spreads over a trading day, which is based on real time data and is calculated every time the bid or ask price is updated. The calculation of the daily average spread-to-price ratio is presented in the following formula:

$$Spread/price\ ratio = \frac{daily\ average\ spread}{\frac{day\ end\ ask\ price + day\ end\ bid\ price}{2}} = \frac{daily\ average\ spread}{day\ end\ mid\ price} \quad (3)$$

The variable that captures the volatility of index total returns, in previous research also labeled as risk, is calculated on an annual basis as the standard deviation of daily index total returns. The annual

dividends of the ETFs are also collected, and are expressed in USD. Subsequently, the number of traded shares are collected for each ETF on a daily basis. The variable volumes is then constructed by taking the average of these daily traded volumes for every single ETF-year, which is expressed in billions. Fund size is measured by the average daily share-class total net assets, and is expressed in billions of USD. In addition, information is gathered about the replication method of all ETFs. Subsequently, the dummy variable full replication is created, which takes a value of one for ETFs with a full replication strategy and zero otherwise. Moreover, two dummy variables are created that highlight the crisis years. The variable financial crisis takes the value of one for the observations that fall in the year 2008, and zero otherwise. The variable COVID-19 is equal to one for the observations in 2020, and equals zero for the remainder of the sample period.

Finally, several control variables are added to the analysis to control for the effect of the type of index the ETF is tracking. For example, ETFs tracking a foreign index tend to have higher tracking errors compared to domestic ETFs (Rompotis, 2011; Svetina & Wahal, 2008). Therefore, a dummy variable is created that indicates whether the ETF is focused on the U.S. or mainly includes foreign investments. This dummy variable foreign is equal to one when the ETF's average exposure to the U.S. over the sample period considered is smaller than 85%, and equals zero otherwise. Subsequently, the dummy variable sector controls for whether the ETF has a sector or broad market orientation. This variable is equal to one if the ETF is classified by Morningstar as sector equity or if the prospectus objective of the ETF is a sector specialty, and is equal to zero otherwise. Furthermore, we account for the market capitalization of the securities the ETFs invest in. The Morningstar equity style box provides the market capitalization per ETF on a monthly basis, which is divided into small cap, mid cap and large cap. Based on this variable, the average market capitalization is determined for each ETF for each year. If the average market capitalization does not result in one of the three categories the lowest market capitalization is assumed. Subsequently, the dummy variables small cap and mid cap are constructed, which are equal to one if the ETF for that year is mainly invested in small cap and mid cap securities respectively. Finally, the dummy variable smart beta is added, which is equal to one if the ETF is following a smart beta strategy[1], and equals zero if it is not.

[1] A smart beta strategy aims to achieve higher risk-adjusted returns than the returns on traditional cap-weighted indices (Cazalet, Grison & Roncalli, 2014). Therefore, assets are weighted differently than based on market capitalization, which is why smart beta is also known as alternative-weighted indexing. The smart beta strategy resembles active management, but the big difference is that smart beta is applied on a rules-based and transparent way.

3.2 Descriptive statistics independent variables

The descriptive statistics of all independent variables are shown in table 1. The annual report net expense ratio of the ETFs ranges from a negative 0.08% to a positive 2.03%, where the average net expense ratio equals 0.42%. The volatility of the underlying indices, which is measured by the standard deviation of daily index total returns, is on average equal to 1.22%. The lowest volatility measured in a year equals 0.23%, and the volatility was at most 4.92%. Dividends have a mean value of 0.94 USD, and vary between 0 and 22.64 USD. The average daily traded volume is equal to 29.4 million. The lowest average traded volume is equal to 2, and the highest average traded volume equals 65.32 billion, resulting in a huge variance in this variable. The same applies to the variable fund size, which is measured by the average daily share-class total net assets. Fund size is 726,223 USD at its lowest and 288.53 billion USD at its highest, with an average of 2.18 billion USD. In about 42.7% of the ETF years, a full replication strategy is applied. This means that a partial replication strategy is applied in about 57.3% of cases. In addition, 3.2% and 9.8% of the observations are measured in the years 2008 and 2020 respectively. For 47.9% of the observations, the ETF mainly includes foreign investments. For the remaining 52.1%, and thus majority of observations, there is a focus on U.S. investments. In 32.4% of the cases there is a sector orientation, and therefore ETFs have a broad market orientation in 67.6% of cases. In 61.9% of the sample ETFs focus on large cap investments. While in 27.0% of the sample the focus is on mid cap investments, and even only 11.1% on small cap investments. Finally, a smart beta strategy is applied in 41.1% of cases.

Table 1 Descriptive statistics of the independent variables

Variable	Obs.	Mean	Median	Std. Dev.	Min	Max	Skewness	Kurtosis
NER	8,896	0.4172	0.4400	0.2110	-0.0800	2.0300	0.1798	3.1249
Volatility	8,896	1.2201	1.0318	0.6166	0.2324	4.9196	1.4852	5.5721
Dividends	8,896	0.9404	0.7256	0.9059	0.0000	22.6400	5.7886	82.5543
Volumes	8,896	0.0294	0.0000	0.9239	0.0000	65.3224	50.6463	3105.0630
Fund size	8,896	2.1835	0.1975	10.3200	0.0007	288.5266	14.3860	289.4267
Full replication	8,896	0.4267	0.0000	0.4946	0.0000	1.0000	0.2964	1.0878
Financial crisis	8,896	0.0320	0.0000	0.1761	0.0000	1.0000	5.3148	29.2471
COVID-19 crisis	8,896	0.0978	0.0000	0.2971	0.0000	1.0000	2.7081	8.3337
Foreign	8,896	0.4789	0.0000	0.4996	0.0000	1.0000	0.0846	1.0072
Sector	8,896	0.3239	0.0000	0.4680	0.0000	1.0000	0.7529	1.5668
Small cap	8,896	0.1109	0.0000	0.3141	0.0000	1.0000	2.4775	7.1380
Mid cap	8,896	0.2701	0.0000	0.4440	0.0000	1.0000	1.0354	2.0721
Smart beta	8,896	0.4113	0.0000	0.4921	0.0000	1.0000	0.3605	1.1299

Notes table 1: Where NER stands for net expense ratio.

4 Methodology

The methodology of this study on the tracking efficiency of U.S. ETFs builds on the methodology of Buetow and Henderson (2012), because to my knowledge this is the leading research into the tracking efficiency of U.S. ETFs in which the examination of a crisis period is included. The study examines a large sample of 845 ETFs over a long sample period running from 1994 to 2010, which includes the financial crisis of 2008.

To evaluate the tracking efficiency of the ETFs, I compare the ETF total returns to the total returns of the benchmark index. By using this approach, the two components of tracking error, which are the NAV tracking error and the component caused by variation of the market price around the NAV, are combined into one single measure (Buetow & Henderson, 2012). The difference in total return between the benchmark index and the NAV reflects the performance of the fund management. Whereas the difference in total return between the ETF and the NAV is the result of the efficiency of the creation and redemption process, and the interplay of supply and demand for the ETF itself. Because the focus of this study is on an ETF's ability to provide investors with the same returns as the underlying benchmark index, I use total returns based on market prices instead of net asset values.

4.1 Tracking error measurement

For the entire sample of ETFs three daily tracking error measures are computed and reported for every calendar year. As stated earlier, the three definitions of tracking error suggested by Pope and Yadav (1994) are the ones most widely used in previous literature and are therefore applied in this research. So for each ETF-year the average daily absolute difference in return, the standard deviation of daily return differences and the Standard Error of Regression (SER) for the application of the Capital Asset Pricing Model (CAPM) are determined. In the calculation of these tracking error measures only the observations for which both the ETF and benchmark index total return are known, are included. Furthermore, all ETF-years with less than 30 observations are removed from the sample.

The first definition of tracking error is the average daily absolute difference in return between the ETF and its corresponding benchmark index, $TE1_{i,y}$, which is calculated as shown in the following formula:

$$TE1_{i,y} = \frac{\sum_{t=1}^{n} |TR_{i,t} - TR_{b,t}|}{n} = \frac{\sum_{t=1}^{n} absolute\ difference_{i,b,t}}{n} \quad (4)$$

In this formula, $TE1_{i,y}$ is the first average daily tracking error for ETF i in year y; $TR_{i,t}$ is the total return on ETF i on day t; $TR_{b,t}$ is the total return on benchmark index b on day t, n is the number of daily

returns and $absolute\ difference_{i,b,t}$ is the absolute difference in total return between ETF i and benchmark index b on day t.

Subsequently, tracking errors are measured by the standard deviation of return differences between the ETF and its underlying benchmark index, $TE2_{i,y}$, as shown in the following formula:

$$TE2_{i,y} = \sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(difference_{i,b,t} - \overline{difference}_{i,b,y})^2} \quad (5)$$

Here $TE2_{i,y}$ is the second measure of daily tracking error for ETF i in year y; n is the number of daily returns; $difference_{i,b,t}$ is the difference in total return between ETF i and benchmark index b on day t and $\overline{difference}_{i,b,y}$ is the average daily difference in total return between ETF i and benchmark index b in year y. One shortcoming of this tracking error measure is that it may result in zero when the ETF consistently outperforms or underperforms its benchmark index by the same magnitude (Chu, 2011).

The third and last way in which tracking errors are estimated is by the Standard Error of Regression (SER) resulting from the estimation of the Capital Asset Pricing Model (CAPM) which is given in the following formula:

$$TR_{i,t} = \alpha_i + \beta \cdot TR_{b,t} + \varepsilon_{i,t} \quad (6)$$

In this formula $TR_{i,t}$ is the total return on ETF i at day t; α_i is the return an investor can get on ETF i that is not related to its benchmark index; β reflects the degree to which the ETF return replicates its benchmark index return; $TR_{b,t}$ is the total return on benchmark index b at day t and $\varepsilon_{i,t}$ are the residuals of the regressions. When beta is equal to one, this measure of tracking error is exactly equal to the one calculated as the standard deviation of return differences between ETF and benchmark index returns (Pope & Yadav, 1994). However, when beta is not exactly equal to one, both measures of tracking error, $TE2_{i,y}$ and $TE3_{i,y}$, may have different values. Pope and Yadav (1994) also point out that this measure of tracking error may overstate tracking errors when the relationship between the ETF return and the return of its underlying index is nonlinear.

The definitions of tracking error in this study deviate slightly from those of Buetow and Henderson (2012). They omit the tracking error based on the SER, and use the average daily difference in return between the ETF and its corresponding benchmark index as a tracking error measure instead. When the daily differences in return are averaged, negative and positive differences are (partially) offset, resulting in the underestimation of tracking errors. However, it is not required that tracking errors are close to zero, on average, but that ETF returns consistently match those of their benchmark index. As the average difference in return between the ETF and its benchmark index is still informative regarding the relative performance of the ETF compared to its benchmark, we include it in the analysis,

but refer to it as relative performance instead of tracking error. For each ETF the relative performance compared to its benchmark index is computed for every calendar year according to formula seven:

$$RP_{i,y} = \frac{\sum_{t=1}^{n}(TR_{i,t} - TR_{b,t})}{n} = \frac{\sum_{t=1}^{n} difference_{i,b,t}}{n} \quad (7)$$

In this formula, $RP_{i,y}$ is the relative performance of ETF i in year y compared to its benchmark index b; $TR_{i,t}$ is the total return on ETF i on day t; $TR_{b,t}$ is the total return on benchmark index b on day t, n is the number of daily returns and $difference_{i,b,t}$ is the difference in total return between ETF i and benchmark index b on day t.

I start by analyzing and describing the descriptive statistics of all tracking error measures, the relative performance and the alpha, beta, R-squared and adjusted R-squared resulting from the regressions of formula 6, which are presented in the next subsection. The relative performance and alpha are annualized by multiplying the values by 252, the number of trading days in a year, to facilitate interpretation.

4.1.1 Descriptive statistics tracking errors and performance measures

The descriptive statistics of all three tracking error measures, the relative performance and the alpha, beta, R-squared and the adjusted R-squared resulting from the regressions of formula 6 are shown in table 2. All three tracking error measures show excess kurtosis and a relatively high positive skewness. This means that the distributions of all tracking error measures are heavy tailed and have a fat tail to the right side of the distribution. This can also be observed from the histograms in figures A1 to A3 of appendix A. The average daily absolute difference in return varies from 0.02% to 4.04%, with an average of 0.41%. The average standard deviation of daily return differences is equal to 0.59%, but can be 0.03% at its lowest and 5.40% at its highest. The third tracking error, calculated as the SER, has a mean of 0.55% and varies between 0.03% and 5.22%. Based on these descriptive statistics tracking errors of this sample of U.S. equity ETFs seem substantial, but whether these are statistically significant will be tested and shown in the results section. The annualized average daily difference in return between the ETFs and their benchmark indices is equal to -0.0043%, meaning that these ETFs are slightly underperforming their benchmark indices over the entire sample period considered. In addition, the annualized relative performance was -68.19% at its lowest and 80.32% at its highest. As the annualized average daily difference in return is affected by extreme outliers, the median of -0.22% is more representative for the relative performance of U.S. equity ETFs.

The annualized alpha, which is the return an investor can get on an ETF that is not related to its benchmark index, is on average equal to 0.78%, which implies outperformance. However, some ETFs show an underperformance of -103.95% or even an outperformance of 97.58% compared to their benchmark index in particular years. Hence, the annualized alpha is strongly affected by outliers, and the median of 0.05% is more informative.

The mean and median beta are equal to 0.88 and 0.96 respectively, which shows that on average daily ETF total returns closely, but certainly not perfectly, replicate the total returns of the underlying benchmark index. Also, the average and median R-squared of 72.61% and 83.23%, and average and median adjusted R-squared of 72.49% and 83.16%, indicate that daily benchmark index total returns explain a fairly large portion of ETF daily total return variation.

Table 2 Descriptive statistics of tracking errors and other performance measures

Variable	Obs.	Mean	Median	Std. Dev.	Min	Max	Skewness	Kurtosis
TE1	8,896	0.4082	0.3047	0.3748	0.0203	4.0376	1.6641	7.8904
TE2	8,896	0.5864	0.4449	0.5338	0.0260	5.4030	1.6415	7.5057
TE3	8,896	0.5546	0.4308	0.4915	0.0253	5.2240	1.5132	6.7007
RP	8,896	-0.0043	-0.2185	3.8142	-68.1904	80.3233	-0.8608	61.6070
Alpha	8,896	0.7771	0.0508	6.8185	-103.9464	97.5766	-0.6388	38.6769
Beta	8,896	0.8761	0.9611	0.2135	-0.2786	1.7341	-2.1817	8.4924
R-squared	8,896	0.7261	0.8323	0.2881	0.0000	0.9995	-0.8772	2.6580
Adjusted R-squared	8,896	0.7249	0.8316	0.2894	-0.0168	0.9995	-0.8781	2.6606

Notes: This table presents the descriptive statistics of the three tracking error measures, the relative performance (RP) and the alpha, beta, R-squared and adjusted R-squared resulting from the regressions of formula 6. All variables are calculated per ETF-year. TE1 is the average daily absolute difference in return between the ETF and its benchmark index; TE2 measures the standard deviation of return differences between the ETF and its benchmark index; TE3 is estimated by the Standard Error of Regression (SER) resulting from the application of the following regression to each ETF-year: $TR_{i,t} = \alpha_i + \beta \cdot TR_{b,t} + \varepsilon_{i,t}$; RP, the relative performance, is the annualized average daily difference in return between the ETF and its benchmark index; Alpha is an annualized measure that represents the return an investor can get on the ETF that is not related to its benchmark index; Beta reflects the degree to which the ETF total return replicates its benchmark index total return; R-squared and adjusted R-squared indicate the extent to which the ETF total returns can be explained by the benchmark index total returns.

4.2 Methodology hypothesis 1: tracking errors

To test the first hypothesis, one-tailed one-sample t-tests are applied to test whether tracking errors are significantly greater than zero. If the results of these t-tests are statistically significant, assuming a significance level of 5%, it means that the ETFs are generally unable to perfectly replicate the returns of their benchmark indices, which is what is expected. Statistically insignificant results, on the other hand, lead to the conclusion that the ETFs track their benchmark indices well.

Subsequently, the size of the tracking errors is interpreted and compared with comparable studies into the tracking efficiency of U.S. ETFs. Also, the mean and median second and third tracking

error measures are annualized by multiplying these values by the square root of 252. In addition, as in previous studies, the descriptive statistics of the tracking errors are compared for the subsamples of domestic and foreign ETFs. Furthermore, the descriptive statistics of the tracking errors and performance measures are examined when the smallest 20% ETFs, based on the average fund size per ETF, are excluded from the dataset. This is done because these are the typical ETFs that investors would select. These descriptive statistics are then compared with those of the entire sample. In addition, the mean and median annualized second and third tracking error measure of both samples are compared. To give investors more insight into the effect of fund size on the tracking errors and performance measures of ETFs, the mean and median tracking errors and performance measures are shown over the five quintiles based on fund size. Finally, the effect of fund size on tracking errors is shown graphically in a scatter plot.

4.3 Methodology hypotheses 2 to 8: factors affecting tracking errors

In section 2.3 the main factors influencing tracking errors are discussed, resulting in hypotheses two to eight. Next, the way these factors are defined and measured is included in the data section. To test the significance of these variables in explaining tracking errors, and thus test hypotheses two to eight, single and multiple regressions are performed. Since the spread-to-price ratios are only available from 2014 onwards, this variable is initially excluded from the analysis. However, later on the variable is added to the analysis, as described in section 4.3.2. For the entire analysis, a significance level of 5% is assumed.

4.3.1 Methodology hypotheses 2 to 7

Before regression analysis can be performed, it must first be tested whether all time-series variables are stationary. This is done using the Fisher-type unit-root test based on augmented Dickey-Fuller tests, which does not require the data to be strongly balanced. For this test the number of lags used to remove the higher-order autoregressive components of the series is equal to 1, because of the limited number of annual observations per panel. Moreover, this test is performed for each time-series variable both with and without the inclusion of a drift term. The null hypothesis of this test says that all panels contain unit roots, and the alternative hypothesis is that at least one panel is stationary. Regression analysis can be performed without further adjustments to the data if this test is significant for all time-series variables.

When all time-series variables turn out to be stationary, the tracking errors are regressed on all independent variables together, excluding the spread-to-price ratio, but including the control variables stated in the data section. For this multiple regression analysis, a random-effects GLS regression model with time-fixed effects and robust clustered standard errors is applied. As time fixed

effects are added to the regression model, the dummy variables that represent the financial crisis and COVID-19 crisis are omitted from the regression. Clustered standard errors are specified to account for heteroskedasticity and autocorrelation. For all three tracking error measures a multiple regression is performed according to the following model equation:

$$TE_{m,i,t} = \alpha + \beta_1 \cdot NER_{i,t} + \beta_2 \cdot Index\ return\ volatility_{i,t} + \beta_3 \cdot Dividends_{i,t} + \beta_4 \cdot Volumes_{i,t} + \beta_5 \cdot Fund\ size_{i,t} + \beta_6 \cdot Full\ replication_{i,t} + \beta_7 \cdot Foreign_i + \beta_8 \cdot Sector_i + \beta_9 \cdot Small\ cap_{i,t} + \beta_{10} \cdot Mid\ cap_{i,t} + \beta_{11} \cdot Smart\ beta_i + Time\ fixed\ effects + \varepsilon_{i,t} \quad (8)$$

Subsequently, the Breusch and Pagan Lagrangian multiplier test for random effects is performed. If the null hypothesis, that variances across entities (ETFs) are equal to zero, is not rejected, then the conclusion is that the application of a random effect model is not appropriate and a simple OLS regression should be applied. If the null hypothesis is rejected, then there are significant differences between the ETFs, and the random effects model is more appropriate than OLS. Then a Wald test is performed on the year dummy variables to test whether time fixed effects should be included in the regression. If the null hypothesis, that the coefficients of all year dummy variables are jointly equal to zero, is not rejected, no time fixed effects should be included in the regression. If the null hypothesis is rejected, time fixed effects should be included in the regression. Based on these results, the most appropriate model is applied to each tracking error measure.

Subsequently, the same regressions are performed several times, but then based on the ML random effects model with log transformations to subsets of those independent variables that can be log transformed. That is, all independent variables except the dummy variables. To start with, a log transformation is applied to those variables where a log transformation clearly results in a more linear relationship between the variable and the tracking error measure. This is determined based on scatterplots that show the relationship between the (log transformed) variable and the tracking error measure. Each time an additional variable is log transformed and the influence of this transformation on the Akaike's Information Criteria (AIC) and Bayesian Information Criteria (BIC) is examined. The lower the values of these criteria, the better the fit of the model. For every tracking error measure the model with the lowest value for the AIC and BIC is selected as the optimal model. Subsequently, the specification of these optimal models is applied to the random-effects GLS regression model including time-fixed effects and robust clustered standard errors.

The variance inflation factor (VIF) is then calculated for all independent variables in the model. This measure reflects the degree of multicollinearity in the multiple regression model, which is the degree of correlation between the independent variables. Because of multicollinearity the coefficients of the correlated independent variables may be sensitive to small changes in the model, which reduces

the precision of the coefficients and the statistical power of the model. VIFs of five or higher indicate high correlation between the predictors, and are therefore seen as a potential cause for concern. If and only if the coefficients of the variables with the highest VIF, provided that these VIFs are higher than five, are insignificant, the effect on the AIC and BIC of omitting the variable with the highest VIF is considered. In the end, this variable is only removed from the regression model if this results in a lower value of the AIC and BIC, and thus in a better fit of the model. Conclusions about the effect of the independent variables on the tracking error measures are then drawn based on the multiple regression results of the selected optimal models of all three tracking error measures.

However, due to potentially high correlations between the independent variables, it is also interesting to perform single regressions. The results of these regressions provide information about the effect of an individual variable on the tracking errors, without taking into account the influence of any related variables. Therefore, single regressions are performed for all three tracking error measures on all individual independent variables, except for the spread-to-price ratio, over the entire sample period considered. Similar to the multiple regression analysis, a random-effects GLS regression model with time-fixed effects and robust clustered standard errors is applied. The tracking error measures are regressed on the individual independent variables, and where possible also on their log transformed form, on an annual basis for the entire sample of ETFs, according to the following model:

$$TE_{m,i,t} = \alpha + \beta \cdot X_{i,t} + Time\ fixed\ effects + \varepsilon_{i,t} \quad (9)$$

Where $TE_{m,i,t}$ is the tracking error of ETF i in year t for measurement method m, which has a range of 1 to 3; α is the amount of tracking error that is not related to independent variable X; β is the amount of tracking error that is related to independent variable X; $X_{i,t}$ is the value of independent variable X for ETF i in year t and $\varepsilon_{i,t}$ are the residuals of the regressions.

4.3.2 Methodology hypothesis 8

Finally, the methodology as described in above section 4.3.1 is applied once again, but then with the inclusion of the spread-to-price ratio over the period running from 2014 to 2020. This of course results in a slightly different model for the multiple regression, which can be presented as follows:

$$TE_{m,i,t} = \alpha + \beta_1 \cdot Spread/price\ ratio_{i,t} + \beta_2 \cdot NER_{i,t} + \beta_3 \cdot Index\ return\ volatility_{i,t} + \beta_4 \cdot Dividends_{i,t} + \beta_5 \cdot Volumes_{i,t} + \beta_6 \cdot Fund\ size_{i,t} + \beta_7 \cdot Full\ replication_i + \beta_8 \cdot$$

$$Foreign_i + \beta_9 \cdot Sector_i + \beta_{10} \cdot Small\ cap_{i,t} + \beta_{11} \cdot Mid\ cap_{i,t} + \beta_{12} \cdot Smart\ beta_i + Time\ fixed\ effects + \varepsilon_{i,t} \quad (10)$$

4.4 Methodology hypothesis 9: tracking errors in crisis times

To begin with, one-tailed one-sample t-tests are applied to the average of all three tracking error measures per year to test whether tracking errors are significantly greater than zero for each individual year in the sample period. Based on these average tracking errors per year, a pattern may already be observed in how tracking errors have evolved over time, and in which years tracking errors were highest. Also, the average tracking errors in 2008 and 2020 can be compared with the average tracking errors over the entire sample period. To test whether tracking errors are significantly higher during the financial crisis of 2008 and the COVID-19 crisis of 2020, the average tracking error measures of those years are individually compared to the average tracking errors of the remainder of the sample period. One-tailed two-sample t-tests on the differences in tracking error will show whether tracking errors are higher in times of crisis. However, before these t-tests can be applied, it must be determined for each tracking error measure whether both groups have equal variances. For this, the Levene's test and Brown-Forsythe tests are applied, which are test statistics for equality of variance that are found to be robust under nonnormality. If these test statistics show that there is a significant difference in variance between the two groups, the Welch's one-tailed two-sample t-test is applied, which relaxes the assumption of equal variances. When both groups have a similar variance, a regular one-tailed two-sample t-test is applied. Statistical significant results would confirm the ninth hypothesis that tracking errors are larger in times of crisis. When the results of the t-tests are not significant, tracking errors in times of crisis are not significantly higher compared to other years. As an addition to the analysis, the sign and significance of the coefficients of the dummy variables representing the financial crisis year 2008 and COVID-19 crisis year 2020 in the single regressions of formula nine are examined. For both variables, a positively significant coefficient is expected.

5 Results

5.1 Results hypothesis 1: tracking errors

The first hypothesis states that U.S. equity ETFs exhibit significant tracking errors over the period from January 2, 2001 to December 31, 2020. The results of the one-tailed one-sample t-tests that are applied to the average of all three tracking error measures are presented in table 3. The first tracking error measure, which is calculated as the average daily absolute difference in total return between the ETF and its benchmark index, is on average significantly bigger than zero (M = 0.408, SD = 0.375), t(8,895)

= 102.726, *p* = .000. The second tracking error measure, which measures the standard deviation of total return differences between the ETFs and their benchmark indices, is on average significantly bigger than zero (*M* = 0.586, *SD* = 0.534), *t*(8,895) = 103.604, *p* = .000. And also the third and last tracking error measure, estimated by the SER, is on average significantly bigger than zero (*M* = 0.555, *SD* = 0.491), *t*(8,895) = 106.441, *p* = .000.

Table 3 T-statistics of the tracking error measures

Variable	Obs.	Mean	T-statistic	P-value
TE1	8,896	0.408	102.726***	0.000
TE2	8,896	0.586	103.604***	0.000
TE3	8,896	0.555	106.441***	0.000

Notes: This table presents the results of the one-tailed one-sample t-tests that are applied to the average of all three tracking error measures. TE1 is the average daily absolute difference in return between the ETF and its benchmark index; TE2 measures the standard deviation of return differences between the ETF and its benchmark index; TE3 is estimated by the Standard Error of Regression (SER) resulting from the application of the following regression to each ETF-year: $TR_{i,t} = \alpha_i + \beta \cdot TR_{b,t} + \varepsilon_{i,t}$. ***p<0.001.

Based on these results, hypothesis 1, which states that U.S. equity ETFs exhibit significant tracking errors over the period from January 2, 2001 to December 31, 2020, is confirmed. From these results it can be concluded that the ETFs are generally unable to perfectly replicate the total returns of their benchmark indices over the period from 2001 to 2020.

5.1.1 Interpretation size of tracking errors and performance measures

Besides the conclusion that these U.S. equity ETFs exhibit significant tracking errors, the size of these tracking errors is also substantial. For example, the annualized second tracking error has an average of 9.31% and a median of 7.06%. The annualized third tracking error measure is slightly lower with an average of 8.80% and a median of 6.84%. The tracking errors found are lower than those from the study by Buetow and Henderson (2012). For example, the median of the first tracking error measure, which is calculated as the average daily absolute difference in total return between the ETF and its benchmark index, is equal to 0.30% versus a 0.40% found by Buetow and Henderson (2012). And the median second tracking error measure, which measures the standard deviation of total return differences between the ETFs and their benchmark indices, is equal to 0.44% against the 0.59% found by Buetow and Henderson (2012). Also, the mean second tracking error measure, which is equal to 0.59%, is lower than that found by Rompotis (2011) for his sample of 50 iShares ETFs over the period from 2002 to 2007, which is equal to 0.63%. However, the median of tracking error two, which is equal to 0.44%, is higher than that found by Rompotis (2011), which is equal to 0.35%. In addition, all three

average tracking error measures are considerably higher than those found by Shin and Soydemir (2010). They find an average first, second and third tracking error measure of 0.062%, 0.134% and 0.133% respectively. This difference in result might be explained by the fact that the study of Shin and Soydemir (2010) was only conducted on a limited sample of 26 ETFs and only over a 3-year period, from July 2004 to June 2007, in which no substantial crisis occurred. An alternative explanation is the fact that our tracking errors have fat tails to the right side of their distributions.

Moreover, Buetow and Henderson (2012), as well as Svetina and Wahal (2008), divide their sample of ETFs listed on U.S. exchanges in a subsample of ETFs invested in U.S. securities and a subsample of ETFs invested in non-U.S. securities. The tracking errors of our subsamples of domestic and foreign ETFs and those of Buetow and Henderson (2012) and Svetina and Wahal (2008) are presented in table 4. The results of this study show that tracking errors of ETFs tracking a foreign index are more than four times higher than tracking errors of ETFs that are invested in U.S. securities. Besides, tracking errors of both the domestic and foreign ETFs resulting from our study are considerably lower than those resulting from the studies of Buetow and Henderson (2012) and Svetina and Wahal (2008).

Furthermore, the descriptive statistics of the tracking errors and performance measures are examined when the smallest 20% ETFs, based on the average fund size per ETF, are excluded from the dataset. When these descriptive statistics, presented in table 5a, are compared with the descriptive statistics of the entire dataset, as shown again in table 5b, it appears that the mean and in particular the median tracking errors are considerably lower. Similarly, the mean and median annualized second tracking error measure decreased from 9.31% to 8.03% and from 7.06% to 5.35% respectively. The same applies to the mean and median annualized third tracking error, which decreased from 8.80% to 7.65% and from 6.84% to 5.20% respectively. In addition, the average annualized relative performance increased from -0.0043% to 0.0154%. However, as mentioned earlier, the annualized relative performance and annualized alpha are strongly affected by outliers, making their median more informative. The decrease in the median annualized alpha indicates that instead of an outperformance of 0.0508%, there is an underperformance of 0.0165%. Lastly, the mean and median beta, R-squared and adjusted R-squared have increased.

Table 4 Comparison of tracking errors domestic versus foreign ETFs

	TE1		**TE2**		**TE3**	
	Domestic	Foreign	Domestic	Foreign	Domestic	Foreign
This study	0.11%	0.55%	0.17%	0.77%	0.17%	0.74%
Buetow and Henderson (2012)	0.24%	0.88%	0.35%	1.19%		
Svetina and Wahal (2009)			0.47%	1.13%		

Notes table 4: TE1 is the average daily absolute difference in return between the ETF and its benchmark index; TE2 measures the standard deviation of return differences between the ETF and its benchmark index; TE3 is estimated by the Standard Error of Regression (SER) resulting from the application of the following regression to each ETF-year: $TR_{i,t} = \alpha_i + \beta \cdot TR_{b,t} + \varepsilon_{i,t}$.

Table 5a Descriptive statistics of tracking errors and other performance measures, excluding 20% smallest ETFs based on the average fund size per ETF

Variable	Obs.	Mean	Median	Std. Dev.	Min	Max	Skewness	Kurtosis
TE1	7,111	0.3517	0.2258	0.3606	0.0203	4.0376	1.9935	9.4458
TE2	7,111	0.5057	0.3368	0.5127	0.0260	5.4030	1.9717	9.0504
TE3	7,111	0.4822	0.3276	0.4767	0.0253	5.2240	1.8564	8.3333
RP	7,111	0.0154	-0.2219	3.4317	-41.6488	80.3233	0.6238	69.8280
Alpha	7,111	0.5815	-0.0165	6.2042	-103.9464	78.4641	-1.1947	44.2384
Beta	7,111	0.9082	0.9760	0.1899	-0.2786	1.7341	-2.9035	13.6444
R-squared	7,111	0.7826	0.9174	0.2603	0.0000	0.9995	-1.2049	3.5557
Adjusted R-squared	7,111	0.7816	0.9170	0.2614	-0.0168	0.9995	-1.2054	3.5574

Table 5b Descriptive statistics of tracking errors and other performance measures

Variable	Obs.	Mean	Median	Std. Dev.	Min	Max	Skewness	Kurtosis
TE1	8,896	0.4082	0.3047	0.3748	0.0203	4.0376	1.6641	7.8904
TE2	8,896	0.5864	0.4449	0.5338	0.0260	5.4030	1.6415	7.5057
TE3	8,896	0.5546	0.4308	0.4915	0.0253	5.2240	1.5132	6.7007
RP	8,896	-0.0043	-0.2185	3.8142	-68.1904	80.3233	-0.8608	61.6070
Alpha	8,896	0.7771	0.0508	6.8185	-103.9464	97.5766	-0.6388	38.6769
Beta	8,896	0.8761	0.9611	0.2135	-0.2786	1.7341	-2.1817	8.4924
R-squared	8,896	0.7261	0.8323	0.2881	0.0000	0.9995	-0.8772	2.6580
Adjusted R-squared	8,896	0.7249	0.8316	0.2894	-0.0168	0.9995	-0.8781	2.6606

Notes table 5a and 5b: Table 5a presents the descriptive statistics of the three tracking error measures, the relative performance (RP) and the alpha, beta, R-squared and adjusted R-squared resulting from the regressions of formula 6, excluding the 20% smallest ETFs based on the average fund size per ETF. Table 5b presents the same thing, but for the entire dataset. All variables are calculated per ETF-year. TE1 is the average daily absolute difference in return between the ETF and its benchmark index; TE2 measures the standard deviation of return differences between the ETF and its benchmark index; TE3 is estimated by the Standard Error of Regression (SER) resulting from the application of the following regression to each ETF-year: $TR_{i,t} = \alpha_i + \beta \cdot TR_{b,t} + \varepsilon_{i,t}$; RP, the relative performance, is the annualized average daily difference in return between the ETF and its benchmark index; Alpha is an annualized measure that represents the return an investor can get on the ETF that is not related to its benchmark index; Beta reflects the degree to which the ETF total return replicates its benchmark index total return; R-squared and adjusted R-squared indicate the extent to which the ETF total returns can be explained by the benchmark index total returns.

Subsequently, table 6 shows the mean and median tracking errors and performance measures per quintile based on fund size. This clearly shows that the mean and especially the median tracking errors decrease as the fund size increases. This finding is supported by the scatter plot presented in figure 1, which shows that the second tracking error measure decreases with fund size. The same applies to the first and third tracking error measure, but because of redundancy these graphs are not presented. The median relative performance decreases sharply from the first to the second quintile, after which it gradually increases. The median relative performance increases the most from the fourth to the fifth quintile, still resulting in an underperformance of 0.1584% for the 20% largest ETFs. The median alpha decreases steadily across the quintiles, with a small increase from the fourth to the fifth quintile. The difference in the median alpha between the 20% smallest and 20% largest ETFs is equal to 1.62%, with an outperformance of 1.54% for the 20% smallest ETFs and an underperformance of 0.08% for the 20% largest ETFs. Finally, the mean and median beta, R-squared and adjusted R-squared generally increase across the quintiles, with the strongest increase observed from the first to the second quintile.

Table 6 Mean and median tracking errors and performance measures per quintile based on fund size

	Quintile (fund size)	TE1	TE2	TE3	RP	Alpha	Beta	R-squared	Adjusted R-squared
Mean	1	0.6775	0.9777	0.9014	-0.2544	1.5890	0.7193	0.4742	0.4717
	2	0.4447	0.6343	0.6092	-0.0617	0.5673	0.8899	0.7238	0.7226
	3	0.3465	0.4988	0.4798	0.0607	0.6280	0.9218	0.7953	0.7945
	4	0.3117	0.4480	0.4268	0.0572	0.5473	0.9177	0.8059	0.8051
	5	0.2606	0.3729	0.3557	0.1771	0.5532	0.9316	0.8312	0.8305
Median	1	0.6234	0.8942	0.8151	-0.1912	1.5432	0.7831	0.4795	0.4773

2	0.3622	0.5242	0.5070	-0.2761	0.3240	0.9324	0.7864	0.7856
3	0.2198	0.3388	0.3310	-0.2759	0.0432	0.9712	0.9162	0.9158
4	0.1384	0.2068	0.2037	-0.2720	-0.1277	0.9846	0.9688	0.9686
5	0.0745	0.1253	0.1224	-0.1584	-0.0794	0.9909	0.9917	0.9917

Notes: This table presents the mean and median of the three tracking error measures, the relative performance (RP) and the alpha, beta, R-squared and adjusted R-squared resulting from the regressions of formula 6, per quintile based on fund size. All variables are calculated per ETF-year. TE1 is the average daily absolute difference in return between the ETF and its benchmark index; TE2 measures the standard deviation of return differences between the ETF and its benchmark index; TE3 is estimated by the Standard Error of Regression (SER) resulting from the application of the following regression to each ETF-year: $TR_{i,t} = \alpha_i + \beta \cdot TR_{b,t} + \varepsilon_{i,t}$; RP, the relative performance, is the annualized average daily difference in return between the ETF and its benchmark index; Alpha is an annualized measure that represents the return an investor can get on the ETF that is not related to its benchmark index; Beta reflects the degree to which the ETF total return replicates its benchmark index total return; R-squared and adjusted R-squared indicate the extent to which the ETF total returns can be explained by the benchmark index total returns.

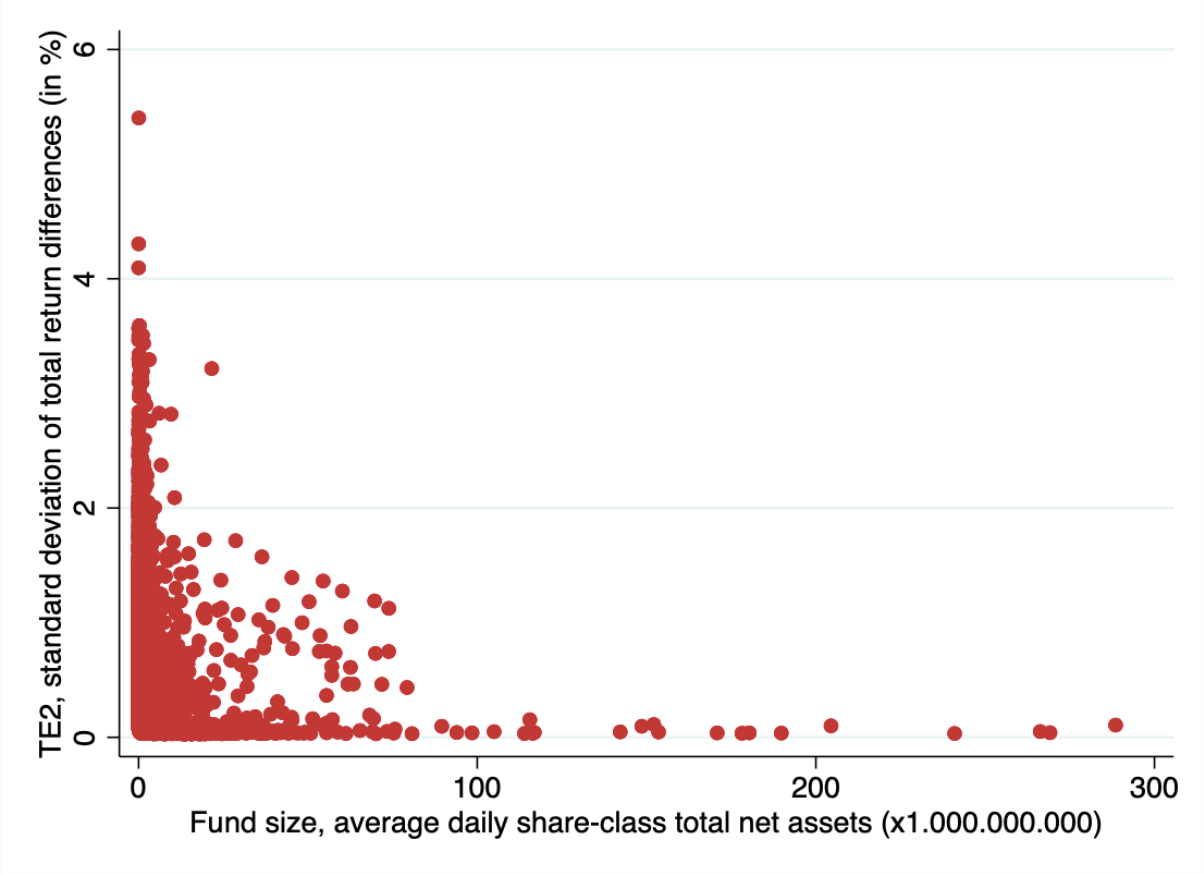

Figure 1 Scatterplot TE2 against fund size

From these findings it appears that tracking errors of domestic ETFs are lower than those of foreign ETFs. In addition, it is found that tracking errors of U.S. equity ETFs decrease with fund size and that tracking errors of the 20% smallest ETFs are considerably higher compared to the rest of the ETFs. Based on the descriptive statistics of the relative performance and alpha, no unequivocal conclusion can be drawn about the effect of fund size on the performance of an ETF. The median relative performance is highest for the 20% largest ETFs, while the median alpha is highest for the 20% smallest ETFs. However, the beta, R-squared and adjusted R-squared do increase with fund size. These results imply that the tracking efficiency of U.S. equity ETFs increases with fund size.

5.2 Results hypotheses 2 to 8: factors affecting tracking errors

5.2.1 Results hypotheses 2 to 7

In this section, hypotheses two to seven will be tested and answered. These hypotheses are as follows:

Hypothesis 2: The expense ratio has a positive effect on tracking errors of U.S. equity ETFs.
Hypothesis 3: The volatility of benchmark index total returns has a positive effect on tracking errors of U.S. equity ETFs.
Hypothesis 4: Dividends have a positive effect on tracking errors of U.S. equity ETFs.
Hypothesis 5: The trading volume has a negative effect on tracking errors of U.S. equity ETFs.
Hypothesis 6: The fund size has a negative effect on tracking errors of U.S. equity ETFs.
Hypothesis 7: The application of a full replication strategy has a negative effect on tracking errors of U.S. equity ETFs.

The results of the Fisher-type unit root tests based on augmented Dickey-Fuller tests, displayed in table B1 of appendix B, show that all time-series variables are stationary. Therefore, regression analysis can be applied without further adjustments to the data. The results of the multiple regressions based on the random-effects GLS regression model with time-fixed effects and robust clustered standard errors, including all independent variables except the spread-to-price ratio, are shown in table B2 of appendix B. The results of the Breusch and Pagan Lagrangian multiplier tests for random effects are shown in table 7, and are significant for all three tracking error measures. This means that there are significant differences between the ETFs, and therefore the random effects model is more appropriate than OLS. Subsequently, the results of the Wald tests on the year dummy variables are shown in table 8. The null hypothesis, that the coefficients of all year dummy variables are jointly equal to zero, is rejected for all three tracking error measures, which means that time fixed effects should be included in the regression analysis.

Table 7 Results of the Breusch and Pagan Lagrangian multiplier tests for random effects

	$\overline{\chi}^2$**(1, *N* = 8,896)**	**p**
TE1	7,812.36***	0.000
TE2	6,875.23***	0.000
TE3	6,973.82***	0.000

Notes: ***p<0.001.

Table 8 Results of the Wald tests for time fixed effects

	χ^2**(19, *N* = 8,896)**	**p**
TE1	1,141.86***	0.000
TE2	1,256.26***	0.000
TE3	1,346.39***	0.000

Notes: ***p<0.001.

The random effects model with time fixed effects is therefore the most appropriate model specification for all three tracking error measures. Next, the ML random effects model is used to determine whether log transformations to particular variables can further improve the fit of the model. Tables B3a to B3c of appendix B show the values of the model selection criteria, AIC and BIC, for models in which log transformations are applied to various variables. For each tracking error measure, the model with the lowest value for the AIC and BIC is selected as the optimal model. For the first tracking error measure, this results in a random-effects GLS regression model with time-fixed effects, robust clustered standard errors and log transformations to the variables volatility, volumes and fund size. For the second and third tracking error measure, this results in the same model specification with the addition of a log transformation to the net expense ratio. Subsequently, the variance inflation factors of the independent variables are shown in table B4 of appendix B. None of the VIFs is greater than five, indicating only moderate correlation between the independent variables. Thus, for all three tracking error measures, the optimal model remains intact, and no variables are omitted from the regression model. The results of the multiple regression analysis, on the basis of which hypotheses two to seven are tested, are presented in table 9.

Table 9 Multiple regression results - random-effects GLS model with time fixed effects and clustered standard errors

Variable	**TE1**	**TE2**	**TE3**
NER	0.048		
	(0.030)		

Log NER		0.012	0.017
		(0.012)	(0.011)
Log volatility	0.058*	0.096**	0.071*
	(0.025)	(0.037)	(0.031)
Dividends	0.009	0.012	0.011
	(0.005)	(0.007)	(0.006)
Log volumes	-0.005	-0.004	-0.001
	(0.004)	(0.006)	(0.005)
Log fund size	-0.048***	-0.076***	-0.066***
	(0.005)	(0.007)	(0.006)
Full replication	-0.028*	-0.035*	-0.031*
	(0.012)	(0.017)	(0.015)
Foreign	0.305***	0.403***	0.406***
	(0.014)	(0.020)	(0.018)
Sector	-0.105***	-0.150***	-0.133***
	(0.014)	(0.020)	(0.018)
Small cap	0.034	0.049	0.054*
	(0.020)	(0.027)	(0.023)
Mid cap	-0.011	-0.016	-0.007
	(0.012)	(0.016)	(0.014)
Smart beta	-0.038**	-0.049**	-0.048**
	(0.012)	(0.018)	(0.016)
2002	-0.073	-0.157	0.085
	(0.189)	(0.345)	(0.157)
2003	-0.142	-0.291	-0.057
	(0.189)	(0.345)	(0.161)
2004	-0.241	-0.436	-0.216
	(0.188)	(0.345)	(0.154)
2005	-0.311	-0.502	-0.283
	(0.188)	(0.344)	(0.155)
2006	-0.282	-0.461	-0.238
	(0.188)	(0.344)	(0.154)
2007	-0.196	-0.335	-0.100
	(0.187)	(0.342)	(0.152)
2008	0.261	0.372	0.558***
	(0.185)	(0.338)	(0.152)
2009	-0.011	-0.096	0.127
	(0.185)	(0.339)	(0.150)

2010	-0.265	-0.440	-0.197
	(0.186)	(0.341)	(0.151)
2011	-0.211	-0.344	-0.100
	(0.186)	(0.341)	(0.151)
2012	-0.355	-0.574	-0.331*
	(0.188)	(0.343)	(0.153)
2013	-0.371*	-0.571	-0.335*
	(0.188)	(0.344)	(0.154)
2014	-0.394*	-0.619	-0.377*
	(0.188)	(0.344)	(0.154)
2015	-0.333	-0.543	-0.299
	(0.187)	(0.343)	(0.153)
2016	-0.329	-0.531	-0.299
	(0.187)	(0.343)	(0.152)
2017	-0.440*	-0.682*	-0.443**
	(0.190)	(0.346)	(0.156)
2018	-0.336	-0.529	-0.290
	(0.187)	(0.342)	(0.152)
2019	-0.460*	-0.721*	-0.469**
	(0.188)	(0.343)	(0.153)
2020	-0.301	-0.452	-0.183
	(0.185)	(0.339)	(0.150)
Constant	0.485*	0.823*	0.582***
	(0.206)	(0.371)	(0.176)
Observations	8,896	8,876	8,876
Overall R-squared	0.568	0.559	0.590

Notes: This table presents the results of the multiple random-effects GLS regressions with time fixed effects and clustered standard errors. The regressions are performed for all three tracking error measures, which are TE1, TE2 and TE3. TE1 is the average daily absolute difference in return between the ETF and its benchmark index; TE2 measures the standard deviation of return differences between the ETF and its benchmark index; TE3 is estimated by the Standard Error of Regression (SER) resulting from the application of the following regression to each ETF-year: $TR_{i,t} = \alpha_i + \beta \cdot TR_{b,t} + \varepsilon_{i,t}$. Standard errors are shown in parentheses. * p<0.05; ** p<0.01; ***p<0.001.

To start with, no significant effect is found of the (log transformed) net expense ratio, dividends and log transformed volumes on any of the tracking errors. The log transformed volatility has a positive significant effect on all three tracking error measures. The most positive effect is found for the second

tracking error measure, β = 0.096, p = .009. This is followed by the positive effect of the log transformed volatility on tracking error three, β = 0.071, p = .021. Finally, the least positive effect was found for the first tracking error measure, β = 0.058, p = .022. Starting from an average volatility of 2.2201%, a one percentage point increase in volatility is associated with an increase of 0.035%, 0.057% and 0.043% in tracking error one, two and three respectively. The log transformed fund size, on the other hand, has a significant negative effect on all three tracking error measures. The biggest effect is again found for the second tracking error, β = -0.076, p = .000, followed by the third tracking error measure, β = -0.066, p = .000. The negative effect of the log transformed fund size is smallest for tracking error one, β = -0.048, p = .000. A 1% increase in fund size results in a decrease of 0.00048%, 0.00076% and 0.00066% in tracking error measure one, two and three respectively. Also, a significant negative effect of the variable full replication on tracking errors is found. For ETFs with a full replication strategy tracking errors one, two and three are 0.028%, 0.035% and 0.031% lower respectively. Besides, significantly higher tracking errors are found for ETFs that track a foreign index. The largest positive effect is found for the third tracking error measure, β = 0.406, p = .000, followed by tracking error two, β = 0.403, p = .000. But also for the first tracking error measure a significant positive effect is found for the variable foreign, β = 0.305, p = .000. Tracking a foreign index is thus associated with a 0.305%, 0.403% and 0.406% increase in tracking errors one to three respectively. Having a sector orientation, on the other hand, results in significantly lower tracking errors, with β = -0.105, p = .000 for TE1, β = -0.150, p = .000 for TE2 and β = -0.133, p = .000 for TE3. Furthermore, for ETFs that have a focus on small cap investments, the third tracking error is 0.054% higher, β = 0.054, p = .020. Finally, following a smart beta strategy is associated with significantly lower tracking errors, with β = -0.038, p = .002 for TE1, β = -0.049, p = .007 for TE2 and β = -0.048, p = .002 for TE3.

Due to the correlations between the independent variables, as presented in table B5 of appendix B, also single regression analysis is applied. The results of these single regressions, according to the random-effects GLS regression model with time-fixed effects and robust clustered standard errors, are shown in table 10.

Table 10 Single regression results - random-effects GLS model with time fixed effects and clustered standard errors

	TE1		**TE2**		**TE3**	
Variable	**Coefficient**	**R-squared**	**Coefficient**	**R-squared**	**Coefficient**	**R-squared**
NER	0.260***	0.236	0.371***	0.253	0.357***	0.266
Log NER	0.074***	0.219	0.108***	0.237	0.099***	0.247
Volatility	-0.008	0.173	-0.002	0.193	-0.019	0.197
Log volatility	0.027	0.173	0.053	0.193	0.031	0.197

Dividends	-0.007	0.175	-0.014	0.196	-0.011	0.199
Log dividends	-0.016**	0.174	-0.027***	0.196	-0.022**	0.199
Volumes	-0.006***	0.173	-0.008***	0.194	-0.007***	0.197
Log volumes	-0.037***	0.250	-0.055***	0.275	-0.047***	0.271
Fund size	0.000	0.174	0.000	0.195	0.000	0.198
Log fund size	-0.058***	0.311	-0.087***	0.338	-0.075***	0.336
Full replication	-0.124***	0.188	-0.163***	0.205	-0.157***	0.211
Financial crisis (2008)	0.194	0.173	0.262	0.193	0.453*	0.197
Covid19 crisis (2020)	-0.473*	0.173	-0.726	0.193	-0.429*	0.197
Foreign	0.385***	0.485	0.513***	0.467	0.503***	0.507
Sector	-0.128***	0.206	-0.180***	0.225	-0.162***	0.228
Small cap	0.003	0.175	0.011	0.195	0.014	0.198
Mid cap	-0.033*	0.175	-0.045*	0.195	-0.037*	0.198
Smart beta	-0.118***	0.188	-0.151***	0.204	-0.153***	0.211

Notes: This table presents the coefficients and overall R-squared values of the single random-effects GLS regressions with time fixed effects and clustered standard errors. The regressions are performed for all three tracking error measures, which are TE1, TE2 and TE3. TE1 is the average daily absolute difference in return between the ETF and its benchmark index; TE2 measures the standard deviation of return differences between the ETF and its benchmark index; TE3 is estimated by the Standard Error of Regression (SER) resulting from the application of the following regression to each ETF-year: $TR_{i,t} = \alpha_i + \beta \cdot TR_{b,t} + \varepsilon_{i,t}$. * p<0.05; ** p<0.01; ***p<0.001.

Based on the results of the multiple regression analysis, hypotheses two to seven are tested. Table 11 provides a clear overview of the results for these hypothesis tests. No significant effect is found of the net expense ratio on tracking errors. Therefore, hypothesis 2, which states that the expense ratio has a positive effect on tracking errors of U.S. equity ETFs, is not supported. A positive significant effect of the volatility of benchmark index total returns on tracking errors is found, so that the third hypothesis is supported. No significant effect of dividends on tracking errors is found, which means no support for hypothesis four is found. The same applies to the fifth hypothesis, which states that trading volume has a significant negative effect on tracking errors of U.S. equity ETFs. No significant effect of traded volumes on tracking errors is found. On the other hand, a significant negative effect is found of the log transformed fund size on all three tracking error measures. Thus, the sixth hypothesis is supported. Finally, the results show that the application of a full replication strategy results in significantly lower tracking errors. Therefore, hypothesis seven is supported.

Table 11 Overview of answers to hypotheses two to seven

	Hypothesis	Answer
2	The expense ratio has a positive effect on tracking errors of U.S. equity ETFs.	Not supported
3	The volatility of benchmark index total returns has a positive effect on tracking errors of U.S. equity ETFs.	Supported
4	Dividends have a positive effect on tracking errors of U.S. equity ETFs.	Not supported
5	The trading volume has a negative effect on tracking errors of U.S. equity ETFs.	Not supported
6	The fund size has a negative effect on tracking errors of U.S. equity ETFs.	Supported
7	The application of a full replication strategy have a negative effect on tracking errors of U.S. equity ETFs.	Supported

5.2.2 Results hypothesis 8

In this section, the eighth hypothesis, which states that the bid-ask spread has a positive effect on tracking errors of U.S. equity ETFs, is tested.

The results of the Fisher-type unit root tests based on augmented Dickey-Fuller tests, displayed in table C1 of appendix C, show that all time-series variables are stationary. Therefore, regression analysis can be applied without further adjustments to the data. The results of the multiple regressions based on the random-effects GLS regression model with time-fixed effects and robust clustered standard errors including all independent variables are shown in table C2 of appendix C. The results of the Breusch and Pagan Lagrangian multiplier tests for random effects are shown in table 12, and are significant for all three tracking errors. This means that there are significant differences between the ETFs, and therefore the random effects model is more appropriate than OLS. Subsequently, the results of the Wald tests on the year dummy variables are shown in table 13. The null hypothesis, that the coefficients of all year dummy variables are jointly equal to zero, is rejected for all three tracking error measures, which means that time fixed effects should be included in the regression analysis.

Table 12 Results of the Breusch and Pagan Lagrangian multiplier tests for random effects

	$\overline{\chi}^2$**(1, *N* = 5,648)**	**p**
TE1	3,324.01***	0.000
TE2	2,813.53***	0.000
TE3	2,909.21***	0.000

Notes: ***p<0.001.

Table 13 Results of the Wald tests for time fixed effects

	χ^2(6, *N* = 5,648)	p
TE1	683.89***	0.000
TE2	650.66***	0.000
TE3	683.89***	0.000

Notes: ***p<0.001.

Thus, the random effects model with time fixed effects is the most appropriate model specification for all three tracking error measures. Next, the ML random effects model is used to determine whether log transformations to particular variables can further improve the fit of the model. Tables C3a to C3c of appendix C show the values of the model selection criteria, AIC and BIC, for models in which log transformations are applied to various variables. For each tracking error measure, the model with the lowest value for the AIC and BIC is selected as the optimal model. For the first tracking error measure, this results in a random-effects GLS regression model with time-fixed effects, robust clustered standard errors and log transformations to the variables volatility, volumes, fund size and dividends. For the second and third tracking error measure, this results in the same model specification with the addition of a log transformation to the net expense ratio. Subsequently, the variance inflation factors of the independent variables are presented in table C4 of appendix C. The log transformed volumes and log transformed fund size have VIFs that are greater than five, indicating high correlation between these predictors. But since the coefficient of the log transformed fund size is significant, neither variable is omitted from the regression model. Therefore, for all three tracking error measures, the optimal model remains intact, and no variables are omitted from the regressions. The results of the multiple regression analysis, on the basis of which the eighth hypothesis is tested, are shown in table 14.

Table 14 Multiple regression results - random-effects GLS model with time fixed effects and clustered standard errors

Variable	**TE1**	**TE2**	**TE3**
Spread/price ratio	0.079**	0.110**	0.114**
	(0.027)	(0.039)	(0.040)
NER	0.082**		
	(0.031)		
Log NER		0.010	0.019*
		(0.011)	(0.009)
Log volatility	0.054***	0.075***	0.041*
	(0.015)	(0.022)	(0.020)
Log dividends	-0.025***	-0.034***	-0.031***

	(0.005)	(0.008)	(0.007)
Log volumes	-0.000	0.001	0.010
	(0.004)	(0.006)	(0.006)
Log fund size	-0.033***	-0.053***	-0.051***
	(0.005)	(0.007)	(0.006)
Full replication	-0.036***	-0.044**	-0.035**
	(0.011)	(0.015)	(0.013)
Foreign	0.266***	0.365***	0.357***
	(0.013)	(0.018)	(0.016)
Sector	-0.123***	-0.168***	-0.153***
	(0.013)	(0.019)	(0.017)
Small cap	0.009	0.014	0.024
	(0.014)	(0.020)	(0.018)
Mid cap	-0.014	-0.022	-0.015
	(0.009)	(0.013)	(0.012)
Smart beta	-0.038**	-0.049**	-0.047**
	(0.012)	(0.017)	(0.015)
2015	0.057***	0.074***	0.079***
	(0.006)	(0.009)	(0.008)
2016	0.060***	0.086***	0.079***
	(0.006)	(0.010)	(0.009)
2017	-0.052***	-0.077***	-0.078***
	(0.006)	(0.009)	(0.009)
2018	0.051***	0.082***	0.084***
	(0.006)	(0.010)	(0.009)
2019	-0.062***	-0.097***	-0.084***
	(0.005)	(0.008)	(0.007)
2020	0.085***	0.170***	0.207***
	(0.015)	(0.023)	(0.022)
Constant	0.154***	0.276***	0.337***
	(0.038)	(0.062)	(0.054)
Observations	5,647	5,627	5,627
Overall R-squared	0.612	0.593	0.618

Notes table 14: This table presents the results of the multiple random-effects GLS regressions with time fixed effects and clustered standard errors. The regressions are performed for all three tracking error measures, which are TE1, TE2 and TE3. TE1 is the average daily absolute difference in return between the ETF and its benchmark index; TE2 measures the standard deviation of return differences between the ETF and its benchmark index; TE3 is estimated by the Standard Error of Regression (SER) resulting from the application of the following regression to each ETF-year: $TR_{i,t} = \alpha_i + \beta \cdot TR_{b,t} + \varepsilon_{i,t}$. Standard errors are shown in parentheses. * p<0.05; ** p<0.01; ***p<0.001.

Based on the results presented in table 14 it can be concluded that the spread-to-price ratio has a positive significant effect on all three tracking error measures. The spread-to-price ratio has the greatest positive effect on the third tracking error measure, β = 0.114, *p* = .004. This is followed by an almost equally positive effect on the second tracking error measure, β = 0.110, *p* = .005. Finally, the least positive effect is found for the first tracking error measure, β = 0.079, *p* = .003. If the spread-to-price ratio, which is expressed in percentages, increases by 1(%), this results in an increase of 0.079%, 0.110% and 0.114% in tracking error measure one, two and three respectively.

Due to the correlations between the independent variables, as presented in table C5 of appendix C, also single regression analysis is applied. The results of these single regressions, according to the random-effects GLS regression model with time-fixed effects and robust clustered standard errors, are presented in table 15.

Table 15 Single regression results - random-effects GLS model with time fixed effects and clustered standard errors

	TE1		**TE2**		**TE3**	
Variable	Coefficient	R-squared	Coefficient	R-squared	Coefficient	R-squared
Spread-to-price ratio	0.125**	0.200	0.181**	0.208	0.177**	0.219
Log spread-to-price ratio	0.080***	0.261	0.117***	0.271	0.114***	0.281
NER	0.341***	0.170	0.492***	0.177	0.472***	0.194
Log NER	0.063***	0.131	0.092***	0.141	0.091***	0.154
Volatility	-0.043**	0.068	-0.068**	0.082	-0.091***	0.087
Log volatility	0.010	0.067	0.006	0.081	-0.023	0.086
Dividends	-0.019*	0.081	-0.027*	0.096	-0.028**	0.102
Log dividends	-0.042***	0.090	-0.060***	0.106	-0.056***	0.110
Volumes	-0.929	0.068	-2.085	0.082	-1.844	0.087
Log volumes	-0.038***	0.158	-0.057***	0.179	-0.048***	0.170
Fund size	0.000	0.071	-0.001*	0.087	-0.001**	0.093
Log fund size	-0.051***	0.259	-0.076***	0.280	-0.066***	0.270
Full replication	-0.119***	0.093	-0.158***	0.103	-0.149***	0.109

Covid19 crisis (2020)	0.104***	0.067	0.193***	0.081	0.205***	0.086
Foreign	0.367***	0.471	0.497***	0.447	0.482***	0.480
Sector	-0.151***	0.129	-0.214***	0.141	-0.195***	0.145
Small cap	0.000	0.071	-0.004	0.086	0.002	0.090
Mid cap	-0.021	0.071	-0.039*	0.086	-0.033*	0.090
Smart beta	-0.109***	0.091	-0.142***	0.100	-0.141***	0.108

Notes: This table presents the coefficients and overall R-squared values of the single random-effects GLS regressions with time fixed effects and clustered standard errors. The regressions are performed for all three tracking error measures, which are TE1, TE2 and TE3. TE1 is the average daily absolute difference in return between the ETF and its benchmark index; TE2 measures the standard deviation of return differences between the ETF and its benchmark index; TE3 is estimated by the Standard Error of Regression (SER) resulting from the application of the following regression to each ETF-year: $TR_{i,t} = \alpha_i + \beta \cdot TR_{b,t} + \varepsilon_{i,t}$. * p<0.05; ** p<0.01; ***p<0.001.

Based on the results of the multiple regression analysis, a significant positive effect of the spread-to-price ratio on all three tracking error measures is found. Therefore, hypothesis eight, which states that the bid-ask spread has a positive effect on tracking errors of U.S. equity ETF, is supported.

5.3 Results hypothesis 9: tracking errors in crisis times

The ninth hypothesis states that tracking errors of U.S. equity ETFs are higher during the financial crisis of 2008 and the COVID-19 crisis of 2020. I start by looking at the results of the one-tailed one-sample t-tests that are applied to the average of all three tracking error measures per year, which are presented in table 16. From these results it can be concluded that ETF tracking errors are significantly greater than zero for every year from 2002 to 2020, *p* = .000. Only in the year 2001 tracking errors one to three are not significantly greater than zero, with *p* = .196, *p* = .207 and *p* = .185 respectively. For all three tracking error measures, tracking errors are highest in 2008, which can be seen from table 16 and figure 2. However, looking at figure 2, tracking errors in 2020, which is the year of the coronavirus stock market crash, do not seem higher than in the rest of the sample period.

Table 16 T-statistics of tracking errors per year

Years	Obs.	TE1	T-statistic	TE2	T-statistic	TE3	T-statistic
2001	2	0.679	1.417	1.077	1.315	0.748	1.527
2002	42	0.519	10.047***	0.767	10.816***	0.730	11.995***
2003	66	0.554	10.226***	0.766	9.589***	0.723	9.838***
2004	87	0.438	14.818***	0.596	14.616***	0.538	16.424***

2005	106	0.319	14.313***	0.457	14.049***	0.407	16.015***
2006	135	0.332	15.180***	0.477	15.255***	0.433	16.490***
2007	209	0.463	18.416***	0.667	20.034***	0.630	20.431***
2008	285	1.029	23.998***	1.544	25.748***	1.433	26.538***
2009	320	0.759	24.296***	1.069	25.294***	1.002	26.433***
2010	389	0.494	25.861***	0.702	25.326***	0.666	26.111***
2011	495	0.559	29.134***	0.818	30.495***	0.776	31.576***
2012	554	0.412	31.780***	0.573	31.672***	0.542	32.549***
2013	558	0.356	29.482***	0.518	30.762***	0.490	30.975***
2014	625	0.324	30.651***	0.454	30.729***	0.433	31.052***
2015	705	0.405	32.425***	0.565	32.787***	0.537	33.498***
2016	816	0.415	36.184***	0.590	36.260***	0.543	37.642***
2017	826	0.241	37.615***	0.338	36.492***	0.320	37.300***
2018	893	0.366	39.306***	0.529	40.019***	0.498	41.297***
2019	913	0.227	34.395***	0.312	34.275***	0.301	35.694***
2020	870	0.421	33.041***	0.647	34.620***	0.633	35.279***

Notes: This table presents the results of the one-tailed one-sample t-tests that are applied to the average of all three tracking error measures per year. The columns labeled as TE1, TE2 and TE3 show the average tracking errors per year. Where TE1 is the average daily absolute difference in return between the ETF and its benchmark index; TE2 measures the standard deviation of return differences between the ETF and its benchmark index; TE3 is estimated by the Standard Error of Regression (SER) resulting from the application of the following regression to each ETF-year: $TR_{i,t} = \alpha_i + \beta \cdot TR_{b,t} + \varepsilon_{i,t}$. In addition, the corresponding T-statistics are reported, where ***p<0.001.

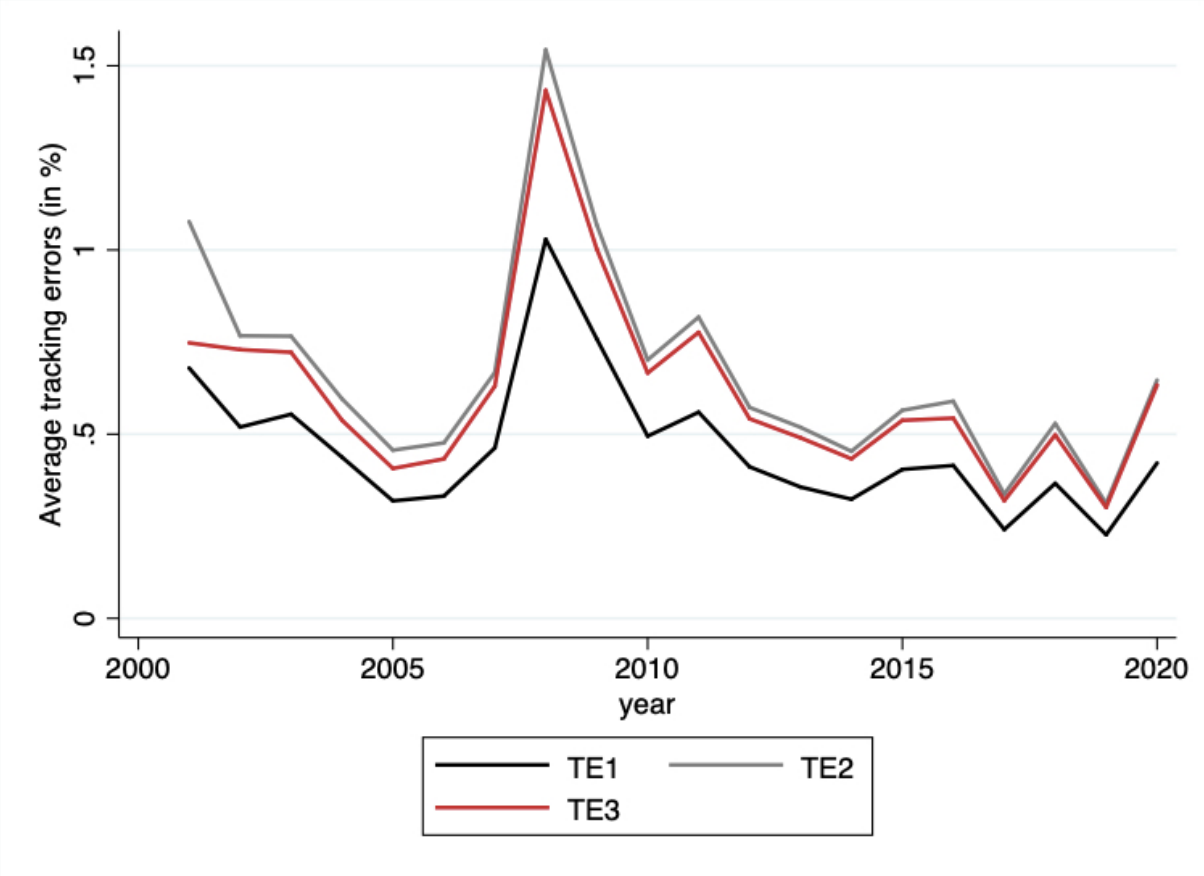

Figure 2 Average tracking errors, 2001-2020

Based on the results of the Levene's test and Brown-Forsythe tests, which are presented in table D1a and D1b of appendix D, it appears that there is a difference in variance between the defined groups for all tracking error measures for the examination of both 2008 and 2020. Therefore, for the examination of all tracking error measures for both 2008 and 2020, the one-tailed two-sample Welch's t-test is applied, of which the results are presented in table 17a and 17b respectively. The first tracking error measure, which is measured as the average daily absolute difference in total return between the ETF and its benchmark index, is on average a significant 0.641% higher in 2008 (M = 1.029, SD = 0.724) compared to the remainder of the sample period (M = 0.388, SD = 0.339), t(288.159) = 14.902, p = .000. The second tracking error measure, that is calculated as the standard deviation of return differences between the ETFs and their benchmark indices, gives a 0.989% higher average tracking error for 2008 (M = 1.543, SD = 1.012) compared to the rest of the sample period (M = 0.555, SD = 0.479), t(288.255) = 16.433, p = .000. This difference in average tracking error is statistically significant. According to the third tracking error measure, estimated by the SER, tracking errors are on average 0.908% higher during 2008 (M = 1.433, SD = 0.912) in comparison to the other years in the sample period (M = 0.526, SD = 0.442), t(288.473) = 16.743, p = .000. Also this result is statistically significant. According to the first tracking error measure tracking errors are not significantly higher in 2020 (M =

0.422, *SD* = 0.376) compared to the rest of the sample period (*M* = 0.407, *SD* = 0.375), *t*(1,064.799) = 1.106, *p* = 0.134. The second tracking error measure is on average 0.067% higher in 2020 (*M* = 0.647, *SD* = 0.551) in comparison to the other years in the sample period (*M* = 0.580, *SD* = 0.532), *t*(1,052.422) = 3.414, *p* = .000. This result is statistically significant. Eventually, the third and last tracking error measure gives a significant 0.087% higher average tracking error for 2020 (*M* = 0.633, *SD* = 0.529) compared to the remainder of the sample period (*M* = 0.546, *SD* = 0.486), *t*(1,034.816) = 4.642, *p* = .000. Thus, for all three tracking error measures it can be concluded that the tracking errors of these U.S. equity ETFs are significantly higher in 2008, the year of the financial crisis. The second and third tracking error measure indicate significantly higher tracking errors for the year 2020, the year of the coronavirus stock market crash. However, based on the first tracking error measure, tracking errors in 2020 are not significantly higher than those in other years.

Table 17a Results one-tailed two-sample Welch's t-test, 2008 compared to the rest of the sample period

	Obs (2008)	Mean (2008)	Obs (rest)	Mean (rest)	Diff.	T-statistic	P-value	Welch's df
TE1	285	1.029	8,611	0.388	0.641	14.902***	0.000	288.159
TE2	285	1.543	8,611	0.555	0.989	16.433***	0.000	288.255
TE3	285	1.433	8,611	0.526	0.908	16.743***	0.000	288.473

Table 17b Results one-tailed two-sample Welch's t-test, 2020 compared to the rest of the sample period

	Obs (2020)	Mean (2020)	Obs (rest)	Mean (rest)	Diff.	T-statistic	P-value	Welch's df
TE1	870	0.422	8,026	0.407	0.015	1.106	0.134	1,064.799
TE2	870	0.647	8,026	0.580	0.067	3.414***	0.000	1,052.422
TE3	870	0.633	8,026	0.546	0.087	4.642***	0.000	1,034.816

Notes: Table 17a presents the results of the one-tailed two-sample Welch's t-test that is applied to the differences in average tracking errors over 2008 and the remainder of the sample period. Table 17b presents the results of the one-tailed two-sample Welch's t-test that is applied to the differences in the average tracking errors over 2020 and the remainder of the sample period. The number of observations and the mean are reported for both groups. Also the differences in average TE, the corresponding T-statistics, one-sided P-values and degrees of freedom are reported. TE1 is the average daily absolute difference in return between the ETF and its benchmark index; TE2 measures the standard deviation of return differences between the ETF and its benchmark index; TE3 is estimated by the Standard Error of Regression (SER) resulting from the application of the following regression to each ETF-year: $TR_{i,t} = \alpha_i + \beta \cdot TR_{b,t} + \varepsilon_{i,t}$. ***p<0.001.

Finally, the sign and significance of the coefficients of the dummy variables representing the financial crisis year 2008 and COVID-19 crisis year 2020 in the single regressions, as shown in tables 10 and 15, are examined. No significant positive effect is found for the dummy variable representing the year

2008 on the first and second tracking error measure, with β = 0.194, p = .334 and β = 0.262, p = .483 respectively. For the third tracking error, a significant positive effect is found of the dummy variable representing the year 2008 on the tracking error measure, β = 0.453, p = .014. The results for the year 2020 are highly dependent on the sample period considered. First, the results based on the longest sample period, that is from 2001 to 2020, are examined. For the first and third tracking error measure, a negative significant effect is found for the dummy variable that represents the COVID-19 crisis year 2020, with β = -0.473, p = .020 for TE1 and β = -0.429, p = .021 for TE3. In the examination of the second tracking error measure, the coefficient of the dummy variable representing the year 2020 is not significantly different from zero, β = -0.726, p = .055. However, when looking at the results based on the sample period from 2014 to 2020, where the year 2020 is compared to the years 2014 to 2019, a positive significant effect is found for the year 2020 on tracking errors. The most positive effect is found for the third tracking error measure, β = 0.205, p = .000. This is followed by the positive effect of the dummy variable of the year 2020 on the second tracking error measure, β = 0.193, p = .000. Finally, the least positive but significant result is found for tracking error one, β = 0.104, p = .000.

Since the ninth hypothesis simply states that tracking errors from U.S. equity ETFs are higher in the years 2008 and 2020, the conclusion regarding this hypothesis is drawn based on the results of the one-tailed two-sample t-tests. Therefore, the ninth hypothesis, which states that tracking errors of U.S. equity ETFs are higher during the financial crisis of 2008 and the COVID-19 crisis of 2020, is supported.

6 Discussion

All relevant results are discussed in the results section, based on which the hypotheses are tested. This section further elaborates on these results, and whether these are in line with expectations and findings of previous studies.

To begin with, significant tracking errors were found for the sample of U.S. equity ETFs over the period running from 2001 to 2020. Based on this result, hypothesis 1, which states that U.S. equity ETFs exhibit significant tracking errors over the period from January 2, 2001 to December 31, 2020, is supported. This finding is in line with expectations and is consistent with the findings of Svetina and Wahal (2008), Shin and Soydemir (2010) and Rompotis (2011), that also find significant tracking errors for U.S. ETFs.

Subsequently, conclusions about the effect of the examined independent variables on tracking errors are drawn based on the results of the optimal multiple regression model per tracking error measure. Although the effect of some variables appears to be insignificant based on the multiple regression results, in some cases single regression analysis points out that these variables do

significantly influence tracking errors. However, due to high correlations between the independent variables, the significant effect disappears when multiple regression analysis is applied. This shows that the variable in question is indeed important for tracking errors, but that it has no additional explanatory power when other factors are also taken into account.

For example, no significant effect was found of the annual report net expense ratio on tracking errors. This result is not in line with expectations and is not consistent with the results of previous studies such as those of Rompotis (2011), Chu (2011) and Naumenko and Chystiakova (2015). However, the simple regression results, as presented in table 10, do show that the net expense ratio has a positive significant effect on all three tracking error measures, even at a significance level of 0.1%. The net expense ratio thus seems to have a positive influence on tracking errors, but has no additional explanatory power compared to the other independent variables. This may be due to the strong negative correlation between the (log transformed) net expense ratio and the log transformed fund size and the strong positive correlation between the (log transformed) net expense ratio and the variable foreign. It is probable that the lack of a significant positive effect of the (log transformed) net expense ratio on tracking errors in the multiple regression analysis is caused by this multicollinearity. Another possible explanation for the absence of a significant result is that the annual report net expense ratio is not a total cost measure. For example, brokerage costs and sales charges are not included in the ratio.

As expected, a positive significant effect of the volatility of benchmark index total returns on tracking errors is found, which means that the third hypothesis is supported. Market volatility, also explained as risk, thus results in an increase in tracking errors, which is in line with the findings of, among others, Chiang (1998), Frino and Gallagher (2002) and Rompotis (2011).

The fourth hypothesis, which states that dividends have a positive effect on tracking errors of U.S. equity ETFs, is not supported, because no significant effect of dividends on tracking errors is found. Also based on the simple regression analysis, no significant effect of dividends on tracking errors is found, and the simple regressions of the log transformed dividends on tracking errors even show a significant negative effect. These results conflict with expectations. However, although in previous studies a positive significant effect was expected to be found, it was not always found. Although Osterhoff and Kaserer (2016) do find the expected positive significant effect of dividends on tracking errors, no significant result is found in the studies by Frino and Gallagher (2002) and Shin and Soydemir (2010).

In addition, the hypothesis which states that trading volume has a negative effect on tracking errors of U.S. equity ETFs, is not supported. The multiple regressions show that there is no significant effect of traded volumes on any of the tracking error measures. This result is not in line with the expectation that a high degree of liquidity has a lowering effect on tracking errors, and is inconsistent

with the findings of Buetow and Henderson (2012), Mateus and Rahmani (2017) and Meinhardt, Mueller and Schoene (2015). However, the single regressions indicate that both the variable volumes and log transformed volumes has a significant negative effect on all three tracking error measures. These results are significant at a significance level of 0.1%. Therefore, the trading volume of an ETF seems to have a negative influence on tracking errors, but has no additional explanatory power in comparison with the other independent variables. The lack of a significant negative effect of the log transformed volumes on tracking errors in the multiple regression analysis is probably caused by the inclusion of the log transformed fund size. The log transformed fund size does have a significant negative effect on tracking errors. Given the strong positive correlation between both variables, the significance of the coefficient of the log transformed volumes disappears with the inclusion of the log transformed fund size to the multiple regressions.

As discussed, a significant negative effect is found of the log transformed fund size on all three tracking error measures. Based on this result the sixth hypothesis, which states that fund size has a negative effect on tracking errors of U.S. equity ETFs, is supported. This result is in line with expectations and the findings of the studies of Chu (2011) and Buetow and Henderson (2012).

Moreover, the results show that the application of a full replication strategy results in significantly lower tracking errors, as a result of which the seventh hypothesis is supported. This result is consistent with the findings of Frino and Gallagher (2002) and Blitz and Huij (2012) and in line with expectations.

And the eighth hypothesis, which says that the bid-ask spread has a positive effect on tracking errors of U.S. equity ETFs, is also supported. That is because, as expected, a significant positive effect of the spread-to-price ratio on all three tracking error measures is found. This conclusion corresponds to the findings of Frino and Gallagher (2002), Delcoure and Zhong (2007), Mateus and Rahmani (2017) and Meinhardt, Mueller and Schoene (2015).

Finally, the results show that tracking errors in the financial crisis year of 2008 are significantly higher compared to the rest of the years in the sample period. Except for the first tracking error measure, significantly higher tracking errors are also found for 2020 in comparison to the remainder of the sample period. These results support the finding from previous studies that tracking errors are higher in times of crisis. Only the finding that the first tracking error measure is not significantly higher in 2020 compared to the years 2001 to 2019 is not in line with expectations.

7 Conclusion

In this section, the hypotheses test results are summarized one by one, resulting in the answer to the research question. Subsequently, the implications of this study are discussed. Finally, the limitations

of this study are identified and recommendations for future research in this field are given. The research question of this study is:

"How does the tracking efficiency of U.S. equity ETFs evolve over time, and by which factors can it be explained?"

7.1 Answers to the hypotheses & research question

- *Hypothesis 1: U.S. equity ETFs exhibit significant tracking errors over the period from January 2, 2001 to December 31, 2020.*

All three tracking error measures are significantly bigger than zero for this sample of U.S. equity ETFs over the period from January 2, 2001 to December 31, 2020. This outcome supports the first hypothesis. It can be concluded that the ETFs are generally unable to replicate the total returns of their benchmark indices over the period from 2001 to 2020.

The test results of hypotheses 2 to 9 are presented in table 18. Table 18 provides a clear overview of all hypothesis test results. Based on the tests of the hypotheses, the research question can be answered concisely. U.S. equity ETFs exhibit significant tracking errors over the period from January 2, 2001 to December 31, 2020, which means that they are generally unable to perfectly replicate the total returns of their benchmark indices. However, the tracking efficiency of these ETFs fluctuates over time. For example, tracking errors are significantly higher in times of crisis, such as during the financial crisis of 2008 and the year of the coronavirus stock market crash in 2020. Several factors have a significant impact on tracking errors. The volatility of the underlying benchmark index total returns and the bid-ask spread of an ETF have a significant positive effect on tracking errors. Also, tracking a foreign index results in higher tracking errors. And when ETFs focus on small cap investments, it only results in an increased third tracking error measure. In contrast, having a larger fund size, applying a full replication strategy, following a smart beta strategy and having a sector orientation results in significantly lower tracking errors. Lastly, the net expense ratio, dividends and trading volume do not appear to have a significant effect on tracking errors.

Table 18 Overview of answers to the hypotheses

	Hypothesis	Answer
1	U.S. equity ETFs exhibit significant tracking errors over the period from January 2, 2001 to December 31, 2020.	Supported
2	The expense ratio has a positive effect on tracking errors of U.S. equity ETFs.	Not supported

3	The volatility of benchmark index total returns has a positive effect on tracking errors of U.S. equity ETFs.	Supported
4	Dividends have a positive effect on tracking errors of U.S. equity ETFs.	Not supported
5	The trading volume has a negative effect on tracking errors of U.S. equity ETFs.	Not supported
6	The fund size has a negative effect on tracking errors of U.S. equity ETFs.	Supported
7	The application of a full replication strategy has a negative effect on tracking errors of U.S. equity ETFs.	Supported
8	The bid-ask spread has a positive effect on tracking errors of U.S. equity ETFs.	Supported
9	Tracking errors of U.S. equity ETFs are higher during the financial crisis of 2008 and the COVID-19 crisis of 2020.	Supported

7.2 Implications

The findings of this study are of interest for both ETF investors and ETF providers. They learn from this study that ETFs are generally unable to perfectly replicate the total returns of their benchmark indices. In addition, they find out that certain factors have a positive or negative influence on tracking errors, and that tracking errors are higher in times of crisis. This could be a reason for investors not to invest in ETFs during periods when ETFs generally exhibit higher tracking errors. Investors could also capitalize on the factors influencing ETF tracking errors by only including ETFs in their portfolios that generally exhibit lower tracking errors, and excluding ETFs with relatively high tracking errors. For example, a strategy could be to invest in ETFs with a large fund size and a full replication strategy. Finally, ETF providers can use these insights to improve the tracking efficiency of the ETFs they offer, for example by adjusting their replication strategy.

7.3 Limitations & recommendations

This study investigates the tracking errors of U.S. equity ETFs over the period from 2001 to 2020. In addition, it is examined whether tracking errors are significantly higher in times of crisis, during the financial crisis of 2008 and the corona crisis of 2020, and which factors influence tracking errors. However, this research does have its limitations. To begin with, in this study the average daily tracking errors are calculated per year, because the data of some independent variables is only available on an annual basis. If the data for the independent variables could somehow be collected on a monthly basis, it would be interesting to also calculate the average daily tracking errors per month. It could then be examined whether this results in different conclusions regarding the hypotheses and research question. Besides, this research focuses only on U.S. equity ETFs. Expanding the sample with, for example, U.S. fixed income ETFs or ETFs traded in other countries would enhance the validity of the study. Furthermore, as mentioned earlier, the sample does not contain ETFs with a synthetic

replication strategy due to the specific regulations set by the U.S. Securities and Exchange Commission (SEC) in 2010. Thus, if the sample would be expanded with ETFs traded in other countries, the effect of synthetic replication on tracking errors could be examined. This would be of added value, because despite the many studies conducted, there is still no consensus about the effect of synthetic replication on the tracking efficiency of ETFs.

Furthermore, the part of the research that focuses on the factors influencing tracking errors could be further developed in future studies. For example, contrary to expectations, no significant positive effect is found of the annual report net expense ratio on tracking errors. This may be due to the fact that this variable is not a total cost measure, because, among other things, the brokerage costs and sales charges are not included in the ratio. Therefore, for future studies in this field we would recommend to use a more refined cost measure, which reflects the total costs of a fund as closely as possible. Moreover, we expected to find a positive significant effect of the annual dividends on tracking errors, due to the delay in the receipt and reinvestment of dividends and the associated transaction costs. The presence and size of the effect of dividends on tracking errors therefore depends, among other things, on the size of the cash holdings that result from it. It may therefore also be interesting for future studies to include the effect of cash holdings on tracking errors in the analysis. Besides, it could be of added value to investigate whether the impact of certain factors differs over time, because the effect and significance of particular variables could depend on the sample period considered. For example, in future studies it could be interesting to compare the effect of the variables on tracking errors during a crisis period with the effect in a non-crisis period.

8 References

Agapova, A. (2011). Conventional mutual index funds versus exchange-traded funds. *Journal of Financial Markets*, *14*(2), 323–343. https://doi.org/10.1016/j.finmar.2010.10.005

Blitz, D., & Huij, J. (2012). Evaluating the performance of global emerging markets equity exchange-traded funds. *Emerging Markets Review*, *13*(2), 149–158. https://doi.org/10.1016/j.ememar.2012.01.004

Blitz, D., Huij, J., & Swinkels, L. (2012). The Performance of European Index Funds and Exchange-Traded Funds. *European Financial Management*, *18*(4), 649–662. https://doi.org/10.1111/j.1468-036X.2010.00550.x

Buetow, G. W., & Henderson, B. J. (2012). An Empirical Analysis of Exchange-Traded Funds. *The Journal of Portfolio Management*, *38*(4), 112–127. https://www-proquest-com.eur.idm.oclc.org/docview/1033047189?accountid=13598

Cazalet, Z., Grison, P., & Roncalli, T. (2014). The Smart Beta Indexing Puzzle. *The Journal of Index Investing*, *5*(1), 97–119. https://doi.org/https://doi.org/10.3905/jii.2014.5.1.097

Chen, J., Chen, Y., & Frijns, B. (2017). Evaluating the tracking performance and tracking error of New Zealand exchange traded funds. *Pacific Accounting Review*, *29*(3), 443–462. https://doi.org/10.1108/PAR-10-2016-0089

Chiang, W. (1998). *Optimizing performance. Indexing for maximum investment results.*

Chu, P. K.-K. (2011). Study on the tracking errors and their determinants: evidence from Hong Kong exchange traded funds. *Applied Financial Economics*, *21*(5), 309–315. https://web-a-ebscohost-com.eur.idm.oclc.org/ehost/pdfviewer/pdfviewer?vid=1&sid=e71cdd85-d488-44bf-b1b3-d4a0320e59d3%40sessionmgr4007

Cresson, J. E., Cudd, R. M., & Lipscomb, T. J. (2002). The Early Attraction of S&P 500 Index Funds: Is Perfect Tracking Performance An Illusion? *Managerial Finance*, *28*(7), 1–8. https://doi.org/https://doi-org.eur.idm.oclc.org/10.1108/03074350210767933

Delcoure, N., & Zhong, M. (2007). On the premiums of iShares. *Journal of Empirical Finance*, *14*, 168–195. https://doi.org/10.1016/j.jempfin.2005.12.004

Drenovak, M., Urošević, B., & Jelic, R. (2014). European Bond ETFs: Tracking Errors and the Sovereign Debt Crisis. *European Financial Management*, *20*(5), 958–994. https://doi.org/10.1111/j.1468-036X.2012.00649.x

Falato, A., Goldstein, I., & Hortaçsu, A. (2020). *Financial Fragility in the COVID-19 Crisis: The Case of Investment Funds in Corporate Bond Markets*. http://www.nber.org/papers/w27559

Frino, A., & Gallagher, D. R. (2001). Tracking S&P 500 Index Funds. *Journal of Portfolio Management*, *28*(1), 44–55.

Frino, A., & Gallagher, D. R. (2002). Is index performance achievable? An analysis of Australian equity index funds. *Abacus*, *38*(2), 200–214. https://doi.org/10.1111/1467-6281.00105

Horch, A. J. (2020). *Here's why investors started pouring trillions into exchange-traded funds*. CNBC. https://www.cnbc.com/2020/05/29/why-investors-are-pouring-trillions-into-exchange-traded-funds.html

Johnson, B., Bioy, H., Kellett, A., & Davidson, L. (2013a). On the Right Track: Measuring Tracking Efficiency in ETFs. *The Journal of Index Investing*, *4*(3), 35–41. https://doi.org/10.3905/JII.2013.4.3.035

Johnson, B., Bioy, H., Kellett, A., & Davidson, L. (2013b). On the Right Track: Measuring Tracking Efficiency in ETFs. *The Journal of Index Investing*, *4*(3), 35–41. https://doi.org/10.3905/jii.2013.4.3.035

Johnson, W. F. (2009). Tracking errors of exchange traded funds. *Journal of Asset Management*, *10*(4), 253–262. https://doi.org/10.1057/jam.2009.10

Kostovetsky, L. (2003). Index Mutual Funds and Exchange-Traded Funds. *The Journal of Portfolio Management*, *29*(4), 80–92. https://www.proquest.com/scholarly-journals/index-mutual-funds-exchange-traded/docview/195582533/se-2?accountid=13598

Mateus, C., & Rahmani, Y. (2017). Physical versus Synthetic Exchange Traded Funds. Which One Replicates Better? *Journal of Mathematical Finance*, *7*(4), 975–989. https://doi.org/10.4236/jmf.2017.74054

Meinhardt, C., Mueller, S., & Schoene, S. (2015). Physical and Synthetic Exchange-Traded Funds: The Good, the Bad, or the Ugly? *The Journal of Investing*, *24*(2), 35–44.

Milonas, N. T., & Rompotis, G. G. (2006). Investigating European ETFs: The Case of the Swiss Exchange Traded Funds. In *Conference of HFAA in Thessaloniki, Greece.* (pp. 1–27).

Naumenko, K., & Chystiakova, O. (2015). An Empirical Study on the Differences between Synthetic and Physical ETFs. *International Journal of Economics and Finance*, *7*(3), 24–35. https://doi.org/10.5539/ijef.v7n3p24

O'Hara, M., & Zhou, X. (Alex). (2021). Anatomy of a liquidity crisis: Corporate bonds in the COVID-19 crisis. *Journal of Financial Economics*. https://doi.org/10.1016/j.jfineco.2021.05.052

Osterhoff, F., & Kaserer, C. (2016). Determinants of tracking error in German ETFs - the role of market liquidity. *Managerial Finance*, *42*(5), 417–437. https://doi.org/10.1108/MF-04-2015-0105

Pope, P. F., & Yadav, P. K. (1994). Discovering Errors in Tracking Error. *Journal of Portfolio Management*, *20*(2), 27–32. https://www-proquest-com.eur.idm.oclc.org/scholarly-journals/discovering-errors-tracking-error/docview/195574097/se-2?accountid=13598

Qadan, M., & Yagil, J. (2012). On the dynamics of tracking indices by exchange traded funds in the presence of high volatility. *Managerial Finance*, *38*(9), 804–832. https://doi.org/10.1108/03074351211248162

Rompotis, G. G. (2011). Predictable patterns in ETFs' return and tracking error. *Studies in Economics and Finance*, *28*(1), 14–35. https://doi.org/10.1108/10867371111110534

Saha, S. (2021). *Inside the Growing Popularity of ETFs*. Yahoo Finance. https://finance.yahoo.com/news/inside-growing-popularity-etfs-120012702.html?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAC1AobUw793M1wyOmQr_ITnzczQ0l5373YJhRaukY5MjjJp9Ngx7fEliNkX0-SiXm5lgo_-nkrIBezZHtgSv5J1Jmve1aASdNk3ypBTCqjD88WkR6l98DxuqlAjiA3itakgQJLKwG38CUFf-2tOPI7M4eAs43BM88B0dCRTYKpMJ

Shin, S., & Soydemir, G. (2010). Exchange-traded funds, persistence in tracking errors and information dissemination. *Journal of Multinational Financial Management*, *20*(4–5), 214–234. https://doi.org/10.1016/j.mulfin.2010.07.005

Svetina, M., & Wahal, S. (2008). Exchange Traded Funds: Performance and Competition. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1303643

9 Appendix

9.1 Appendix A – figures

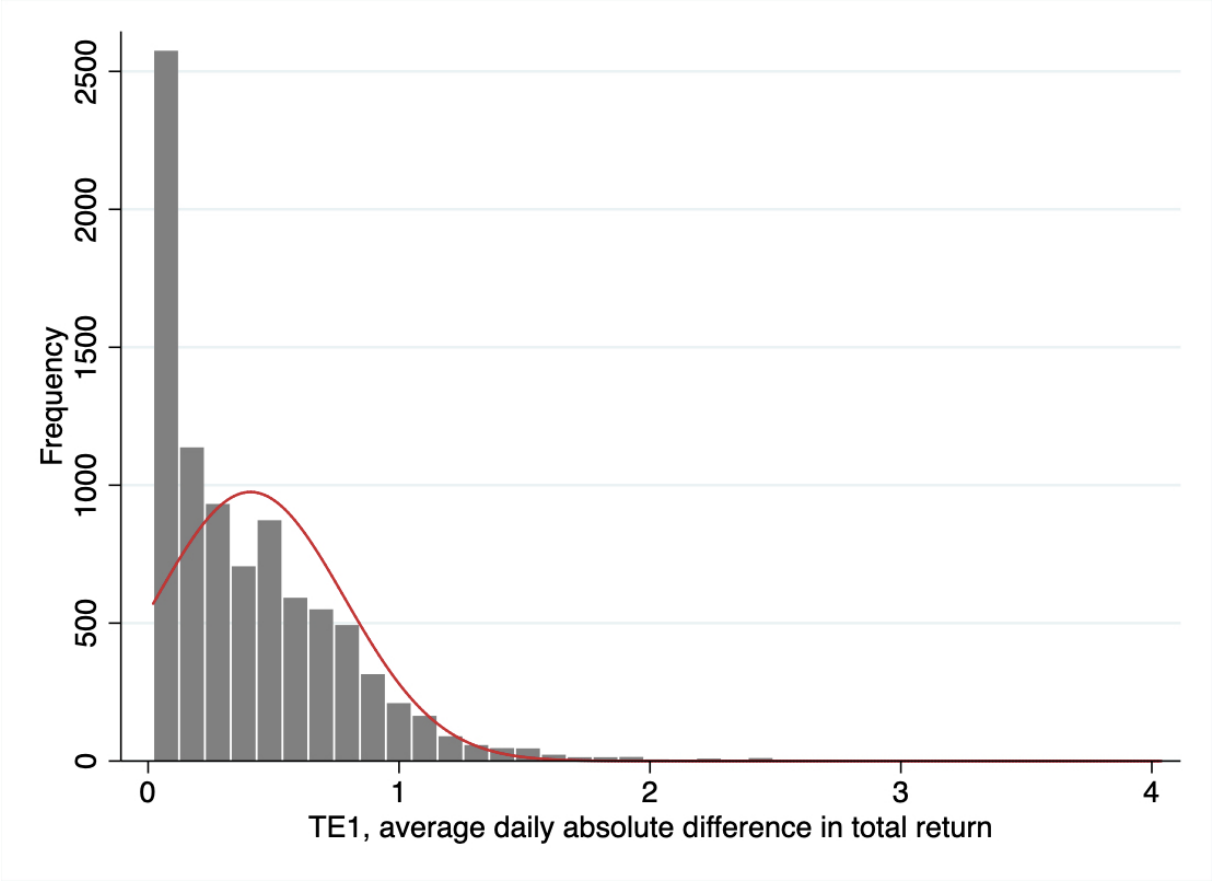

Figure A1 Histogram TE1 (average daily absolute difference in total return)

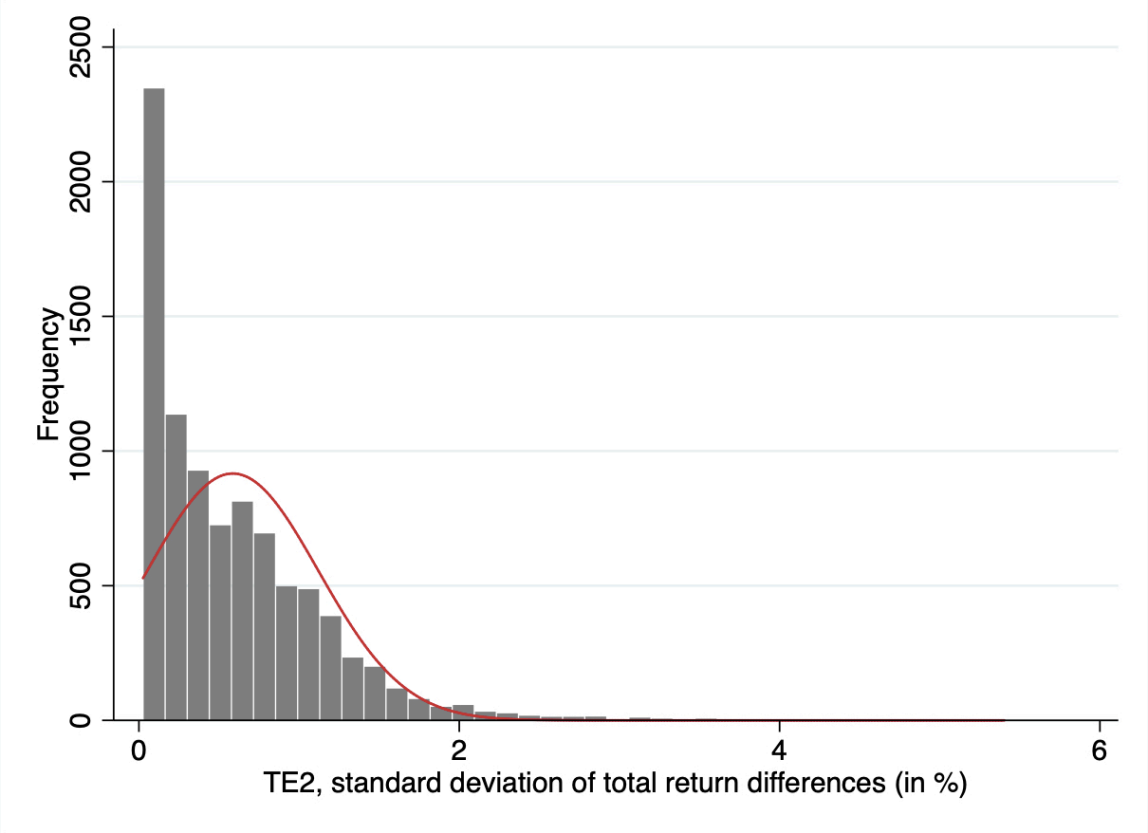

Figure A2 Histogram TE2 (standard deviation of total return differences)

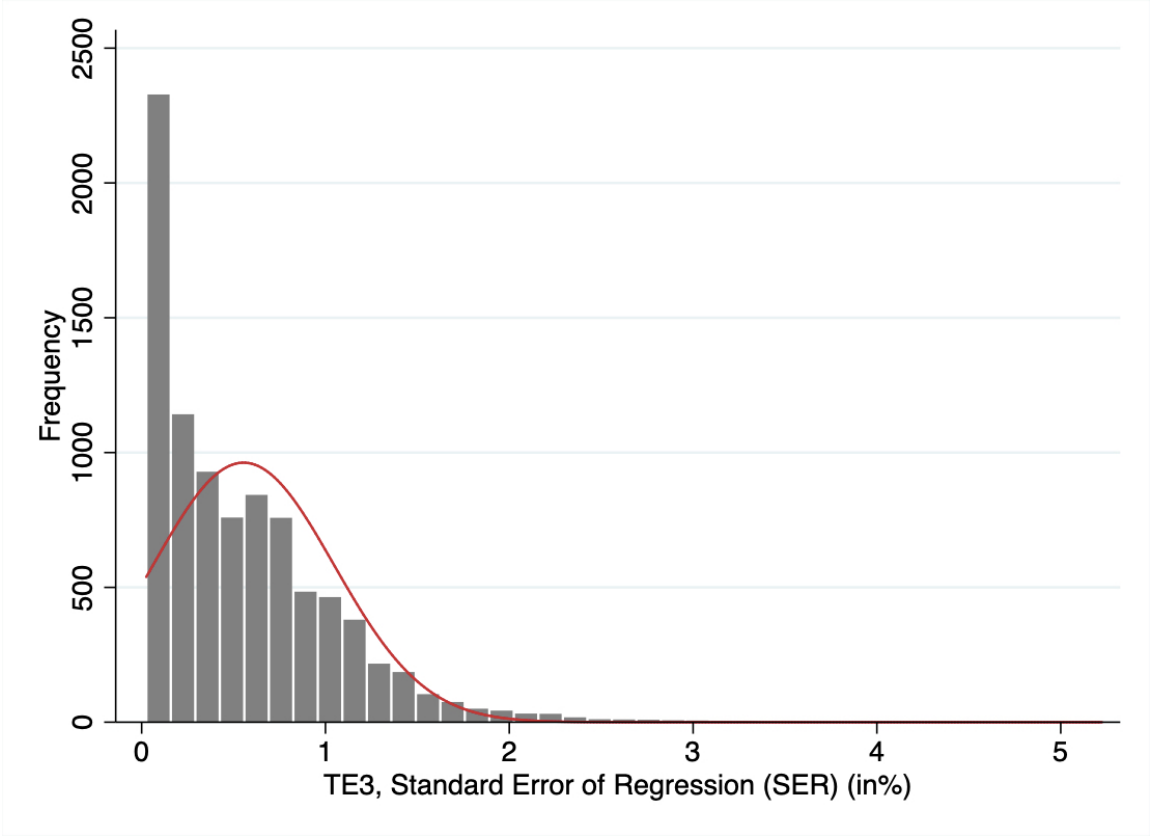

Figure A3 Histogram TE3 (Standard Error of Regression)

9.2 Appendix B – supporting hypotheses 2 to 7

Table B1 Results of the Fisher-type unit root tests based on augmented Dickey-Fuller tests

	Inverse normal (Z)		**Inverse logit (L*)**		**Modified inv. chi-squared (Pm)**	
Variable	Without drift	With drift	Without drift	With drift	Without drift	With drift
TE1	-35.721***	-46.333***	-57.403***	-51.752***	74.562***	71.182***
TE2	-33.034***	-45.370***	-51.635***	-50.205***	65.790***	68.564***
TE3	-31.139***	-44.843***	-48.239***	-49.240***	60.903***	66.845***
NER	-9.851***	-28.557***	-27.881***	-32.631***	28.252***	43.593***
Volatility	-0.123	-31.847***	-1.142	-32.111***	-2.346	38.345***
Dividends	-1.353	-26.455***	-7.839***	-27.345***	16.152***	33.941***
Volumes	-4.057***	-23.891***	-11.230***	-24.540***	18.923***	35.716***
Fund size	-1.545	-21.172***	-10.482***	-21.260***	21.269***	30.337***
Log NER	-5.403***	-26.015***	-19.590***	-29.281***	20.644***	38.592***
Log volatility	-0.746	-33.123***	-1.020	-32.749***	-6.660	38.087***
Log volumes	-11.433***	-33.431***	-21.202***	-35.041***	29.019***	44.365***
Log fund size	-18.208***	-35.612***	-31.738***	-37.573***	43.453***	48.534***

Notes: This table presents the test statistics of the Fisher-type unit root tests based on augmented Dickey-Fuller tests with one lag, both with and without drift term included. ***p<0.001.

Table B2 Results of the multiple random-effects GLS regression model with time-fixed effects and robust clustered standard errors

Variable	**TE1**	**TE2**	**TE3**
NER	0.147***	0.206***	0.207***
	(0.031)	(0.044)	(0.039)
Volatility	0.004	0.015	-0.003
	(0.026)	(0.036)	(0.030)
Dividends	-0.008	-0.015*	-0.013*
	(0.005)	(0.008)	(0.006)
Volumes	-0.007***	-0.008***	-0.007***
	(0.001)	(0.001)	(0.001)
Fund size	-0.000	-0.000	-0.000
	(0.001)	(0.001)	(0.001)
Full replication	-0.030*	-0.039	-0.034*
	(0.014)	(0.020)	(0.017)
Foreign	0.344***	0.458***	0.450***
	(0.016)	(0.023)	(0.020)

Sector	-0.101***	-0.144***	-0.126***
	(0.017)	(0.024)	(0.022)
Small cap	0.052*	0.077*	0.079**
	(0.022)	(0.030)	(0.025)
Mid cap	-0.003	-0.004	0.003
	(0.013)	(0.018)	(0.015)
Smart beta	-0.030*	-0.037	-0.039*
	(0.014)	(0.021)	(0.018)
2002	-0.086	-0.186	0.053
	(0.190)	(0.357)	(0.175)
2003	-0.198	-0.388	-0.153
	(0.190)	(0.357)	(0.179)
2004	-0.311	-0.553	-0.329
	(0.189)	(0.356)	(0.172)
2005	-0.404*	-0.653	-0.424*
	(0.189)	(0.356)	(0.172)
2006	-0.384*	-0.626	-0.391*
	(0.188)	(0.355)	(0.171)
2007	-0.295	-0.495	-0.247
	(0.187)	(0.353)	(0.169)
2008	0.211	0.283	0.488**
	(0.186)	(0.351)	(0.171)
2009	-0.085	-0.221	0.019
	(0.185)	(0.350)	(0.168)
2010	-0.375*	-0.616	-0.357*
	(0.186)	(0.352)	(0.168)
2011	-0.310	-0.505	-0.244
	(0.187)	(0.353)	(0.169)
2012	-0.475*	-0.766*	-0.507**
	(0.188)	(0.355)	(0.170)
2013	-0.516**	-0.800*	-0.543**
	(0.188)	(0.355)	(0.171)
2014	-0.548**	-0.864*	-0.598***
	(0.188)	(0.355)	(0.171)
2015	-0.479*	-0.773*	-0.506**
	(0.187)	(0.354)	(0.169)
2016	-0.472*	-0.758*	-0.503**
	(0.187)	(0.354)	(0.169)

2017	-0.628***	-0.981**	-0.710***
	(0.189)	(0.355)	(0.171)
2018	-0.505**	-0.796*	-0.530**
	(0.187)	(0.353)	(0.169)
2019	-0.639***	-1.004**	-0.723***
	(0.187)	(0.354)	(0.170)
2020	-0.442*	-0.678	-0.378*
	(0.185)	(0.351)	(0.167)
Constant	0.735***	1.139**	0.846***
	(0.196)	(0.365)	(0.180)
Observations	8,896	8,896	8,896
Overall R-squared	0.511	0.494	0.534

Notes: This table presents the results of the multiple random-effects GLS regressions with time fixed effects and clustered standard errors, including all independent variables, except the spread-to-price ratio. The regressions are performed for all three tracking error measures, which are TE1, TE2 and TE3. TE1 is the average daily absolute difference in return between the ETF and its benchmark index; TE2 measures the standard deviation of return differences between the ETF and its benchmark index; TE3 is estimated by the Standard Error of Regression (SER) resulting from the application of the following regression to each ETF-year: $TR_{i,t} = \alpha_i + \beta \cdot TR_{b,t} + \varepsilon_{i,t}$. Standard errors are shown in parentheses. * p<0.05; ** p<0.01; ***p<0.001.

Table B3a AIC and BIC of ML random effects models with time fixed effects and various log transformations to the variables – TE1

TE1		
Log transformed variables	**AIC**	**BIC**
None	-2,317.2866	-2,083.2058
Volatility, volumes, fund size	-3,033.6608	-2,799.5800
Volatility, volumes, fund size, NER	-3,031.8991	-2,797.8926
Volatility, volumes, fund size, Dividends	-3,026.2168	-2,792.1397

Table B3b AIC and BIC of ML random effects models with time fixed effects and various log transformations to the variables – TE2

TE2		
Log transformed variables	**AIC**	**BIC**
None	4,575.2831	4,809.3638
Volatility, volumes, fund size	3,804.3584	4,038.4392
Volatility, volumes, fund size, NER	3,788.6655	4,022.6720
Volatility, volumes, fund size, Dividends	3,809.8498	4,043.9269

Table B3c AIC and BIC of ML random effects models with time fixed effects and various log transformations to the variables – TE3

TE3		
Log transformed variables	**AIC**	**BIC**
None	2,348.3923	2,582.4731
Volatility, volumes, fund size	1,634.5125	1,868.5933
Volatility, volumes, fund size, NER	1,627.3577	1,861.3642
Volatility, volumes, fund size, Dividends	1,640.2688	1,874.3458

Table B4 Variance inflation factors of the independent variables included in the optimal regression model per tracking error measure

	TE1	**TE2**	**TE3**
Variable	**VIF**	**VIF**	**VIF**
NER	1.54	-	-
Log NER	-	1.49	1.49
Log volatility	1.18	1.18	1.18
Dividends	1.22	1.21	1.21
Log volumes	4.50	4.50	4.50
Log fund size	4.99	5.00	5.00
Full replication	1.14	1.13	1.13
Foreign	1.70	1.61	1.61
Sector	1.33	1.36	1.36
Small cap	1.17	1.17	1.17
Mid cap	1.19	1.18	1.18
Smart beta	1.42	1.46	1.46

Table B5 Pairwise correlation table of all dependent and independent variables

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
(1) TE1	1.000														
(2) TE2	0.990*	1.000													
(3) TE3	0.985*	0.992*	1.000												
(4) NER	0.321*	0.307*	0.330*	1.000											
(5) Log NER	0.269*	0.257*	0.279*	0.924*	1.000										
(6) Log volatility	0.276*	0.303*	0.313*	0.131*	0.132*	1.000									
(7) Dividends	-0.096*	-0.106*	-0.105*	-0.197*	-0.189*	-0.100*	1.000								
(8) Log volumes	-0.225*	-0.229*	-0.216*	-0.216*	-0.222*	0.129*	0.129*	1.000							
(9) Log fund size	-0.388*	-0.394*	-0.387*	-0.349*	-0.355*	-0.036*	0.300*	0.849*	1.000						
(10) Full replication	-0.166*	-0.151*	-0.160*	0.029*	-0.020	0.011	-0.026*	-0.051*	0.014	1.000					
(11) Foreign	0.544*	0.509*	0.541*	0.397*	0.339*	-0.053*	0.002	-0.147*	-0.229*	-0.232*	1.000				
(12) Sector	-0.170*	-0.165*	-0.165*	0.157*	0.185*	0.166*	0.016	0.011	-0.058*	0.086*	-0.130*	1.000			
(13) Small cap	-0.056*	-0.051*	-0.055*	-0.041*	-0.020	0.117*	-0.029*	-0.023*	-0.009	0.070*	-0.201*	-0.098*	1.000		
(14) Mid cap	-0.046*	-0.039*	-0.040*	0.179*	0.169*	0.119*	-0.075*	-0.053*	-0.079*	0.058*	-0.110*	0.204*	-0.215*	1.000	
(15) Smart beta	-0.150*	-0.134*	-0.146*	-0.022*	0.026*	-0.122*	-0.050*	-0.141*	-0.020	0.256*	-0.283*	-0.292*	0.130*	-0.027*	1.000

Notes: *p<0.05.

9.3 Appendix C – supporting hypothesis 8

Table C1 Results of the Fisher-type unit root tests based on augmented Dickey-Fuller tests

Variable	Inverse normal (Z)		Inverse logit (L*)		Modified inv. chi-squared (Pm)	
	Without drift	With drift	Without drift	With drift	Without drift	With drift
TE1	1.837	-19.834***	-6.130***	-18.703***	12.295***	18.364***
TE2	4.741	-19.005***	-2.315*	-17.981***	8.081***	16.802***
TE3	5.533	-18.884***	-1.473	-17.907***	7.069***	16.679***
Spread/price ratio	-6.207***	-21.231***	-20.325***	-20.677***	34.063***	22.380***
NER	-2.663**	-15.963***	-15.856***	-15.759***	7.421***	16.286***
Volatility	28.635	-8.829***	28.472	-8.111***	-22.255	3.186***
Dividends	-1.317	-18.818***	-13.126***	-18.308***	25.668***	19.282***
Volumes	-3.256***	-17.082***	-15.954***	-16.813***	31.904***	19.367***
Fund size	-1.392	-16.057***	-14.808***	-15.705***	31.891***	17.887***
Log NER	-0.658	-14.801***	-13.605***	-14.671***	7.009***	15.251***
Log volatility	22.726	-13.755***	21.677	-12.531***	-21.481	7.594***
Log dividends	-3.549***	-20.920***	-15.395***	-20.217***	27.296***	21.150***
Log volumes	-5.794***	-21.221***	-20.804***	-20.648***	35.487***	21.747***
Log fund size	-5.303***	-19.941***	-20.547***	-19.447***	37.625***	20.726***

Notes: This table presents the test statistics of the Fisher-type unit root tests based on augmented Dickey-Fuller tests with one lag, both with and without drift term included. * p<0.05; ** p<0.01; ***p<0.001.

Table C2 Results of the multiple random-effects GLS regression model with time-fixed effects and robust clustered standard errors

Variable	TE1	TE2	TE3
Spread/price ratio	0.108**	0.155**	0.151**
	(0.034)	(0.049)	(0.047)
NER	0.196***	0.284***	0.275***
	(0.033)	(0.049)	(0.042)
Volatility	-0.022	-0.036	-0.056**
	(0.015)	(0.023)	(0.021)
Dividends	-0.015*	-0.021*	-0.021*
	(0.007)	(0.011)	(0.009)
Volumes	0.349	0.049	0.607
	(1.075)	(1.383)	(1.362)

Fund size	0.000	-0.000	-0.000
	(0.000)	(0.000)	(0.000)
Full replication	-0.026*	-0.033	-0.027
	(0.012)	(0.017)	(0.015)
Foreign	0.276***	0.366***	0.357***
	(0.014)	(0.021)	(0.019)
Sector	-0.120***	-0.172***	-0.152***
	(0.015)	(0.022)	(0.020)
Small cap	0.024	0.032	0.041*
	(0.015)	(0.022)	(0.020)
Mid cap	-0.005	-0.012	-0.005
	(0.009)	(0.013)	(0.012)
Smart beta	-0.042**	-0.058**	-0.059***
	(0.013)	(0.019)	(0.016)
2015	0.074***	0.101***	0.103***
	(0.006)	(0.009)	(0.008)
2016	0.078***	0.113***	0.104***
	(0.007)	(0.010)	(0.009)
2017	-0.084***	-0.126***	-0.120***
	(0.005)	(0.008)	(0.008)
2018	0.047***	0.076***	0.078***
	(0.006)	(0.010)	(0.009)
2019	-0.079***	-0.124***	-0.108***
	(0.006)	(0.009)	(0.008)
2020	0.140***	0.252***	0.290***
	(0.020)	(0.031)	(0.029)
Constant	0.192***	0.290***	0.271***
	(0.021)	(0.031)	(0.028)
Observations	5,648	5,648	5,648
Overall R-squared	0.564	0.539	0.574

Notes: This table presents the results of the multiple random-effects GLS regressions with time fixed effects and clustered standard errors, including all independent variables. The regressions are performed for all three tracking error measures, which are TE1, TE2 and TE3. TE1 is the average daily absolute difference in return between the ETF and its benchmark index; TE2 measures the standard deviation of return differences between the ETF and its benchmark index; TE3 is estimated by the Standard Error of Regression (SER) resulting from the application of the following regression to each ETF-year: $TR_{i,t} = \alpha_i + \beta \cdot TR_{b,t} + \varepsilon_{i,t}$. Standard errors are shown in parentheses. * p<0.05; ** p<0.01; ***p<0.001.

Table C3a AIC and BIC of ML random effects models with time fixed effects and various log transformations to the variables – TE1

TE1		
Log transformed variables	**AIC**	**BIC**
None	-5,370.2615	-5,230.8413
Volatility, volumes, fund size	-5,722.4112	-5,582.9910
Volatility, volumes, fund size, spread/price ratio	-5,592.9755	-5,453.5553
Volatility, volumes, fund size, NER	-5,720.0814	-5,580.7357
Volatility, volumes, fund size, dividends	-5,769.2170	-5,629.8005

Table C3b AIC and BIC of ML random effects models with time fixed effects and various log transformations to the variables – TE2

TE2		
Log transformed variables	**AIC**	**BIC**
None	-650.7933	-511.3731
Volatility, volumes, fund size	-1,018.7819	-879.3617
Volatility, volumes, fund size, spread/price ratio	-907.0089	-767.5887
Volatility, volumes, fund size, NER	-1,035.3986	-896.0530
Volatility, volumes, fund size, dividends	-1,058.2954	-918.8789
Volatility, volumes, fund size, NER, dividends	-1,075.6076	-936.2657

Table C3c AIC and BIC of ML random effects models with time fixed effects and various log transformations to the variables – TE3

TE3		
Log transformed variables	**AIC**	**BIC**
None	-1,757.2904	-1,617.8702
Volatility, volumes, fund size	-2,037.8801	-1,898.4599
Volatility, volumes, fund size, spread/price ratio	-1,919.5154	-1,780.0952
Volatility, volumes, fund size, NER	-2,038.3106	-1,898.9649
Volatility, volumes, fund size, dividends	-2,075.7055	-1,936.2890
Volatility, volumes, fund size, NER, dividends	-2,077.4898	-1,938.1479

Table C4 Variance inflation factors of the independent variables included in the optimal regression model per tracking error measure

	TE1	TE2	TE3
Variable	**VIF**	**VIF**	**VIF**
Spread/price ratio	1.44	1.43	1.43
NER	1.54	-	-
Log NER	-	1.44	1.44
Log volatility	1.15	1.15	1.15
Log dividends	1.40	1.37	1.37
Log volumes	6.67	6.67	6.67
Log fund size	7.53	7.58	7.58
Full replication	1.15	1.14	1.14
Foreign	1.77	1.68	1.68
Sector	1.32	1.34	1.34
Small cap	1.17	1.17	1.17
Mid cap	1.24	1.22	1.22
Smart beta	1.43	1.47	1.47

Table C 5 Pairwise correlation table of all dependent and independent variables

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
(1) TE1	1.000															
(2) TE2	0.991*	1.000														
(3) TE3	0.985*	0.991*	1.000													
(4) Spread/price ratio	0.421*	0.413*	0.421*	1.000												
(5) NER	0.327*	0.312*	0.330*	0.254*	1.000											
(6) Log NER	0.269*	0.256*	0.272*	0.215*	0.922*	1.000										
(7) Log volatility	0.163*	0.192*	0.203*	-0.010	0.088*	0.075*	1.000									
(8) Log dividends	-0.163*	-0.171*	-0.169*	-0.179*	-0.324*	-0.277*	-0.140*	1.000								
(9) Log volumes	-0.284*	-0.293*	-0.268*	-0.468*	-0.201*	-0.215*	0.133*	0.193*	1.000							
(10) Log fund size	-0.439*	-0.444*	-0.425*	-0.516*	-0.322*	-0.324*	0.024	0.351*	0.893*	1.000						
(11) Full replication	-0.168*	-0.153*	-0.155*	-0.037*	0.089*	0.044*	0.081*	-0.136*	-0.051*	-0.015	1.000					
(12) Foreign	0.638*	0.607*	0.630*	0.281*	0.337*	0.280*	-0.087*	0.002	-0.115*	-0.245*	-0.239*	1.000				
(13) Sector	-0.247*	-0.243*	-0.242*	-0.043*	0.139*	0.158*	0.169*	-0.083*	0.074*	0.015	0.093*	-0.156*	1.000			
(14) Small cap	-0.065*	-0.062*	-0.060*	0.027*	0.003	0.021	0.135*	-0.033*	-0.029*	-0.009	0.074*	-0.192*	-0.062*	1.000		
(15) Mid cap	-0.077*	-0.074*	-0.072*	0.033*	0.207*	0.184*	0.120*	-0.177*	-0.038*	-0.052*	0.079*	-0.112*	0.246*	-0.211*	1.000	
(16) Smart beta	-0.160*	-0.145*	-0.155*	-0.046*	0.022	0.082*	-0.116*	-0.004	-0.167*	-0.039*	0.235*	-0.285*	-0.282*	0.090*	-0.038*	1.000

Notes: *p<0.05.

9.4 Appendix D – supporting hypothesis 9

Table D1a Results of the Levene's test and Brown-Forsythe tests, 2008 compared to the rest of the sample period

	Std dev (2008)	Std dev (rest)	W0	P-value	W50	P-value	W10	P-value	df
TE1	0.724	0.339	782.212***	.000	616.783***	.000	708.791***	.000	(1; 8,894)
TE2	1.012	0.479	757.410***	.000	619.182***	.000	697.654***	.000	(1; 8,894)
TE3	0.912	0.442	728.100***	.000	606.010***	.000	672.431***	.000	(1; 8,894)

Table D1b Results of the Levene's test and Brown-Forsythe tests, 2020 compared to the rest of the sample period

	Std dev (2020)	Std dev (rest)	W0	P-value	W50	P-value	W10	P-value	df
TE1	0.376	0.375	29.648***	.000	11.014***	.001	24.277***	.000	(1; 8,894)
TE2	0.551	0.532	42.030***	.000	19.368***	.000	35.113***	.000	(1; 8,894)
TE3	0.529	0.486	85.118***	.000	41.362***	.000	71.562***	.000	(1; 8,894)

Notes: Table D1a presents the results of the Levene's test and Brown-Forsythe tests that are applied to the differences in tracking error standard deviations in 2008 and the remainder of the sample period. Table D1b presents the results of the Levene's test and Brown-Forsythe tests that are applied to the differences in tracking error standard deviations in 2020 and the remainder of the sample period. The standard deviations are reported for both groups. W0 is the test statistic for Levene's test centered at the mean; W50 is the first of Brown-Forsythe's tests centered at the median; W10 is the second Brown-Forsythe test centered using the 10% trimmed mean. P-values are reported for all three test statistics. Finally, the degrees of freedom are shown in the last column. TE1 is the average daily absolute difference in return between the ETF and its benchmark index; TE2 measures the standard deviation of return differences between the ETF and its benchmark index; TE3 is estimated by the Standard Error of Regression (SER) resulting from the application of the following regression to each ETF-year: $TR_{i,t} = \alpha_i + \beta \cdot TR_{b,t} + \varepsilon_{i,t}$. ***p<0.001.