Outperformance in Exchange Traded Fund Pricing Deviations: Generalised Control of Data Snooping Bias

Fearghal Kearney*
Finbarr Murphy†
Mark Cummins‡

December 20, 2012

Abstract

An investigation into Exchange Traded Fund (ETF) outperformance during the period 2008-2012 is undertaken utilising a data set of 288 US traded securities. ETFs are tested for Net Asset Value (NAV) premium, underlying index and market benchmark outperformance, with Sharpe, Treynor and Sortino ratios employed as risk adjusted performance measures. A key contribution is the application of an innovative generalised stepdown procedure in controlling for data snooping bias. It is found that a large proportion of optimized replication and debt asset class ETFs display risk adjusted premiums with energy and precious metals focused funds outperforming the S&P500 market benchmark.

Keywords: Exchange Traded Funds; ETF Performance; Multiple Hypothesis Testing; Data Snooping Bias

*DCUBS, Dublin City University, Dublin 9, Ireland. Email: *Fearghal.kearney2@mail.dcu.ie*
†Kemmy Business School, University of Limerick, Ireland. Email: *Finbarr.Murphy@ul.ie*
‡DCUBS, Dublin City University, Dublin 9, Ireland. Email: *Mark.Cummins@dcu.ie*

1 Introduction

Exchange traded funds (ETFs) are variants of mutual funds which first came to prominence in the early 1990s. ETFs allow market participants to trade index portfolios, similar to how individual investors trade shares of a stock. They seek to track the value and volatility of an underlying benchmark index through the construction of portfolios replicative of the index's constituents. They were first traded on the Toronto Stock Exchange in 1989 and today's market boasts over 1,220 US traded ETFs[1]. Investors seeking ETF outperformance may be tempted to apply a number of performance measures to a large data set of ETFs in order to test for those that are profitable. Given enough tests they are virtually certain to uncover individually significant ETFs and may naively use these as a basis for portfolio selection decisions. However, in such a set up, there is a likelihood that these *seemingly significant* outperformers are due to mere chance alone. As the number of simultaneous tests conducted increases so too does the likelihood of such false discoveries. This issue is known as data snooping bias and must be controlled for when studying ETF outperformance. A key contribution in this study is the use of an innovative procedure, proposed in the literature, to control for this problem.

The reason for the growth in popularity of ETFs over recent years can be attributed to a number of advantages that they offer over other index linked products. Tax efficiency and lower expenses are the two most frequently mooted draws for investors, with another being smaller transaction quantities than equivalent futures products, a feature allowing retail investors the opportunity to participate in the market. Empirical studies on active mutual funds have found that, on average, they do not produce above normal returns. Malkiel (1995) and Gruber (1996) both show that this inability to beat the market is primarily due to the level of management expenses charged. This phenomenon has increased interest in passive market tracking funds. ETFs aim to replicate index performance but with lower transaction costs and greater tax efficiency than observed in comparable mutual funds. Actively managed ETFs, whose goal is to realize above market returns only release information on their specific holdings at an end of day frequency, whereas the weighted constituents of the passively managed ETFs are always known. Rompotis (2011) cites this as a major reason both why passive ETFs are advantageous in the eyes of potential arbitrageurs and for their retention as the more popular ETF type. Other miscellaneous strengths of ETFs that have contributed to their rise in popularity have been explicitly identified. Firstly, ETFs provide diversification satisfying broad exposure, be it marketwide or sectoral coverage, with sectoral ETFs facilitating bespoke hedging requirements. Secondly, Alexander and Barbosa (2008) and Yu (2005) observe that ETFs do not have short selling restrictions in the same manner as regular stocks so they may be more useful for hedging. Lastly, ETFs are not subject to the uptick rule which Curcio et al. (2004) cite as another benefit for shareholders.

[1] http://www.ici.org/etf_resources/research/etfs_06_12 (Accessed 30/10/12)

A set of 288 US traded ETFs is considered in this study with hypothesis tests constructed that seek to identify those that outperform their Net Asset Value (NAV), their underlying index or a market benchmark. A major contribution to the literature here is the utilisation of a generalised data snooping bias procedure in the ETF performance appraisal setting. Data snooping bias, in this context, is the problem whereby under naive analysis statistically significant outperformance relationships may be identified by pure chance alone. The false discovery of such random artifacts can greatly mislead an investor's portfolio selection and links directly to the broader issue of multiple hypothesis testing in statistical and econometric applications. The operative balanced stepdown procedure of Romano and Wolf (2010) is applied here which serves as an improvement over the more conservative seminal reality check bootstrap test of White (2000) and the superior predictive ability test of Hansen (2005). It boasts a greater ability to reject false null hypotheses as well as offering balance in the sense that all hypotheses are treated equally in terms of power.

A number of quantitative studies employ such procedures to control for data snooping bias. Sullivan and Timmermann (1999), Hsu and Kuan (2005), Park and Irwin (2007), Marshall et al. (2008) and Qui and Wu (2008) apply the reality check bootstrap test of White (2000) to evaluate the profitability of a wide range of technical trading rules commonly used in industry. Qui and Wu (2008) analyse foreign exchange markets whilst Marshall et al. (2008) considering a data set of 15 commodities. Hsu et al. (2009) employ a stepwise extension of the superior predictability test of Hansen (2005) to re-evaluate the profitability of technical trading rules, with Bajgrowicz and Scaillet (2009) utilising a false discovery rate (i.e. the proportion of false discoveries to the total number of significant hypothesis tests identified) approach to analyse technical trading rules applied to stock returns. Controlling for data snooping bias in a statistical arbitrage setting, Cummins and Bucca (2012) provide a practical comparison of the stepwise procedure of Romano and Wolf (2007) and the balanced stepdown procedure of Romano and Wolf (2010). They find that the balanced stepdown procedure is unbiased in its approach and is shown to identify many more profitable trading strategies compared to the non-balanced stepdown procedure. To asses the performance of hedge funds, Criton and Scaillet (2011) use the false discovery rate to control for data snooping bias. Barras et al. (2010) and Cuthbertson et al. (2008) also utilise the false discovery rate in order to find the proportion of "lucky" mutual funds amongst those with significant individual alphas. However, unless the false discovery rate is zero it is not possible to identify which of the individual funds are genuinely outperforming. This study significantly extends this literature, incorporating the more recent balanced stepdown procedure of Romano and Wolf (2010) and applying this in the ETF space to identify both individual ETFs and ETF cohorts that outperform. The Romano and Wolf (2010) procedure further works on the generalised familywise error rate rather than the false discovery rate - the former being the actual number of false discoveries from the set of all true hypotheses.

The majority of research conducted to date has centred on data sets comprising small numbers of large ETFs, single ETF families or industries, with measurements being applied

inconsistently across the differing studies, inhibiting effective cross comparison. This body of work amends that, primarily through the use of a large, diverse sample size which incorporates many sectoral and internationally focused indices. It investigates numerous ETF attributes and their ability to dictate outperformance, alongside including a recent time period to ascertain the current validity of inferences made in previous studies. The effect of replication type and asset class focus on ETF performance for instance has not been rigorously tested in the literature to date and as such this study incrementally contributes in this way. This work may be of interest to a variety of stakeholders. Firstly, investigating ETF outperformance is significant from an academic perspective as it furthers our understanding of the market's pricing dynamics. Secondly, the wider investment community would benefit from the work as an aid in identifying specific ETF cohorts suitable to individual portfolio requirements. Lastly, arbitrageurs may be interested in the exploitation of any uncovered deviations.

The remainder of the paper is organised as follows. Section 2 discusses in-kind deviations along with performance differences between ETF prices, underlying indices and a market benchmark. Section 3 discusses the issue of data snooping bias and links this to the broader issue of multiple hypothesis testing. The details of the balanced stepdown procedure of Romano and Wolf (2010), along with the associated operative method that allows for computational efficiency. The empirical analysis conducted in this study is outlined in Section 4, describing the data set, defining the formal hypothesis tests and discussing performance. Section 5 presents the results of the empirical analysis and considers various attributes of outperforming funds. Section 6 concludes.

2 Outperformance

This paper seeks to examine ETF outperformance on three levels:

- ETF NAV premium;
- ETF price versus its tracked underlying index;
- ETF price versus a market return benchmark.

NAV premium refers to the amount that the secondary market price of the ETF trades above its calculated NAV. If the amount is negative it is referred to as a NAV discount. The creation/redemption/deletion procedure facilitates exploitation in such situations, whereby the investor can exchange units of trust for the underlying index's stock and vice versa. The return to optimal Law of One Price levels would occur if there were no limits to arbitrage with the most notable observed limitations being market frictions (redemption fees and bid/ask spreads). There is empirical evidence of an inconsistency in premium levels between domestic and non-domestic funds, whereby non-domestic funds display persistent

premiums with US domestic funds tracking their NAVs relatively well. Elton et al. (2002) and Ackert and Tian (2008) both observe that US ETFs are priced close to NAVs, while Engle and Sarkar (2006) and Jares and Lavin (2004) report that some country ETFs display premiums/discounts.Elton et al. (2002) report an average annual return from holding Spiders[2] of 21.91% between the years 1994 and 1998, with the NAV return being slightly lower at 21.89%. They highlight however, that the figures may overstate the true difference as Spiders continue to trade for up to 15 minutes after the New York Stock Exchange closes. Engle and Sarkar (2006) use both daily and intra-day data to investigate short term deviations between the traded price and NAVs of 21 domestic (US) and 16 international ETFs between April and September 2000. They find that ETFs trade in a premium range of between -0.1bps and 4.6bps. US ETFs show minute premiums which are smaller than typical bid-ask spreads whereas international ETFs are less accurately priced due to higher tax and creation/redemption costs. Jares and Lavin (2004) consider mispricings in two Asian ETFs, namely Hong Kong and Japan country funds. They conclude that the non-synchronised trading hours between the US and foreign markets induces the presence of premiums. This study incorporates ETFs from both of these geographic locations.

An ETF is said to have an index tracking error if a fund does not perfectly mirror its underlying benchmark index. Elton et al. (2002) find that the Spider NAV return is on average 28 basis points lower than the return on the S&P500 index. However, when dividends are added to and management expenses deducted from the NAV figure it closely replicates the index return. The influence of expense ratio on ETF outperformance is one of the many factors addressed later in section 5. Harper et al. (2006) provide a comparison of ETFs and closed-end country funds, observing no significant tracking error between iShares ETFs and MSCI [3] indices during the period April 1996 to December 2001. DeFusco et al. (2011) study the three most liquid ETFs, the Spiders, Diamonds and Cubes[4]. Through setting out 5 hypothesis tests on the non-synchronous price deviations between the ETFs and the notional price of the index, they conclude that the tracking error is a non-zero, non-normal, stationary process that is dependent on both the accumulation of dividends and on the size of the benchmark index. This paper deals with the size issue through the proxy of each ETF's Total Assets Under Management.

Market tracking error in this context refers to how much an ETF under/outperforms a broad market index. The majority of mutual fund and ETF studies to date utilise the S&P500 as their US benchmark index proxy alongside incorporating risk adjusted returns into the analysis.Phengpis and Swanson (2009), using monthly data and incorporating the

[2]Standard & Poor's Depository Receipts ("Spiders" or SPDRs) track the performance of the S&P 500 Index

[3]MSCI is an abbreviation of Morgan Stanley Capital International. iShares are ETFs tracking the performance of MSCI individual country market indices.

[4]Diamonds and Cubes are ETFs tracking the performance the Dow Jones Industrial Average and NASDAQ 100 indices respectively.

Wilshire 3000 index to represent the US market return, find that country iShares are not heavily exposed to US market risk. The results are obtained using a new two factor test specification with the iShares typically mirroring their underlying market indices up to the end of March 2007. The relationship between a US market benchmark and country iShares is revisited in this study. Mateus and Kuo (2008) also study ETF performance, providing a comparative analysis of 20 country-specific ETFs with the S&P500 equity index over a five year period. Risk adjusted measures are used, namely, Sharpe, Treynor and Sortino Ratios. Sharpe and Sortino ratios are again calculated by Rompotis (2011), who shows that the majority of the 50 selected iShares in his sample outperform the S&P500 on both an annual and aggregate basis over the 2002 to 2007 period.

3 Multiple Hypothesis Testing: Data Snooping Bias

The objective of the study is to formally and statistically test for the presence of outperformance in ETF markets. This will inevitably involve the testing of a large number of performance measure implementations simultaneously. In particular, 11 pricing deviations are considered for each of the 288 ETFs, leading to the simultaneous assessment of 3,168 performance measures. This introduces the well-established issue of data snooping bias, which in this context, is the problem whereby under naive analysis statistically significant outperformance relationships may be identified by pure chance alone. The false discovery of such random artifacts can greatly mislead an investor's portfolio selection and links directly to the broader issue of multiple hypothesis testing in statistical and econometric applications.

The issue with multiple hypothesis testing is that the probability of false discoveries, i.e. the rejection of true null hypotheses by chance alone, is often significant. There are a number of approaches described in the literature to deal with this multiple comparisons problem and control for the familywise error rate (FWER) and related variants. Romano et al. (2010) provide an excellent summary of the issues and the literature. The FWER is defined as the probability that at least one or more false discoveries occur. Consistent with the notation of Romano et al. (2010), the following definition is made:

$$FWER_\theta = P_\theta \left\{ \text{reject at least one null hypothesis } H_{0,s} : s \in \mathcal{I}(\theta) \right\},$$

where $H_{0,s}, s = 1, \ldots, S$, is a set of null hypotheses; and $\mathcal{I}(\theta)$ is the set of true null hypotheses. Controlling the FWER involves setting a significance level α and requiring that $FWER_\theta \leq \alpha$. This approach is particularly conservative given that it does not allow even for one false discovery and so is criticised for lacking *power*, where power is loosely defined as the ability to reject false null hypotheses, i.e. identify true discoveries (Romano et al. (2010)). The greater S, the more difficult it is to make true discoveries.

To deal with this weakness, generalised FWER approaches have been proposed in the literature. The generalised FWER seeks to control for k (where $k \geq 1$) or more false

discoveries and, in so doing, allows for greater power in multiple hypothesis testing. The generalised k-FWER is defined as follows:

$$k\text{-}FWER_\theta = P_\theta \left\{\text{reject at least } k \text{ null hypothesis } H_{0,s} : s \in \mathcal{I}(\theta)\right\}.$$

Towards building a framework to identify outperforming ETFs, with statistical significance, the following one-sided hypothesis tests will be considered:

$$H_{0,s} : \theta_s \leq 0 \quad \text{vs.} \quad H_{1,s} : \theta_s > 0.$$

The objective is to control for the multiple comparisons in this scenario through the generalised FWER, which offers greater power while also implicitly accounting for the dependence structure that exists between the tests. Before outlining the balanced stepdown procedure of Romano and Wolf (2010) , it is first necessary to present the (inferior) single step procedure designed around the generalised FWER. The advantages of the Romano and Wolf (2010), procedure are better appreciated with this context.

3.1 Single-Step Procedure

Assume a set of test statistics $T_{n,s} = \hat{\theta}_{n,s}$ associated with the hypothesis tests, where n is introduced to denote the sample size of the data used for estimation. Letting $A \equiv \{1, \ldots, S\}$, the single step procedure proceeds by rejecting all hypotheses where $T_{n,s} \geq c_{n,A}(1-\alpha, k)$, and where $c_{n,A}(1-\alpha, k)$ represents the $(1-\alpha)$-quantile of the distribution of k-max $\left(\hat{\theta}_{n,s} - \theta_s\right)$ under P_θ. With P_θ unknown, the critical value $c_{n,A}(1-\alpha, k)$ is also unknown. However, an estimate critical value may be determined using appropriate bootstrapping techniques. That is, the critical value $\hat{c}_{n,A}(1-\alpha, k)$ is estimated as the $(1-\alpha)$-quantile of the distribution of k-max $\left(\hat{\theta}^*_{n,s} - \hat{\theta}_{n,s}\right)$ for $\hat{P}_\theta$ an unrestricted estimate of P_θ. See Romano et al. (2010) for further technical details.

3.2 Balanced Stepdown Procedure

The single step procedures is improved upon with the balanced stepdown procedure of Romano and Wolf (2010) by allowing for subsequent iterative steps to identify additional hypothesis rejections. It also offers balance by construction in the sense that each hypothesis is treated equally in terms of power. The stepdown procedure is constructed such that at each stage, information on the rejected hypotheses to date is used in re-testing for significance on the remaining hypotheses.

Again assume a set of test statistics $T_{n,s} = \hat{\theta}_{n,s}$ associated with the hypothesis tests, where n is again the sample size of the data used for estimation. Introducing some notation,

let $H_{n,s}(\cdot, P_\theta)$ denote the distribution function of $\left(\hat{\theta}_{n,s} - \theta_s\right)$ and let $c_{n,s}(\gamma)$ denote the γ-quantile of this distribution. The confidence interval

$$\left\{\theta_s : \hat{\theta}_{n,s} - \theta_s \leq c_{n,s}(\gamma)\right\}$$

then has coverage probability γ. Balance is the property that the marginal confidence intervals for a population of S simultaneous hypothesis tests have the same probability coverage. Within the context of controlling the generalised k-FWER, the overall objective is to ensure that the simultaneous confidence interval covers all parameters $\theta_s, s = 1, ..., S$, except for at most $(k-1)$ of them, for a given limiting probability $(1-\alpha)$, while at the same time ensuring balance (at least asymptotically). So, what is sought is that

$$P_\theta\left\{\hat{\theta}_{n,s} - \theta_s \leq c_{n,s}(\gamma) \quad \text{for all but at most } (k-1) \text{of the hypotheses}\right\}$$

$$\equiv P_\theta\left\{H_{n,s}\left(\hat{\theta}_{n,s} - \theta_s, P_\theta\right) \leq \gamma \quad \text{for all but at most } (k-1) \text{of the hypotheses}\right\}$$

$$\equiv P_\theta\left\{k\text{-}\max\left(H_{n,s}\left(\hat{\theta}_{n,s} - \theta_s, P_\theta\right)\right) \leq \gamma\right\} = 1 - \alpha.$$

Letting $L_{n,\{1,...,S\}}(k, P_\theta)$ denote the distribution of $k\text{-}\max\left(H_{n,s}\left(\hat{\theta}_{n,s} - \theta_s, P_\theta\right)\right)$, the appropriate choice of the coverage probability γ is then $L^{-1}_{n,\{1,...,S\}}(1-\alpha, k, P_\theta)$.

Given that P_θ is unknown, it necessary to use appropriate bootstrapping techniques to generate an estimate of the coverage probability $L^{-1}_{n,\{1,...,S\}}\left(1-\alpha, k, \hat{P}_\theta\right)$, under $\hat{P}_\theta$. Therefore, from this development it is possible to define the simultaneous confidence interval

$$\left\{\theta_s : \hat{\theta}_{n,s} - \theta_s \leq H^{-1}_{n,s}\left(L^{-1}_{n,\{1,...,S\}}\left(1-\alpha, k, \hat{P}_\theta\right), \hat{P}_\theta\right)\right\}.$$

The right hand side of the above inequality will form the basis of the critical value definitions used within the stepdown procedure. See Romano and Wolf (2010) for further technical details. Note that although the above development was made assuming the full set of hypothesis tests, it equally applies to any subset $K \subseteq \{1, \ldots, S\}$. Hence, the balanced stepwise algorithm may now be described as follows.

- **Step 1:** Let A_1 denote the full set of hypothesis indices, i.e. $A_1 \equiv \{1, \ldots, S\}$. If for each hypothesis test, the associated test statistic $T_{n,s}$ is less than or equal to the corresponding critical value estimate $\hat{c}_{n,A_1,s}(1-\alpha, k) \equiv H^{-1}_{n,s}\left(L^{-1}_{n,A_1}\left(1-\alpha, k, \hat{P}_\theta\right), \hat{P}_\theta\right)$ then fail to reject all null hypotheses and stop the algorithm. Otherwise, proceed to reject all null hypotheses $H_{0,s}$ for which the associated test statistics exceeds the critical value level, i.e. where $T_{n,s} > \hat{c}_{n,A_1,s}(1-\alpha, k)$.

- **Step 2:** Let R_2 denote the set of indices for the hypotheses rejected in Step 1 and let A_2 denote the indices for those hypotheses not rejected. If the number of elements in R_2 is less than k, i.e. $|R_2| < k$, then stop the algorithm; as the probability of k or more false discoveries is zero in this case. Otherwise, the appropriate critical value to be applied for each hypothesis test s at this stage is calculated as follows:

$$\hat{d}_{n,A_2,s}(1-\alpha,k) = \max_{I \subseteq R_2, |I|=k-1} \{\hat{c}_{n,K,s}(1-\alpha,k) : K \equiv A_2 \cup I\}.$$

  Hence, additional hypotheses from A_2 are rejected if $T_{n,s} > \hat{d}_{n,A_2,s}(1-\alpha,k), s \in A_2$. If no further rejections are made then stop the algorithm.
  $\vdots$

- **Step j:** Let R_j denote the set of indices for the hypotheses rejected up to Step $(j-1)$ and let A_j denote the indices for those hypotheses not rejected. The appropriate critical value to be applied for each hypothesis test s at this stage is calculated as follows:

$$\hat{d}_{n,A_j,s}(1-\alpha,k) = \max_{I \subseteq R_j, |I|=k-1} \{\hat{c}_{n,K,s}(1-\alpha,k) : K \equiv A_j \cup I\}.$$

  Hence, additional hypotheses from A_j are rejected if $T_{n,s} > \hat{d}_{n,A_j,s}(1-\alpha,k), s \in A_j$. If no further rejections are made then stop the algorithm.
  $\vdots$

At each stage j in the stepwise procedure, the hypotheses that are not rejected thus far are re-tested over a smaller population of hypothesis tests than previously. The size of this smaller population is given $(|A_j| + k - 1)$, which includes all the hypotheses within A_j, in addition to $(k-1)$ hypotheses drawn from those hypotheses already rejected, i.e. drawn from R_j. Given that control of the generalised k-FWER is the premise of the procedure, it is expected that there are at most $(k-1)$ false discoveries amongst the set of hypotheses rejected R_j. However, it is not known which of the rejected hypotheses may represent false discoveries. Hence, it is necessary to circulate through all combinations of R_j, of size $(k-1)$, in order to obtain the appropriate critical values. A maximum critical value $\hat{d}_{n,A_j,s}(1-\alpha,k)$ must be determined for each hypothesis test s. This adds an additional layer of computational burden on the algorithm.

3.2.1 Operative Method

In requiring to circulate through all subsets of R_j, of size $(k-1)$, in order to obtain the maximum critical value to apply at each stage of the stepdown procedure, the algorithm can become highly, if not excessively, computationally burdensome. Depending on the $|R_j|$ and

the value of k, the number of combinations ${}^{|R_j|}C_{k-1}$ can become very large. Romano and Wolf (2010) therefore suggest an operative method that reduces this computational burden, while at the same time maintaining much of the attractive properties of the algorithm.[5]

It is first necessary to be able to order the hypothesis tests rejected up to step $(j-1)$ in terms of significance. To this end, it is noted that marginal$p$-values can be obtained as follows:

$$\hat{p}_{n,s} \equiv 1 - H_{n,s}\left(\hat{\theta}_{n,s}, \hat{P}_\theta\right).$$

This gives the following ascending order for the significance of the hypothesis tests:

$$\hat{p}_{n,r_1} \leq \hat{p}_{n,r_2} \leq \ldots \leq \hat{p}_{n,r_{|R_j|}},$$

where $\left\{r_1, r_2, \ldots, r_{|R_j|}\right\}$ is the appropriate permutation of associated hypothesis test indices that gives this ordering. As before, a maximum number of combinations, N_{max}, at each step of the algorithm is defined. Then an integer value M is chosen such that ${}^{M}C_{k-1} \leq N_{max}$, leading to the calculation of the critical values as follows:

$$\hat{d}_{n,A_j,s}\left(1-\alpha,k\right) = \max_{I \subseteq \left\{r_{\max\left(1,\left|R_j\right|-M+1\right)},\ldots,r_{\left|R_j\right|}\right\}, |I|=k-1} \left\{\hat{c}_{n,K,s}\left(1-\alpha,k\right) : K \equiv A_j \cup I\right\}.$$

What this serves to do is to replace circulating through all the hypothesis tests rejected to date with that of circulating through only the M least significant hypothesis tests rejected. Of course, in the case where $M \geq |R_j|$ then this amounts to circulating through all the hypotheses rejected. Although this approach is premised on the assumption that the (up to $k-1$) false discoveries lie within the least significant hypotheses rejected so far, it does offer significant computational efficiencies for the algorithm. It is this operative method that is used for the empirical analysis in subsequent sections.

4 Empirical Analysis: Framework and Data

The balanced stepdown procedure described in the previous section offers a more generalised and flexible approach to controlling data snooping bias than previous methodologies in the literature. In particular, it controls the generalised FWER using a superior stepwise procedure that offers balance by construction. For this reason, it is utilised for the empirical analysis of this study. Firstly, in order to test for ETF premiums, the differences between the mean daily log return of the quoted ETF price and the mean daily log return of its

[5] Attractive properties include: conservativeness; allows for finite sample control of the k-FWER under P_θ; and provides asymptotic control in the case of contiguous alternatives Romano and Wolf (2007)

reported NAV are examined, with the null hypothesis being that the ETF return is less than or equal to the NAV return, i.e. no outperformance.[6] The analysis is extended through the implementation of traditional risk adjusted measures such as the Sharpe, Sortino and Treynor ratio test statistics with the null hypotheses of no outperformance again in place. The same approach is employed in constructing index and market outperformance hypothesis tests, replacing the NAV series with the fund's underlying index and the S&P500 series respectively.

The three risk adjusted ratios are now examined. The Sharpe ratio (Sharpe (1966)), is the most commonly used ex post measure of risk adjusted performance in ETF literature. It is a measure of an investment's performance per unit of risk, whereby standard deviation is used as a proxy for the portfolio's risk. The Treynor Ratio is a variant of the Sharpe Ratio which incorporates a CAPM based excess return component, effectively giving excess return per unit of market risk. Where the normality assumption is not in place for returns it is beneficial to consider the Sortino Ratio, the third risk adjusted measure of performance considered. It is again based on the Sharpe Ratio but differentiates between upside and downside risk whereby it does not penalized for upside volatility. Formally, these risk-adjusted measures are summarised as follows:

$$\rho_p = \frac{R_p - r_f}{\eta_p} \ ,$$

where ρ_p = Portfolio's Sharpe, Sortino or Treynor Ratios, R_p = Portfolio Return, η_p = Standard Deviation of Portfolio for Sharpe, Standard Deviation of Negative Returns for Sortino or Market Beta for Treynor Ratios and r_f = Risk Free Rate.

As referred to previously, for each of the 288 ETFs, 11 pricing deviations are calculated on a daily basis.[7] Formally:

[6] Use of the log return methodology is in line with Engle and Sarkar (2006)

[7] Note that the construction of the Treynor Ratio, which incorporates the market beta figure, is the reason for the omission of a Mkt TE TR measure. TE is an abbreviation of "Tracking Error".

Outperformance Measure	Assigned Name
ETF price log return− ETF NAV log return	Premium
ETF price log return−Underlying index's log return	Index TE
ETF price log return−S&P500 log return	Mkt TE
ETF price log return Sharpe Ratio− ETF NAV log return Sharpe Ratio	Premium SR
ETF price log return Sharpe Ratio−Underlying index's log return Sharpe Ratio	Index TE SR
ETF price log return Sharpe Ratio−S&P500 log return Sharpe Ratio	Mkt TE SR
ETF price log return Sortino Ratio−ETF NAV log return Sortino Ratio	Premium SorR
ETF price log return Sortino Ratio−Underlying index's log return Sortino Ratio	Index TE SorR
ETF price log return Sortino Ratio−S&P500 log return Sortino Ratio	Mkt TE SorR
ETF price log return Treynor Ratio−ETF NAV log return Treynor Ratio	Premium TR
ETF price log return Treynor Ratio−Underlying index's log return Treynor Ratio	Index TE TR

Table 1: Outperformance Measures

To complete the set up of the empirical analysis, it is necessary to discuss the choice of generalizing parameter k and the probability parameter α to be used within the balanced stepdown procedures. To ensure tight control of the number of false discoveries while at the same time offering power to the tests, k is chosen to ensure that no more than 1% of the tests considered represent false discoveries. The significance level α chosen is 5% alongside an N_{max} value of 100 combinations in line with Romano and Wolf (2010).

Attention is now turned to the composition of the sample data. The data set comprises 288 US domiciled equity, commodity and debt ETFs with pre-2008 inception dates. The period of study is 2008-2012, a time span that is chosen to strike an acceptable balance between being sufficiently long to retain power in the proposed econometric tests and recent enough to be representative of the vast array of ETFs which are currently on offer. Historical information on end of day market price, reported NAV and the notional value of the tracked is downloaded from Bloomberg for each fund. Supplementary data on total asset value, underlying asset class, replication strategy, expense ratio, industry and country focus is also assimilated. Table 2 outlines the observed cohort proportions of the data set. It boasts funds in the Assets Under Management range of \$9.72m to \$101,187.40m with a broad industry split; 18 from the energy sector, 14 from technology, 12 from financial services and 11 from health and biotechnology for example[8]. The median expense ratio is 0.51 with a range of 0.09 to 2.55 observed. The sample includes both many US and non US focused funds[9] along with full, optimized and derivative replication types. A major contribution of this study

[8]Unfortunately a large number of the ETFs are classed as N.A and so unidentifuable which is due to being either a cross industry ETF or an ETF that has not provided the information to Bloomberg.

[9]International ETFs refer to investments targeted at multiple geographic locations outside of the home market (US) whereas Global ETFs refer to investments targeted at multiple geographic locations inclusive of the home market (US).

is borne out of the inclusion of these additional factors as they allow for more informed portfolio selection decisions. Average daily risk free rates are downloaded from the website of Kenneth French[10] in a manner similar to Rompotis (2011). These are to be utilised in the calculation of risk adjusted performance measures.

As identified earlier, the use of the Sortino ratio is appropriate and valid where returns are shown to be non-normal. For completeness, the normality of returns is formally tested for each of the 288 ETF price, the 288 NAV and the 288 index series. The hypothesis that the returns are normal is tested using the Jarque-Bera two-sided goodness of fit test[11]. The multiple comparisons problem presents itself here again due to conducting 864 Jarque-Bera normality tests simultaneously. Given the availability of p-values from the Jarque-Bera tests, the use of a p-value based multiple hypothesis testing (MHT) procedure is appropriate here.[12] The MHT framework of Romano and Shaikh (2006) is employed here, controlling for what is referred to as the False Discovery Proportion (FDP). It is defined as:

$$FDP \equiv \begin{cases} \frac{FR}{TR}, & TR > 0 \\ 0, & TR = 0 \end{cases},$$

where FR denotes the number of false rejections and TR denotes the total number of rejections. Romano and Shaikh (2006) propose a stepdown procedure that controls the FDP, whereby for a given proportion $\tilde{\gamma}$ and significance level $\tilde{\alpha}$,

$$P\{FDP > \tilde{\gamma}\} \leq \tilde{\alpha}.$$

For the set of hypothesis tests $H_{0,i}, i = 1, \ldots, 864$, there are available p-values $\hat{p}_i, i = 1, \ldots, s$. The p-values are ordered from the most significant down to the least significant, i.e. $\hat{p}_{(1)} \leq \hat{p}_{(2)} \ldots \leq \hat{p}_{(s)}$, and the associated ordered null hypotheses $H_{0,(i)}$ are rejected if and only if $\hat{p}_{(i)} \leq \tilde{\alpha}'_{(i)}$ with the cut-off values defined as:[13]

$$\tilde{\alpha}'_{(i)} \equiv \tilde{\alpha}_{(i)}/C,$$

where

[10] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html (Accessed 30/06/12)

[11] The null hypothesis is that the deviations are normally distributed with unspecified mean and standard deviation, whereas the alternative is that the deviations are not normally distributed.

[12] There are two classifications of procedure identified in the MHT literature: (i) re-sampling based and (ii) p-value based. The balanced stepdown procedure outlined in Section 4 is of the re-sampling type, involving a bootstrapping component. See Romano et al. (2010) for more details of both classifications.

[13] It is important to emphasise the subtle difference in notation. $H_{0,i}$ is the i-th hypothesis test considered and $\hat{p}_i$ is the associated p-value. In contrast, $H_{0,(i)}$ is used to denote the i-th hypothesis when all hypotheses are ordered in terms of significance from the most significant down to the least significant, with $\hat{p}_{(i)}$ denoting the associated ordered p-value.

$$\tilde{\alpha}_{(i)} = \frac{(\lfloor \tilde{\gamma} i \rfloor + 1)\tilde{\alpha}}{s + \lfloor \tilde{\gamma} i \rfloor + 1 - i}$$

and

$$C \equiv C(\tilde{\gamma}, \tilde{\alpha}, s) = \max_{|I|} S(\tilde{\gamma}, \tilde{\alpha}, |I|),$$

$$S(\tilde{\gamma}, \tilde{\alpha}, |I|) \equiv |I| \sum_{j=1}^{N} \frac{\beta_j - \beta_{j-1}}{j},$$

$$N \equiv N(\tilde{\gamma}, \tilde{\alpha}, |I|) = \min \left\{ \lfloor \tilde{\gamma} s \rfloor + 1, |I|, \left\lfloor \tilde{\gamma} \left(\frac{s - |I|}{1 - \tilde{\gamma}} + 1 \right) \right\rfloor + 1 \right\},$$

and where

$$\beta_0 \equiv 0,$$

$$\beta_m \equiv \frac{m}{max\left\{ s + m - \left\lceil \frac{m}{\tilde{\gamma}} \right\rceil + 1, |I| \right\}}, m = 1, ..., \lfloor \tilde{\gamma} s \rfloor,$$

and

$$\beta_{\lfloor \tilde{\gamma} s \rfloor + 1} \equiv \frac{\lfloor \tilde{\gamma} s \rfloor + 1}{|I|}.$$

This approach boasts robustness to the dependence structure of the p-values. The proportion parameter $\tilde{\gamma}$ is chosen to be 5% with the significance level $\tilde{\alpha}$ set at 5% also. See Romano and Shaikh (2006) for further details.

Upon implementing the procedure, significant non-normality is observed for all price, NAV and underlying index series, confirming the use of the Sortino Ratio as appropriate. Even though the sample ETF returns are not normally distributed, traditional risk adjusted ratios; Sharpe and CAPM based Treynor Ratios are extensively used in previous studies and will again be applied in this body of work. They provide an intuitive way of comparing results between studies and offer numerous practical applications in measuring both ETF and mutual fund performance (Mateus and Kuo (2008)). Plantinga et al. (2001) examine the application of risk adjusted ratios to Euronext mutual funds and find that there is a high correlation between the classic Sharpe ratio and a ratio controlling for downside risk, adding further weight to the applicability of such performance measures. The next section presents the results subsequent to applying the balanced stepdown procedure described in Section 3 to the data set.

Industry Focus	Count
N.A.	198
Energy	18
Technology Sector	14
Financial Services	12
Health & Biotechnology	11
Real Estate Sector	10
Utility Sector	7
Precious Metals Sector	7
Environmentally Friendly	4
Internet/Telecommunications	4
Leisure Industry Sector	2
Food/Beverage Sector	1

Geographic Focus	Count
United States	188
International	34
Global	27
China	5
European Region	3
Japan	3
Asian Pacific Region ex Japan	2
Latin American Region	2
Other	24

Asset Allocation	Count	Derivative Replication	Count
Equity	263	Full	145
Commodity	13	Optimized	62
Debt	9	Unknown	46
Asset Allocation	3	Derivative	35

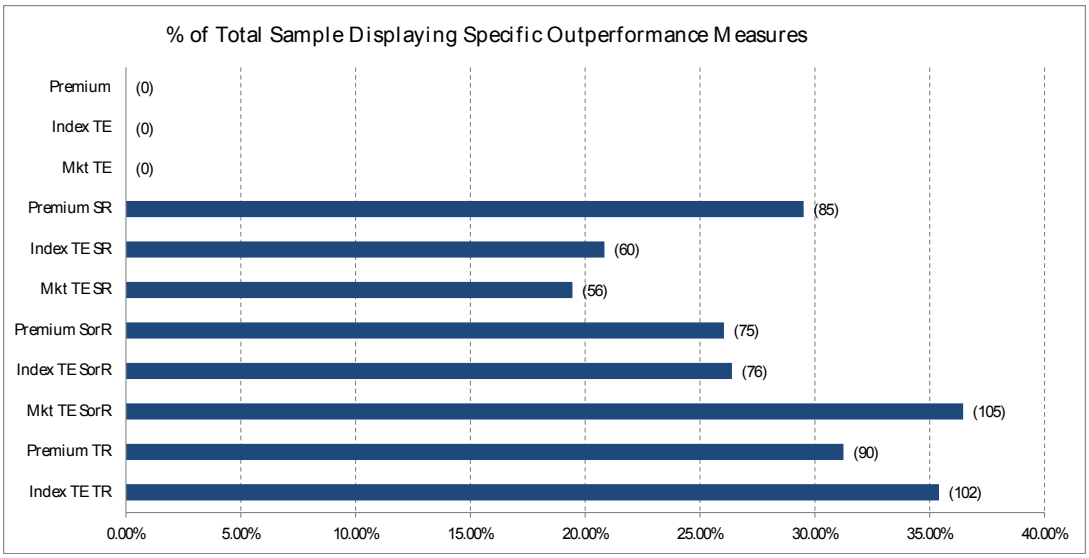
[a]Count of ETFs in data set split by various attributes

Table 2: Data Set Properties

5 Empirical Analysis: Results

The results of the operative balanced stepdown procedure of Romano and Wolf (2010) are presented in Figure 1, giving the percentage (the actual numbers are given in parenthesis) of ETFs in the sample which display specific outperformance measures that can be stood over with statistical confidence. The main item of note is that none of the log return outperformance measures are significant under the balanced stepdown procedure. This leads to relying primarily on inferences made around the risk adjusted measures for the remainder of the paper. The various measures display differing numbers of outperforming funds; for instance 56 funds show market benchmark outperformance under the Sharpe Ratio with almost twice that figure, 105 funds, outperforming the market under the Sortino Ratio measure. Summary statistics for the significant outperforming funds are given in Table 3, providing the average outperformance measure. The results highlight the importance of controlling for data snooping bias. On the basis of the three non-risk adjusted measures, i.e. premium, index tracking error and market tracking error, it is found that none of the funds outperform. Failure to apply the data snooping bias control procedure would have led to the naive identification of outperformance and so investing on such a basis would constitute naive and misinformed portfolio selection.

A number of ETF attributes are now analysed in turn to determine what class of ETFs are most likely to demonstrate risk adjusted outperformance and specifically what outperformance measures they show. Geographic and industry focus are the first to be considered. The geographic focus of ETFs is studied in Figure 2, with a high proportion of Global, In-



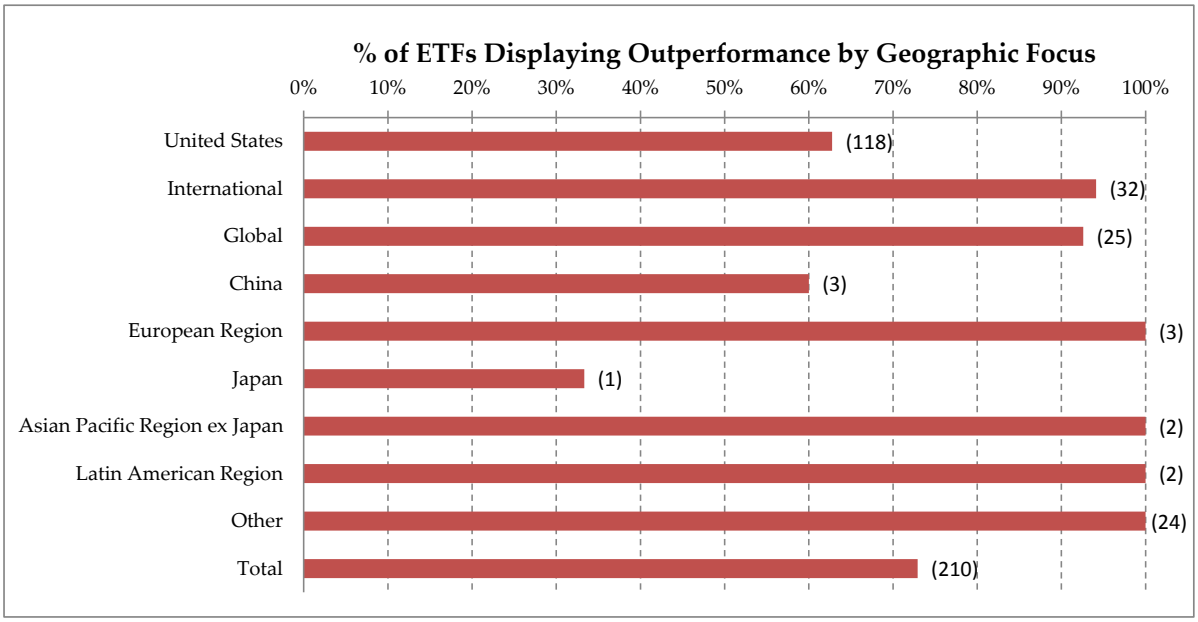
[b]Percentage of ETFs demonstrating specific outperformance measures deemed significant under the balanced stepdown procedure of Romano and Wolf (2010). The figure in brackets gives the fund count in each group.

Figure 1: % of ETFs with Specific Outperformance Measures

	Mean	Standard Dev	Max (ETF Ticker)	Min (ETF Ticker)
Premium	N/A			
Index TE	N/A			
Mkt TE	N/A			
Premium SR	0.02567	0.04385	0.32094 (PCY)	0.00101 (FIW)
Index TE SR	0.02859	0.05158	0.39317 (PCY)	0.00190 (DBC)
Mkt TE SR	0.03228	0.01447	0.06579 (AGG)	0.01090 (PLW)
Premium SorR	0.19738	0.28069	1.88967 (DBS)	0.01190 (SLY)
Index TE SorR	0.21858	0.34545	2.77124 (QLD)	0.01060 (IJH)
Mkt TE SorR	0.25455	0.11153	0.51299 (PCY)	0.07473 (VXF)
Premium TR	0.00751	0.01465	0.08015 (GXC)	0.00002 (RWM)
Index TE TR	0.00861	0.01130	0.06282 (AGG)	0.00001 (IJJ)

[c]Mean (Column 2) refers to the average daily outperformance levels across the 2008-2012 period. Max and Min (Column 4 & 5) identify those ETF tickers which display the highest and lowest aggregated daily outperformance level. All funds are US based with the Bloomberg ticker appendage "US" being omitted for table brevity.

Table 3: Significant Sample Summary Statistics



[d]Percentage of ETFs in each geographic focus category which display at least one significant outperformance measure under the balanced stepdown procedure of Romano and Wolf (2010). The figure in brackets gives the ETF count in each group.

Figure 2: % of ETFs Displaying Outperformance by Geographic Focus

ternational and Other focused funds showing some measure of outperformance. US focused funds on the other hand show a lower proportion of outperformance, although in absolute terms of course the number of funds outperforming is higher at 118. Risk adjusted premium is a primary driver of these results, as seen in Table 4. 63% and 79% of Global and International ETFs respectively, show premium Sharpe Ratio outperformance with only 10% for US funds. These findings are in line with Engle and Sarkar (2006) and Jares and Lavin (2004) who also observe premiums among a high percentage of foreign ETFs and Elton et al. (2002) and Ackert and Tian (2008)who record a low proportion of US focused funds displaying premiums. A lack of synchronization between Net Asset Value calculations and underlying market closes is an oft cited reason for the presence of premiums in ETFs focused over multiple countries/time zones. Further to this, liquidity, latency advantages and reduced market frictions may allow for easier exploitation of deviations among US focused ETFs.

Figure 3 graphs the percentage of ETFs showing some measure of outperformance, split by industry focus this time. Relatively high percentages of Energy, Precious Metals and Real Estate ETFs exhibit outperformance with lower numbers observed for Financial Services ETFs. The high proportion of outperformance observed for these funds are borne out of Market TE Sharpe Ratios, as deduced from Table 5, indicating that 56%, 71% and 50% of Energy, Precious Metals and Real Estate ETFs respectively outperform the market. The two

	United States	International	Global	China	Europe	Japan	Asia Pac	Latin America	Other
	% (Cnt)	% (Cnt)	% (Cnt)	% (Cnt)	% (Cnt)	% (Cnt)	% (Cnt)	% (Cnt)	% (Cnt)
Premium	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)
Index TE	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)
Mkt TE	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)
Premium SR	10% (18)	79% (27)	63% (17)	60% (3)	67% (2)	0% (0)	0% (0)	100% (2)	67% (16)
Index TE SR	11% (20)	44% (15)	63% (17)	20% (1)	0% (0)	0% (0)	0% (0)	50% (1)	25% (6)
Mkt TE SR	27% (51)	0% (0)	7% (2)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	13% (3)
Premium SorR	10% (18)	71% (24)	56% (15)	60% (3)	67% (2)	0% (0)	0% (0)	50% (1)	50% (12)
Index TE SorR	15% (28)	47% (16)	70% (19)	40% (2)	0% (0)	0% (0)	0% (0)	50% (1)	38% (10)
Mkt TE SorR	41% (78)	18% (6)	19% (5)	40% (2)	0% (0)	0% (0)	50% (1)	100% (2)	42% (11)
Premium TR	13% (24)	68% (23)	48% (13)	40% (2)	100% (3)	0% (0)	50% (1)	100% (2)	92% (22)
Index TE TR	14% (27)	82% (28)	74% (20)	20% (1)	100% (3)	33% (1)	0% (0)	100% (2)	83% (20)

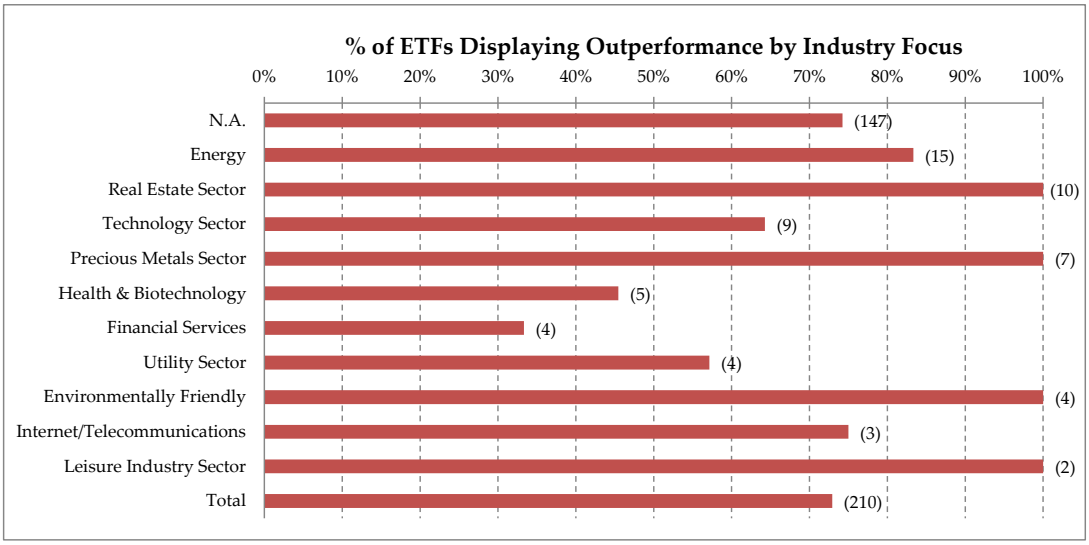
[e]Percentage of ETFs in each geographic focus category which display specific outperformance measures under the balanced stepdown procedure of Romano and Wolf (2010). The figure in brackets gives the ETF count in each group. "Cnt" is an abbreviation of Count and "Asia Pac" is an abbreviation of Asia Pacific excluding Japan, both are used for table brevity.

Table 4: % of ETFs Displaying Specific Outperformance Measures by Geographic Focus

Leisure Industry ETFs in the data set also outperform the market based on Sharpe Ratio. Precious metals became a safe haven for investors due to poor performance in equities over the turbulent 2008-2012 period, with the Energy sector being buoyed by increased manufacturing demand from China. Financial Services in contrast register no ETFs outperforming the market, primarily due to the credit crisis of 2008 and its regulatory legacy.

The next attributes to be analysed are what assets each ETF attempts to replicate and how they conduct the replication. Full replication is the most widely employed strategy in the data set but only 68% of its funds exhibit outperformance, as shown in Figure 4. In comparison, 29 ETFs pursuing derivative replication are seen to display at least one significant outperformance measure, equating to 83% of its sample. Table 6 gives an insight into specifically what outperformance measures are seen in these groups. The main item of note is the presence of significant premium outperformance and absence of significant market outperformance among Optimized ETFs, with 50% of Optimized funds displaying a significant Premium Sharpe Ratio in contrast to just 11% showing Sharpe Ratio market outperformance. An optimized replication strategy involves constructing a portfolio which is a representative subset of the underlying index when full replication of an index's constituents is not possible, be it for cost, liquidity or regulatory reasons. The predominantly illiquid nature of such underlying constituents could be a determining factor in the observation of such redemption in kind inefficiencies.

In relation to asset class, the majority of ETFs in the data set have an equity focus; 263 out of 288 (91%). The prevalence of outperformance is broadly in line with this as seen in



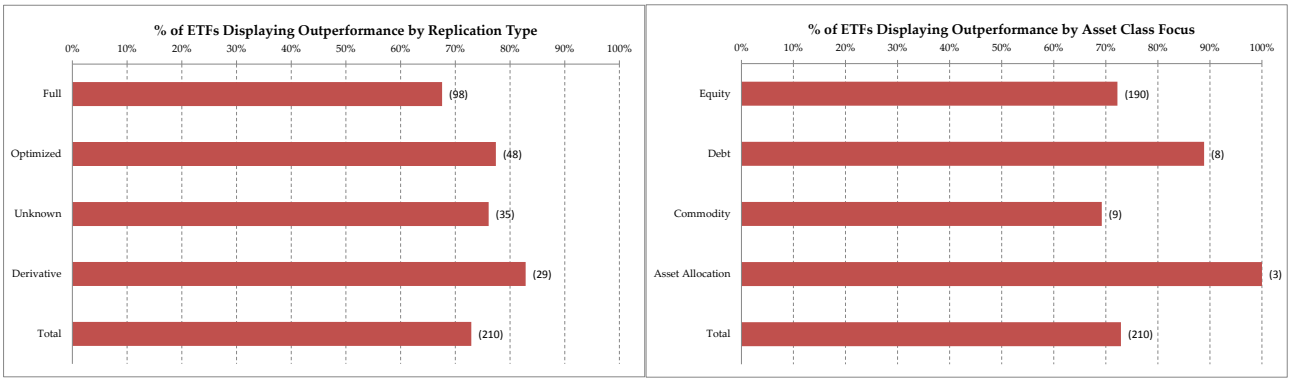
[f]Percentage of ETFs in each industry focus category which display at least one significant outperformance measure under the balanced stepdown procedure of Romano and Wolf (2010). The figure in brackets gives the ETF count in each group.

Figure 3: % of ETFs Displaying Outperformance by Industry Focus

	Energy Sector	Tech Sector	Financial Services	Health & Biotech	Real Estate	Utility Sector	Precious Metals	Environ Friendly	Internet/ Telecoms	Leisure Industry	N.A.
	% (Cnt)	% (Cnt)	% (Cnt)	% (Cnt)	% (Cnt)	% (Cnt)	% (Cnt)	% (Cnt)	% (Cnt)	% (Cnt)	% (Cnt)
Premium	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)
Index TE	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)
Mkt TE	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)
Premium SR	28% (5)	7% (1)	17% (2)	9% (1)	30% (3)	43% (3)	14% (1)	25% (1)	25% (1)	0% (0)	34% (67)
Index TE SR	28% (5)	14% (2)	17% (2)	9% (1)	0% (0)	43% (3)	14% (1)	50% (2)	25% (1)	0% (0)	22% (43)
Mkt TE SR	56% (10)	14% (2)	0% (0)	0% (0)	50% (5)	0% (0)	71% (5)	0% (0)	25% (1)	100% (2)	16% (31)
Premium SorR	17% (3)	7% (1)	17% (2)	18% (2)	10% (1)	43% (3)	14% (1)	25% (1)	25% (1)	50% (1)	30% (59)
Index TE SorR	28% (5)	14% (2)	17% (2)	18% (2)	10% (1)	57% (4)	0% (0)	75% (3)	25% (1)	50% (1)	28% (55)
Mkt TE SorR	61% (11)	50% (7)	0% (0)	36% (4)	60% (6)	14% (1)	100% (7)	0% (0)	50% (2)	100% (2)	33% (65)
Premium TR	6% (1)	14% (2)	17% (2)	18% (2)	40% (4)	43% (3)	0% (0)	25% (1)	25% (1)	0% (0)	37% (74)
Index TE TR	22% (4)	14% (2)	17% (2)	9% (1)	50% (5)	43% (3)	0% (0)	25% (1)	25% (1)	0% (0)	42% (83)

[g]Percentage of ETFs in each industry focus category which display specific outperformance measures under the balanced stepdown procedure of Romano and Wolf (2010). The figure in brackets gives the ETF count in each group. "Cnt" is an abbreviation of Count and "Environ" is an abbreviation of Environmentally, both are used for table brevity.

Table 5: % of ETFs Displaying Specific Outperformance Measures by Industry



[h]Percentage of ETFs in each asset class focus category and also in each replication type category which display at least one significant outperformance measure under the balanced stepdown procedure of Romano and Wolf (2010). The figure in brackets gives the ETF count in each group.

Figure 4: % of ETFs Displaying Outperformance by Replication Type/Asset Class Focus

	Full	Optimized	Derivative	Unknown
	% (Count)	% (Count)	% (Count)	% (Count)
Premium	0% (0)	0% (0)	0% (0)	0% (0)
Index TE	0% (0)	0% (0)	0% (0)	0% (0)
Mkt TE	0% (0)	0% (0)	0% (0)	0% (0)
Premium SR	27% (39)	50% (31)	26% (12)	9% (3)
Index TE SR	21% (31)	19% (12)	17% (8)	26% (9)
Mkt TE SR	21% (30)	11% (7)	22% (10)	26% (9)
Premium SorR	25% (36)	40% (25)	28% (13)	3% (1)
Index TE SorR	24% (35)	35% (22)	22% (10)	26% (9)
Mkt TE SorR	37% (54)	40% (25)	37% (17)	26% (9)
Premium TR	15% (22)	55% (34)	46% (16)	39% (18)
Index TE TR	23% (33)	52% (32)	57% (20)	37% (17)

[i]Percentage of ETFs in each replication strategy which display specific outperformance measures under the balanced stepdown procedure of Romano and Wolf (2010). The figure in brackets gives the ETF count in each group.

Table 6: % of ETFs Displaying Specific Outperformance Measures by Replication Type

	Equity	Commodity	Debt	Asset Allocation
	% (Count)	% (Count)	% (Count)	% (Count)
Premium	0% (0)	0% (0)	0% (0)	0% (0)
Index TE	0% (0)	0% (0)	0% (0)	0% (0)
Mkt TE	0% (0)	0% (0)	0% (0)	0% (0)
Premium SR	28% (73)	23% (3)	78% (7)	67% (2)
Index TE SR	19% (51)	31% (4)	22% (2)	100% (3)
Mkt TE SR	19% (51)	38% (5)	0% (0)	0% (0)
Premium SorR	25% (65)	8% (1)	89% (8)	33% (1)
Index TE SorR	25% (67)	15% (2)	44% (4)	100% (3)
Mkt TE SorR	38% (99)	46% (6)	0% (0)	0% (0)
Premium TR	33% (88)	0% (0)	11% (1)	33% (1)
Index TE TR	38% (99)	0% (0)	22% (2)	33% (1)

[j]Percentage of ETFs in each Asset Class Focus category which display specific outperformance measures under the balanced stepdown procedure of Romano and Wolf (2010). The figure in brackets gives the ETF count in each group.

Table 7: % of ETFs Displaying Specific Outperformance Measures by Asset Class Focus

Figure 4. When looking at the small number of non-Equity ETFs in the sample it can be seen that all of the Asset Allocation focused and almost 90% of the Debt focused funds register significant outperformance measures. A significant Sharpe Ratio Premium is observed for 78% of Debt funds according to Table 7, an indication of redemption in kind inefficiencies.

The final attributes to be examined are the size of the ETF, how much it costs and when it was first traded. Table 8 demonstrates what particular cohorts are most likely to exhibit significant outperformance measures. The results show that ETFs with either high expense ratios or recent inception dates are more likely to display significant outperformance. Table 9 adds an additional layer of granularity to the analysis in showing that the outperformance is primarily due to Index TEs being present, in other words that these funds outperform their underlying indices. The expense ratio result is in line with Harper et al. (2006) and Elton et al. (2002) who find that more expensive ETFs tend to produce greater returns but the difference dissipates once the increased market frictions are accounted for. It is also inferred that larger ETFs have a greater tendency to display significant Premium Sharpe Ratios than smaller ETFs.

To finalise the analysis and provide further insight, the outperformance of individual funds is examined. Table 10 is a list comprising the top 10 funds under each performance measure, compiled and ranked using mean daily outperformance figures. The ETFs in the top 10 for Sharpe and Sortino Ratios across various performance measures highlight the interdependency between these calculations. The distinction between these standard deviation based ratios and the Treynor Ratio which utilizes the CAPM derived β, or correlation between the market and ETF price, as a risk proxy, is apparent when analyzing the cross-measure top

	Assets ($M)	Expense Ratio	Inception Date
Data Set			
Mean	2965.02	0.52%	
Median	421.89	0.51%	15/09/2005
Outperforming ETFs			
Mean	2774.50	0.56%	
Median	429.92	0.52%	01/02/2006
# $\geq$ Data set median (%)	107 (51%)	132 (63%)	121 (58%)
# $<$ Data set median (%)	103 (49%)	78 (37%)	89 (42%)

[k]Rows 2-4 give the mean and median of each attribute for all the ETFs in the data set whereas rows 5-9 give the figures for the subsection of ETFs displaying significant outperformance under the balanced stepdown procedure of Romano and Wolf (2010).

Table 8: Outperformance by Asset/ER/Inception Date

	Total Assets		Expense Ratio		Inception Date	
	$\geq$ $421.89m	$<$ $421.89m	$\geq$ 0.51%	$<$ 0.51%	$\geq$ 15-Sep-05	$<$ 15-Sep-05
	% (Count)	% (Count)	% (Count)	% (Count)	% (Count)	% (Count)
Premium	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)
Index TE	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)
Mkt TE	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)
Premium SR	37% (53)	22% (32)	35% (50)	24% (35)	36% (52)	23% (33)
Index TE SR	34% (49)	18% (26)	27% (39)	25% (36)	28% (41)	24% (34)
Mkt TE SR	8% (11)	6% (8)	10% (14)	3% (5)	9% (13)	4% (6)
Premium SorR	8% (11)	15% (21)	19% (28)	3% (4)	18% (26)	4% (6)
Index TE SorR	29% (42)	24% (34)	33% (47)	20% (29)	33% (48)	20% (28)
Mkt TE SorR	23% (33)	19% (27)	26% (38)	15% (22)	30% (43)	12% (17)
Premium TR	37% (53)	26% (37)	40% (57)	23% (33)	34% (50)	28% (40)
Index TE TR	36% (52)	35% (50)	46% (66)	25% (36)	43% (62)	28% (40)

[l]Percentage of ETFs in each Asset Class Focus/Expense Ratio/Inception Date/Total Assets category which display specific outperformance measures under the balanced stepdown procedure of Romano and Wolf (2010). The figure in brackets gives the ETF count in each group.

Table 9: % of ETFs Displaying Specific Outperformance Measures by Expense Ratio/Inception Date/Total Assets

	Premium SR	Premium SorR	Premium TR	Index TE SR	Index TE SorR	Index TE TR	Mkt TE SR	Mkt TE SorR
1	PCY (0.32094)	PCY (1.88967)	GXC (0.08015)	PCY (0.39317)	PCY (2.77124)	AGG (0.06282)	QLD (0.06579)	DBS (0.51299)
2	LQD (0.18214)	LQD (1.06739)	EWM (0.0721)	LQD (0.13133)	PZA (1.03797)	EWM (0.05754)	DBS (0.05919)	SLV (0.50375)
3	HYG (0.16968)	HYG (1.04502)	AGG (0.06213)	DDM (0.05353)	LQD (0.82621)	HYG (0.04979)	SLV (0.05902)	USD (0.49500)
4	MUB (0.11583)	MUB (0.95198)	EPP (0.05059)	RSU (0.04903)	MUB (0.79061)	EWH (0.04763)	MVV (0.05857)	DIG (0.49411)
5	AGG (0.08780)	EMB (0.43926)	EWH (0.04002)	SSO (0.04833)	DDM (0.43859)	EWT (0.03502)	UWM (0.05600)	QLD (0.49252)
6	EMB (0.05460)	AGG (0.43746)	EWT (0.02934)	MVV (0.04616)	RSU (0.41887)	AIA (0.02783)	DIG (0.05532)	UWM (0.48788)
7	IXJ (0.03775)	PZA (0.40871)	AIA (0.02329)	IXJ (0.03992)	SSO (0.38298)	GMF (0.02546)	SAA (0.05512)	MVV (0.48707)
8	DGS (0.03704)	DGS (0.32111)	EWS (0.02185)	KXI (0.03920)	UUP (0.34833)	ITF (0.02038)	USD (0.05172)	SAA (0.46096)
9	UUP (0.03694)	PBP (0.27846)	DBV (0.01938)	SAA (0.03878)	MVV (0.33571)	EWS (0.02016)	PXI (0.04854)	PXI (0.44491)
10	DEM (0.03583)	IXJ (0.27723)	IYM (0.01879)	GAF (0.03800)	JXI (0.33035)	EEV (0.01875)	PXE (0.04696)	PXE (0.43693)

[m]Significant outperforming ETFs under the balanced stepdown procedure of Romano and Wolf (2010) are ranked in order of the size of their mean daily outperformance measures. The outperformance measure figure is given in brackets. All funds are US based with the Bloomberg ticker appendage "US" being omitted for table brevity.

Table 10: Top 10 by Mean Daily Outperformance

ten ranking composition.

PCY US and LQD US are tickers of particular interest as they appear in the top 3 NAV and index outperformers under both the Sharpe and Sortino Ratio measures. PCY US is the ticker symbol for the PowerShares Emerging Market Sovereign Debt Portfolio which is based on the DB Emerging Market USD Liquid Balanced Index. Its portfolio is comprised of US dollar-denominated government bonds issued by, at present, 64 emerging market countries[14]. It is one of the more recent ETFs in the data set being incepted in October 2007. It follows a full replication strategy at an expense ratio of 0.50%.

LQD US is the ticker symbol for iBoxx $ Investment Grade Corporate Bond Fund which tracks the iBoxx $ Liquid Investment Grade Index. Its portfolio is comprised of liquid, US dollar-denominated, investment grade corporate bonds for sale in the United States. It is a cross-sectoral fund with over 34% currently invested in financial services[15]. Its inception date of the 26th July 2002 is older than the data set median. It also follows a full replication strategy at an expense ratio of 0.15%. This gives an insight into the attribute mix of ETF whose prices *substantially* outperform their NAVs and underlying indices.

DBS US and SLV US are tickers in the top 3 market benchmark outperformers across both Sharpe and Sortino Ratios. DBS is the Powershares DB Silver ETF with SLV the ticker symbol for the iShares Silver Trust. Both funds provide exposure to the market price of silver suggesting that commodity/precious metals *substantially* outperform the market.

[14]http://www.invescopowershares.com/products/holdings.aspx?ticker=PCY (Accessed 30/10/12)

[15]http://us.ishares.com/content/stream.jsp?url=/content/en_us/repository/resource/fact_sheet/lqd.pdf (Accessed 30/10/12)

6 Conclusion

This study seeks to identify ETFs that outperform their calculated NAVs, underlying indices and/or the overall market. Extending the existing ETF literature, an innovative generalised stepwise procedure is used to control for data snooping bias. The balanced stepdown procedure of Romano and Wolf (2010) is applied, serving as an improvement over more conservative single step approaches, such as common techniques like the reality check bootstrap test of White (2000) and the superior predictive ability test of Hansen (2005). Generalised procedures offer greater power to reject false null hypotheses, with the balanced stepdown procedure additionally offering equal treatment in the identification of outperformance. The main item of note from the implementation is that, when performance is analysed on a non-risk adjusted basis only, no funds in the sample are identified as displaying any measure of outperformance. It is only the risk adjusted performance measures that give statistically significant outperformance results and so the insights from these results dominate the commentary.

This paper is the first body of work to test the effect of replication type on performance, finding that 50% of optimized replication ETFs register significant premium Sharpe Ratios. This phenomenon may, in part, be caused by illiquid underlying constituents. This paper is also the first to examine asset class focus, finding that 78% of Debt focused ETFs exhibit significant premium Sharpe Ratios, a figure well above the average and one which gives an indication that Debt focused ETFs are more likely to outperform their NAV than other asset classes. The performance of sectoral ETFs on the other hand has been addressed previously. In this work, Energy, Precious Metals, Real Estate and Leisure are industries which beat the market on a risk adjusted basis. Further to this, precious metal focused funds Powershares DB Silver and the iShares Silver Trust *substantially* outperform the market boasting large mean daily outperformance levels. Precious metals became a safe haven for investors due to poor performance in equities over the turbulent 2008-2012 period, with the Energy sector being buoyed by increased manufacturing demand from China. Financial Services, in contrast, register no market beating funds, primarily due to the credit crisis of 2008 and its legacy.

63% and 79% of Global and International ETFs respectively, show premium Sharpe Ratio outperformance with only 10% for US funds. These findings are in line with Engle and Sarkar (2006) and Jares and Lavin (2004) who also observe premiums among a high percentage of foreign ETFs and Elton et al. (2002) and Ackert and Tian (2008)who record a low proportion of US focused funds displaying premiums. A lack of synchronization between Net Asset Value calculations and underlying market closes is an oft-cited reason for the presence of premiums in funds focused over multiple countries/time zones. Furthermore, liquidity, latency advantages and reduced market frictions allow for easier exploitation of deviations among US focused funds. ETFs exhibiting high expense ratios or recent inception dates have a greater tendency to outperform their index. This expense ratio result is in line

with Harper et al. (2006) and Elton et al. (2002) who find that more expensive ETFs tend to produce greater returns but the difference dissipates once the increased market frictions are accounted for.

This paper succeeds in its stated goals of increasing the understanding of ETF performance alongside providing the wider investment community with an aid in identifying specific ETFs suitable for individual portfolio requirements, along with being of interest to arbitrageurs seeking to exploit the highlighted deviations.

References

Ackert, L., Tian, Y., 2008. Arbitrage, liquidity, and the valuation of exchange traded funds. Financial Markets, Institutions & Instruments 17, 331–362.

Alexander, C., Barbosa, A., 2008. Hedging index exchange traded funds. Journal of Banking & Finance 32, 326–337.

Bajgrowicz, P.G., Scaillet, O., 2009. Technical Trading Revisited: false discoveries, persistence tests, and transaction costs.

Barras, L., Scaillet, O., Wermers, R., 2010. False Discoveries in Mutual Fund Performance: Measuring Luck in Estimated Alphas. Journal of Finance LXV, 179–216.

Criton, G., Scaillet, O., 2011. Time-Varying Analysis in Risk and Hedge Fund Performance: How Forecast Ability Increases Estimated Alpha.

Cummins, M., Bucca, A., 2012. Quantitative spread trading on crude oil and refined products markets. Quantitative Finance , 1–19.

Curcio, R.J., Lipka, J.M., Thornton, J.H.J., 2004. Cubes and individual investors. Financial Services Review, 13, 123–139.

Cuthbertson, K., Nitzsche, D., Sullivan, N.O., 2008. UK mutual fund performance : Skill or luck? Journal of Empirical Finance 15, 613–634. doi:10.1016/j.jempfin.2007.09.005.

DeFusco, R., Ivanov, S., Karels, G., 2011. The Exchange Traded Funds Pricing Deviation: Analysis and Forecasts. Journal of Economics and Finance 35, 181–197.

Elton, E.J., Gruber, M.J., Comer, G., Li, K., 2002. Spiders : Where Are the Bugs ?*. Journal of Business 75, 453–472.

Engle, R., Sarkar, D., 2006. Premiums-Discounts and Exchange Traded Funds. Journal of Derivatives 13, 27–45.

Gruber, M.J., 1996. Another Puzzle: The Growth in Actively Managed Mutual Funds. Journal of Finance 55, 783–810.

Hansen, P., 2005. A test for superior predictive ability. Journal of Business & Economics Statistics 23, 365–380.

Harper, J.T., Madura, J., Schnusenberg, O., 2006. Performance comparison between exchange-traded funds and closed-end country funds. Journal of International Financial Markets, Institutions and Money 16, 104–122. doi:10.1016/j.intfin.2004.12.006.

Hsu, P.h., Hsu, Y.C., Kuan, C.M., 2009. Testing the Predictive Ability of Technical Analysis Using A New Stepwise Test without Data Snooping Bias.

Hsu, P.h., Kuan, C.M., 2005. Re-Examining the Profitability of Technical Analysis with Reality Check.

Jares, T., Lavin, A., 2004. Japan and Hong Kong Exchange-Traded Funds (ETFs): Discounts, Returns, and Trading Strategies. Journal of Financial Services Research 25, 57–66.

Malkiel, B., 1995. Returns from Investing in Equity Mutual Funds 1971 to 1991. Journal of Finance 50, 549–572.

Marshall, B.R., Cahan, R.H., Cahan, J.M., 2008. Can Commodity Futures be Profitably Traded with Quantitative Market Timing Strategies?

Mateus, C., Kuo, T., 2008. The performance and persistence of exchange-traded funds: evidence for iShares MSCI country-specific ETFs, in: Swiss Society for Financial Market Research 11th Conference.

Park, C.H., Irwin, S.H., 2007. What Do We Know About the Profitability of Technical Analysis? Journal of Economic Surveys 21, 786–826. doi:10.1111/j.1467-6419.2007.00519.x.

Phengpis, C., Swanson, P.E., 2009. iShares and the U.S. Market Risk Exposure. Journal of Business Finance and Accounting 36, 972–986.

Plantinga, A., Van der Meer, R., Sortino, F., 2001. The impact of downside risk on risk-adjusted performance of mutual funds in the Euronext Markets.

Qui, M., Wu, Y., 2008. Technical trading-rule profitability, data snooping, and reality check: Evidence from the foreign exchange market. Journal of Money, Credit and Banking 38, 2135–2158.

Romano, J.P., Shaikh, A.M., 2006. On stepdown control of the false discovery proportion. Lecture Notes-Monograph Series , 33–50.

Romano, J.P., Shaikh, A.M., Wolf, M., 2010. Hypothesis Testing in Econometrics. Annual Review of Economics 2, 75–104.

Romano, J.P., Wolf, M., 2007. Control of generalized error rates in multiple testing. Annals of Statistics 35, 1378–1408.

Romano, J.P., Wolf, M., 2010. Balanced Control of Generalized Error Rates. The Annals of Statistics 38, 598–633.

Rompotis, G.G., 2011. Predictable patterns in ETFs return and tracking error. Studies in Economics and Finance 28, 14–35.

Sharpe, W.F., 1966. Mutual Fund Performance. Journal of Business , 119–138.

Sullivan, R., Timmermann, A., 1999. Data-Snooping, Technical Trading Rule Performance and the Bootstrap. Journal of Finance 54, 1647–1691.

White, H., 2000. A reality check for data snooping. Econometrica 68, 1097–1126.

Yu, L., 2005. Basket Securities, Price Formation, and Informational Efficiency.