



“Creating better tracking portfolios with quantiles”

AUTHORS	Mike Aguilar Anessa Custovic Ruyang Chengan Ziming Huang 
ARTICLE INFO	Mike Aguilar, Anessa Custovic, Ruyang Chengan and Ziming Huang (2022). Creating better tracking portfolios with quantiles. <i>Investment Management and Financial Innovations</i> , 19(1), 14-31. doi: 10.21511/imfi.19(1).2022.02
DOI	http://dx.doi.org/10.21511/imfi.19(1).2022.02
RELEASED ON	Monday, 17 January 2022
RECEIVED ON	Monday, 04 October 2021
ACCEPTED ON	Wednesday, 29 December 2021
LICENSE	 This work is licensed under a Creative Commons Attribution 4.0 International License
JOURNAL	"Investment Management and Financial Innovations"
ISSN PRINT	1810-4967
ISSN ONLINE	1812-9358
PUBLISHER	LLC “Consulting Publishing Company “Business Perspectives”
FOUNDER	LLC “Consulting Publishing Company “Business Perspectives”



NUMBER OF REFERENCES

24



NUMBER OF FIGURES

13



NUMBER OF TABLES

5

© The author(s) 2022. This publication is an open access article.



BUSINESS PERSPECTIVES



LLC "CPC "Business Perspectives"
Hryhorii Skovoroda lane, 10,
Sumy, 40022, Ukraine
www.businessperspectives.org

Received on: 4th of October, 2021
Accepted on: 29th of December, 2021
Published on: 17th of January, 2022

© Mike Aguilar, Anessa Custovic,
Ruyang Chengan, Ziming Huang, 2022

Mike Aguilar, Associate Adjunct
Professor at Fuqua Business School,
Duke University, USA. (Corresponding
author)

Anessa Custovic, Quantitative Research
Analyst at Cardinal Retirement
Planning, Inc. in Durham, NC, USA.

Ruyang Chengan, Analyst at
Dimensional Fund Advisors in
Charlotte, NC, USA.

Ziming Huang, Student at the
Economics Department, Duke
University, USA.



This is an Open Access article,
distributed under the terms of the
[Creative Commons Attribution 4.0
International license](https://creativecommons.org/licenses/by/4.0/), which permits
unrestricted re-use, distribution, and
reproduction in any medium, provided
the original work is properly cited.

Conflict of interest statement:
Author(s) reported no conflict of interest

Mike Aguilar (USA), Anessa Custovic (USA), Ruyang Chengan (USA),
Ziming Huang (USA)

CREATING BETTER TRACKING PORTFOLIOS WITH QUANTILES

Abstract

Tracking error is a ubiquitous tool among active and passive portfolio managers, widely used for fund selection, risk management, and manager compensation. This paper shows that traditional measures of the tracking error are incapable of detecting variations in skewness and kurtosis. As a solution, this paper introduces a new class of Quantile Tracking Errors (QuTE), which measures differences in the quantiles of return distributions between a tracking portfolio and its benchmark. Through an extensive simulation study, this paper shows that QuTE is six times more sensitive than traditional tracking measures to skewness and three times more sensitive to kurtosis. The QuTE statistic is robust to various calibrations and can easily be customized. By using the QuTE tracking measure during the Dot Com bubble and the Great Recession, this paper finds differences between the DIA and its benchmark, the DJIA, that otherwise would have gone undetected. Quantile based tracking provides a robust method for relative performance measurement and index portfolio construction.

Keywords tracking error, index tracking, portfolio tracking, index fund, quantile

JEL Classification G11

INTRODUCTION

Index fund managers and risk managers use inadequate tools to track a portfolio's relative performance. The most commonly used tracking error measures are cast as squared deviations between a tracking portfolio and its benchmark, and thus are focused only on the mean and variance of returns. This type of quadratic structure is inconsistent with the linear performance fees through which most managers are compensated (Kritzman, 1987). Instead, managers are incentivized to avoid extreme return deviations (Rudolf et al., 1999), which implies that higher order moments, such as kurtosis, are relevant. Moreover, Beasley et al. (2003) suggest that managers are incentivized to avoid consistently underperforming their benchmark, suggesting that skewness is also relevant.

Doroc'akov'a (2017) and Blume and Edelen (2004) point out that the goal of a tracking error is to measure how closely a portfolio can exactly replicate its associated benchmark. There is a preponderance of evidence that asset returns are non-Gaussian (Mills, 1995; Chung et al., 2006). Therefore, tracking only the first two moments, as do conventional measures, is insufficient.

Other shortcomings of traditional tracking error measures have been cited. For instance, Pope and Yadav (1994) illustrate the bias in the tracking error due to serial correlation in returns. Moreover, Ammann and Tobler (2000) recognize that tracking error variance is subject to the sampling error.

This paper makes two contributions to the literature on portfolio tracking. First, this paper details a previously undocumented shortcoming

of traditional tracking errors. Through a simulation study, this paper shows that traditional tracking errors (such as average tracking error and tracking error volatility) fail to detect situations in which the skewness (and/or kurtosis) of the tracking portfolio differs from that of the associated benchmark.

The second contribution of this paper is to introduce a class of quantile-based tracking errors (QuTE). As this paper will discuss in Section 2.1, there are many variants of the tracking error. Some have symmetric loss functions, structured via absolute or squared deviations. Meanwhile, other variants incorporate asymmetries vis-a-vis semi standard deviations, which are aligned with downside risk. Each have an analogue within the quantile-based measures. This paper shows that even the most basic of these QuTE measures can detect deviations in higher order moments of returns.

This paper begins with a detailed accounting of the traditional measures of tracking error alongside the newly proposed quantile-based measures. Then, the paper conducts an extensive simulation study to explore the relative merits of QuTE. Finally, this paper documents historical episodes where QuTE was able to detect important differences between a tracking portfolio and its benchmark, while the traditional measures were unresponsive.

1. LITERATURE REVIEW

The term “Tracking Error” has evolved over time and is used in myriad contradictory ways by academics and practitioners. To facilitate the discussion, this paper attempts to standardize the terminology and to provide a com-

prehensive list of many variants of the tracking error. Define the price at time t as P_t , and the return from $t - 1$ through t as r_t . Denote r_p as the return on the tracking portfolio, r_b as the associated benchmark, and T as the sample size (e.g., days) over which the portfolio is being tracked.

$$\text{Tracking Error}(TE)_t = r_{P,t} - r_{B,t} , \tag{1}$$

$$\text{Average Tracking Error}(ATE) = \frac{1}{T} \sum_{t=1}^T (r_{P,t} - r_{B,t}) , \tag{2}$$

$$\text{Tracking Error Volatility } (TEV) = \sqrt{\frac{1}{T-1} \sum_{t=1}^T (TE_t - ATE)^2} , \tag{3}$$

$$\text{Tracking Error Risk } (TER) = \sqrt{\frac{1}{T} \sum_{t=1}^T (r_{P,t} - r_{B,t})^2} , \tag{4}$$

$$\text{Root Mean Squared Tracking Error } (RMSTE) = \sqrt{\frac{1}{T-1} \sum_{t=1}^T (TE_t - ATE)^2 + ATE^2} , \tag{5}$$

$$\text{Average Absolute Tracking Error } (AATE) = \frac{1}{T} \sum_{t=1}^T |r_{P,t} - r_{B,t}| , \tag{6}$$

$$\text{Semi Average Tracking Error } (SATE) = \frac{1}{T} \sum_{t=1}^T (r_{P,t} - r_{B,t})_- , \tag{7}$$

$$\text{Semi Tracking Risk } (STR) = \sqrt{\frac{1}{T} \sum_{t=1}^T [(r_{P,t} - r_{B,t})_-]^2} , \tag{8}$$

$$\text{Semi Tracking Volatility } (STV) = \sqrt{\frac{1}{T-1} \sum_{t=1}^T [(TE_t - ATE)_-]^2} , \tag{9}$$

$$\text{Semi Absolute Average Tracking Error } (SAATE) = \frac{1}{T} \sum_{t=1}^T |(r_{P,t} - r_{B,t})_-| , \tag{10}$$

where $(x)_-$ indicates taking only the positive elements of x . This can be annualized by multiplying the above measures by $\sqrt[M]{M}$, where M is the number of periods per year.

Equation (1) was seen first in the academic literature in Franks (1992), which defined it simply “excess of benchmark returns”. Among practitioners, the object in Equation (1) is sometimes referred to as Tracking Difference¹. Roll (1992) refers to this object as “Tracking Error”, which is commonly applied within the proceeding academic literature, and as such reserves that terminology throughout the balance of this paper. Note that the object in Equation (2) is simply an average of the Tracking Error over a period.

The object in Equation (3) is the next most used variant of the term Tracking Error. Franks (1992) refers to this object as Tracking Error, whereas Roll (1992) refers to this as Tracking Error Volatility (TEV). Many proceeding academic studies (Jorion, 2004) use the TEV terminology. Moreover, Equation (3) is commonly referred to as Tracking Error among practitioners². Often this is reported as an annualized value³. Equation (4) is subtly distinct but is less often used in the literature than is Equation (3). Used by Ammann and Tobler (2000), it captures the square root of the sum of the squared tracking error. Root Mean Squared Tracking Error (RMSTE) in Equation (5) was used by Chincarini and Kim (2006) to capture both the variability and the level of the tracking errors.

As noted by Kritzman (1987), portfolio managers are rewarded by linear performance fees based upon the differences between their portfolio and the corresponding benchmark. Rudolf et al. (1999) argue that due to this fact linear deviations between the portfolio and benchmark give a more accurate description of the investors’ risk attitudes than do squared deviations. As such, tracking measures based off absolute, rather than squared differences,

such as those in Equation (6) and Equation (10) are sometimes advocated.

Both the quadratic and absolute measures heretofore are inconsistent with investor loss aversion. Rudolf et al. (1999) advocate the use of semi-variances for downside risk measurement. Equations (7)-(10) reflect this downside risk.

Finally, Beasley et al. (2003) introduce a generalized tracking error written as

$$\left(\frac{1}{T} \sum_{t=1}^T |r_{P,t} - r_{B,t}|^\alpha \right)^{\frac{1}{\alpha}} \quad \text{and semi (downside)} \\ \text{tracking error risk} \quad \left(\frac{1}{T} \sum_{t=1}^T (r_{P,t} - r_{B,t})_-^\alpha \right)^{\frac{1}{\alpha}}.$$

Setting $\alpha = 1$ reproduces AATE and SAATE, while setting $\alpha = 2$ reproduces TER and STR. Also note that the AATE is a special case of the MAPE tracking error of Barro and Canestrelli (2009).

2. METHOD AND SIMULATION STUDIES

This section introduces a class of the tracking error that is based off the difference in the quantiles of the tracking portfolio and respective benchmark, which will be referred to as Quantile Tracking Error (QuTE). After introducing QuTE, this paper explores the differences between QuTE and traditional TE tracking measures by conducting simulation studies. Of particular importance, in subsection 2.2, is the sensitivity of each measure to differences in the empirical distributions of the benchmark and tracking portfolio. Subsections 2.3 and 2.4 focus on robustness of QuTE to various calibrations.

2.1. Method

Set a grid of returns that form $\mathcal{T} - 1$ groups with equal probability of occurring. Then denote $r(\tau)$ to be the τ^{th} $\mathcal{T} -$ quantile of a return distribution.

- 1 See for example, the ESMA https://www.esma.europa.eu/sites/default/files/library/2015/11/2012-832en_guidelines_on_etfs_and_other_ucits_issues.pdf, Morningstar https://media.morningstar.com/uk/MEDIA/Research_Paper/Morningstar_Report_Measuring_Tracking_Efficiency_in_ETFs_February_2013.pdf, and Vanguard <https://www.vanguard.com.hk/documents/understanding-td-and-te-en.pdf>
- 2 CFA Institute <https://www.cfainstitute.org/-/media/documents/support/programs/investment-foundations/19-performance-evaluation.ashx?la=en&hash=F7FF3085AAFADE241B73403142AAE0BB1250B311>, International Organization of Securities Commissions and European Securities and Markets Authority <https://www.iosco.org/library/pubdocs/pdf/IOSCOPD414.pdf>.
- 3 Zephyr <https://www.styleadvisor.com/content/tracking-error>, Vanguard <https://www.vanguard.co.uk/documents/adv/literature/understand-excess.pdf>, Investnet <https://www.investnet.com/sites/default/files/documents/A%20Tracking%20Error%20Primer%20-%20White%20Paper.pdf>.

This paper defines the following tracking error variants inside of the QuTE class,

$$AQuTE = \frac{1}{T} \sum_{\tau \in \mathcal{T}} (r_p(\tau) - r_b(\tau)), \quad (11)$$

$$QuTER = \sqrt{\frac{1}{T} \sum_{\tau \in \mathcal{T}} (r_p(\tau) - r_b(\tau))^2}, \quad (12)$$

$$AAQuTE = \frac{1}{T} \sum_{\tau \in \mathcal{T}} |r_p(\tau) - r_b(\tau)|, \quad (13)$$

$$SAQuTE = \frac{1}{T} \sum_{\tau \in \mathcal{T}} (r_p(\tau) - r_b(\tau))_-, \quad (14)$$

$$SAQuTER = \sqrt{\frac{1}{T} \sum_{\tau \in \mathcal{T}} [(r_p(\tau) - r_b(\tau))_-]^2}, \quad (15)$$

$$SAAQuTER = \frac{1}{T} \sum_{\tau \in \mathcal{T}} |(r_p(\tau) - r_b(\tau))_-|, \quad (16)$$

Intuitively, QuTE compares two assets via differences in the quantiles of their respective return distributions. This is especially useful in finance, given the preponderance of returns with excess skew and kurtosis, and quantile-based methods' ability to capture these distributions (Rostek, 2010). Moreover, a quantile-based approach is consistent with the utility maximization via quantile maximization of Rostek (2010), as well as with Giovannetti (2013), who builds an asset pricing model consistent with CRRA preferences via quantile maximization.

Since the Value-at-Risk (VaR) is merely a quantile of a return distribution, QuTE can be seen as matching on the space of VaRs at various levels. Yamai and Yoshida (2002) show that portfolio ranking via VaR is consistent with expected utility maximization and is free of tail risk. This paper adapts the findings of Rostek (2010), who characterizes the behavior of an agent evaluating different (investment) alternatives by the τ^{th} quantile of the implied (return) distributions and selects the one with the highest quantile payoff. Investor's preferences can be represented via the quantiles of the associated return distribution. In the context of benchmark tracking, the investor's preferences

for deviations from their benchmark can be cast via the differences in the quantiles of the portfolio and benchmark. Portfolio construction with VaR based objective functions is increasingly common (see Gaivoronski & Pflug (2005) for recent examples). Moreover, a quantile-based approach⁴ is especially attractive, given the prevalence of VaR for portfolio risk management (Follmer & Leukert, 1999). Notice the similarities with the tracking error measures defined in Section 1. Importantly, the averaging in the QuTE class is not done over time T , but rather across quantile levels \mathcal{T} . The QuTE measures never force portfolio managers to compare his/her portfolio to the benchmark on a daily basis⁵. This might mitigate the problem of "short termism" as indicated by Ma et al. (2019). Specifically, short evaluation periods for performance-based compensation may damage fund performance by incentivizing managers to engage in activities such as risk shifting and window dressing to boost short-term performance.

Beasley et al. (2003) expand their tracking error to accommodate for the case where someone might want to weigh the importance of the return deviations differently over time. Analogously, this paper introduces a quantile weighted version of QuTE. The case of QuTER is illustrated below, but this approach can easily be extended to any of the measures within the QuTE family:

$$\sqrt{\frac{1}{T} \sum_{\tau \in \mathcal{T}} \lambda(\tau) (r_p(\tau) - r_b(\tau))^2}, \quad (17)$$

where $\lambda(\tau)$ is the importance of quantile τ to the overall tracking error measure. Beasley et al. (2003) do not discuss weighting schemes, but given they are directing the weightings over time, any of the numerous time series lag function might suffice (Almond, etc...). In this paper, the importance weights are linked to the area of the return distribution the user finds most important. Analogous to choosing the quantile level for risk buffers in Basel (e.g., 5% VaR), the importance of specific quantiles can be designated. For tractability and interpretation, this paper recommends scaling such that

4 Note that a natural analogue to QuTE is moment-based matching, rather than quantile based. One could use a method of moments type estimator to match a select set of empirical moments between the benchmark and optimal portfolio. Although potentially attractive, a moment-based approach lacks the flexibility of a nonparametric quantile-based method.

5 The one-to-one mapping between returns and quantile levels permits leveraging the distribution matching literature and cast QuTER within the Fidelity family of similarity measures.

$$\sum_{\tau \in \mathcal{T}} \lambda(\tau) = 1 \quad (18)$$

Section 2.4 offers two approaches to scaling: equal quantile weight and total return attribution.

2.2. Sensitivity to differences in return distributions

In this subsection, a simulation study is conducted to evaluate the traditional tracking error measures of Section 1, as well as the QuTE based measures of Section 2.1. A toy exercise is crafted that, while simple in nature, permits highlighting the sensitivity of the tracking errors to differences in the underlying return distributions. Given the preponderance of evidence citing skewness and kurtosis (see Chung et al. (2006) and Mills (1995), among others) in asset returns, coupled with the calls for linear performance measures (Rudolf et al., 1999; Kritzman, 1987), this paper considers deviations in these “higher order” moments.

The simulations begin by creating a benchmark portfolio. For simplicity, assume the returns of the benchmark follow a standard Normal distribution. Then, calibrate the length and empirical moments of the benchmark to match that of the monthly returns on Dow Jones Industrial Average over the period 1985 through June 2021. This same index is used in a Case Study detailed in Section 3.1. The simulations contain 10,000 paths, each of length 438 months.

Next, generate a tracking portfolio that follows one of five distinct distributions depicted in Table 1. In Case 0, the tracking portfolio has the same distribution as the benchmark portfolio. In Case 1, they differ only in the mean. Similarly, Case 2 varies in terms of variance, Case 3 in terms of skewness, and Case 4 in terms of kurtosis⁶.

Table 1. Simulation study design

Case	Mean	Standard Deviation	Skewness	Kurtosis
Case 0	(0, 0)	(1, 1)	(0, 0)	(3, 3)
Case 1	(0, 0.75)	(1, 1)	(0, 0)	(3, 3)
Case 2	(0, 0)	(1, 4.40)	(0, 0)	(3, 3)
Case 3	(0, 0)	(1, 1)	(0, -1.09)	(3, 3)
Case 4	(0, 0)	(1, 1)	(0, 0)	(3, 7.11)

All return series are generated from a flexible Pearson distribution. Each cell contains the moments for the (Benchmark, Tracking) portfolios. The value 0.75, 4.40, -1.09 and 7.11 in parenthesis are the mean, standard deviation, skewness, and kurtosis of monthly returns on Dow Jones Industrial Average over the period 1985 through June 2021, respectively.

This exercise explores the ability of the various traditional tracking measures to detect differences in the mean (standard deviation, skewness, kurtosis) of the tracking portfolio and benchmark. As noted in Section 1, the TEV depicted in Equation (3) is the most used tracking measure among academics and practitioners. The TEV is compared to ATE, TER, and RMSTE⁷.

First, vary the mean return of the tracking portfolio in excess of the benchmark (i.e., excess mean) in the range $S \in \{-5\% \text{ to } 5\%\}$ ⁸. Next, compute the ATE, TER, RMSTE and TEV for each of these values of excess mean by taking the average over simulation paths. Finally, scale⁹ the values for each of the cases for ease of visual comparison. Panel A of Figure 1 depicts the ATE, TER, RMSTE and TEV values over the range of excess mean values. Panels B, C, and D similarly reflect excess standard deviation, skewness, and kurtosis.

Figure 1 plots the scaled ATE, TER, RMSTE and TEV statistics for a range of means, standard deviation, skewness, and kurtosis. Panel A plots the scaled ATE, TER, RMSTE and TEV statistics for excess mean. Panel B plots the scaled ATE, TER,

6 Each series was simulated within Matlab using the `pearsrnd` function for a Pearson system of random numbers with moments calibrated to match the mean, standard deviation, skewness, and kurtosis of the monthly return of the Dow Jones Industrial Average over the period 1985 through June 2021.

7 The measures of absolute and semi tracking error are beyond the scope of this paper.

8 This paper also considers excess standard deviation in the range -0.9 to 4, excess skewness in the range -1.4 to 1.4, and excess kurtosis in the range -1.5 to 4.5.

9 This paper scales as follows: $[\text{Tracking Measure Value} - \min(\text{Tracking Measure Value})] / [\max(\text{Tracking Measure Value}) - \min(\text{Tracking Error Value})]$.

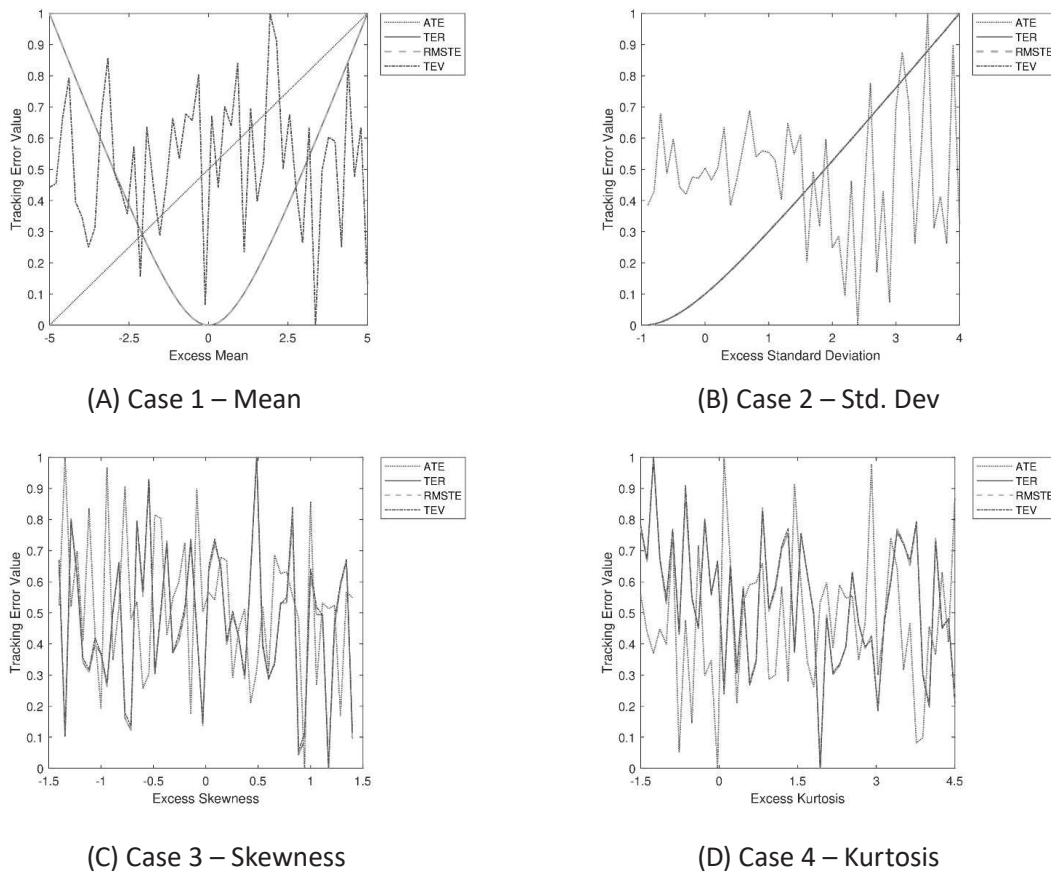


Figure 1. Scaled ATE, TER, RMSTE and TEV sensitivity plots

RMSTE and TEV statistics for the excess standard deviation. Panel C plots the scaled ATE, TER, RMSTE and TEV statistics for the excess skewness. Panel D plots the scaled ATE, TER, RMSTE and TEV statistics for the excess kurtosis. Note that for some cases the tracking measures may be visually indistinguishable on the plot.

A desirable measure of tracking error should achieve a minimum at an excess mean (standard deviation, skewness, kurtosis) of 0, i.e., when there is no difference between the tracking portfolio and benchmark, the tracking error measure should be at its low point. According to Figure 1, ATE is unable to detect changes in any of the four moments. Meanwhile, TEV performs similarly to TER and RMSTE across Cases 2 through 4. In this sense, TEV is roughly equivalent to TER and RMSTE.

Next, compare the traditional and quantile-based tracking measures in terms of their abilities to detect differences in the underlying statistical distributions of the benchmark and tracking portfolio.

The comparison is centered around the TER of Equation (4) and the QuTER of Equation (12). Note the prior findings that TER is roughly equivalent to the popular TEV, which makes this comparison relevant. Moreover, QuTER is a direct analogue of TER, providing a fair comparison.

Table 2 explores these relative sensitivities by computing the percent change in the (Qu)TER statistic relative to Case 0. The greater is the percent change in the (Qu)TER in Case 1 relative to Case 0, the more sensitive is that measure to variations in the means of the two series.

Table 2. Sensitivity of TER and QuTER

Tracking Measure	Case 1	Case 2	Case 3	Case 4
QuTER	613	3124	336	106
TER	13	219	0.10	0.06
PVal	< 0.01	< 0.01	< 0.01	< 0.01

Table 2 reports the sensitivity of TER and QuTER to variations in the distributions of the tracking portfolio and benchmark. Each cell represents the

percent change in the associated tracking measure relative to Case 0, averaged over 10,000 simulated paths. The row labeled PVal reports the p-value from a two-tailed test of equal means.

The p-value of 0 for Case 1 in Table 2 implies that the percent change in the QuTER statistic for Case 1 relative to Case 0 is not equal to the percent change in the TER statistic for Case 1 relative to Case 0. In fact, QuTER and TER have unequal sensitivities to differences in each of the first four statistical moments. Moreover, one-tailed t-tests suggest that the QuTER is in fact more sensitive than TER in all Cases.

These findings are explored further by conducting a sensitivity analysis similar to the exercise above. Again, vary the degree of mean returns in the tracking portfolio in excess of the benchmark (i.e., excess mean) in the range $S \in \{-5\% \text{ to } 5\%\}^{10}$. Next, compute the TER and QuTER for each of these values of excess mean, simulated and averaged over 10,000 paths. Finally, scale the values for each of the cases for visual comparison. Panel A of Figure 2 depicts the TER and QuTER values over the range of excess mean values. Panels B, C, and D similarly reflect excess standard deviation, skewness, and kurtosis. Again, a desirable measure of the tracking error should achieve a minimum at an excess mean (standard deviation, skewness, kurtosis) of 0, i.e., when there is no difference between the tracking portfolio and benchmark, the tracking error measure should be at its low point. The values for each of the cases for visual comparison. Panel A of Figure 2 depicts the TER and QuTER values over the range of excess mean values. Panels B, C, and D similarly reflect excess standard deviation, skewness, and kurtosis. Again, a desirable measure of the tracking error should achieve a minimum at an excess mean (standard deviation, skewness, kurtosis) of 0, i.e., when there is no difference between the tracking portfolio and benchmark, the tracking error measure should be at its low point.

Panel A of Figure 2 suggests that TER and QuTER are both sensitive to variations in the mean return of the tracking portfolio and benchmark. They each reach minimum values near 0 excess mean and rise at values above and below that amount. Similarly,

Panel B illustrates that both TER and QuTER appear sensitive to deviations in excess standard deviation. However, Panels C and D illustrate that TER is not sensitive to deviations in skewness or kurtosis. Meanwhile, QuTER continues to respond to these excess variations. Figure 2 also illustrates that the QuTER values reach minima precisely when one might hope, i.e., when there is no deviation between the mean/variance/skewness/kurtosis of the tracking portfolio and the benchmark. These findings are consistent for ATE/AQuTE, AATE/AAQuTE, and ATR/AQuTER.

Figure 2 plots the scaled TER and QuTER statistics for a range of means, standard deviation, skewness and kurtosis. Panel A plots the scaled TER and QuTER statistics for excess mean. Panel B plots the scaled TER and QuTER statistics for the excess standard deviation. Panel C plots the scaled TER and QuTER statistics for the excess skewness. Panel D plots the scaled TER and QuTER statistics for the excess kurtosis.

To facilitate a statistical comparison between TER and QuTER, this paper conducts a regression that projects the standardized tracking errors upon the absolute moment's differences (excess moment) in the tracking portfolio and benchmark. An error group dummy variable and an interaction term with the moment difference are added to the regression to explore whether there are differences between the different tracking errors.

Consider Case 3 as an illustration. Define the response variable as $Y_i = z\text{-score}(\text{Error}_i)$ for $i \in S$, where Error_i is the QuTER or TER value. The tracking errors are standardized within each error group to ease comparison. For example,

$$\begin{aligned} z\text{-score}(\text{QuTER}_i) &= \\ &= \frac{\text{QuTER}_i - \text{mean}(\text{QuTER})}{\text{std}(\text{QuTER})}. \end{aligned} \tag{20}$$

Now, define the excess moment, excess skewness here, as $Em_i = |\text{skew}_i^P - \text{skew}_i^B|$ for $i \in S$. The excess mean, standard deviation, and kurtosis are all defined analogously. The error group dummy D_{QuTER} is set to 1 if the QuTER is used to measure the tracking error.

10 This paper also considers excess standard deviation in the range -0.9 to 4, excess skewness in the range -1.4 to 1.4, and excess kurtosis in the range -1.5 to 4.5.

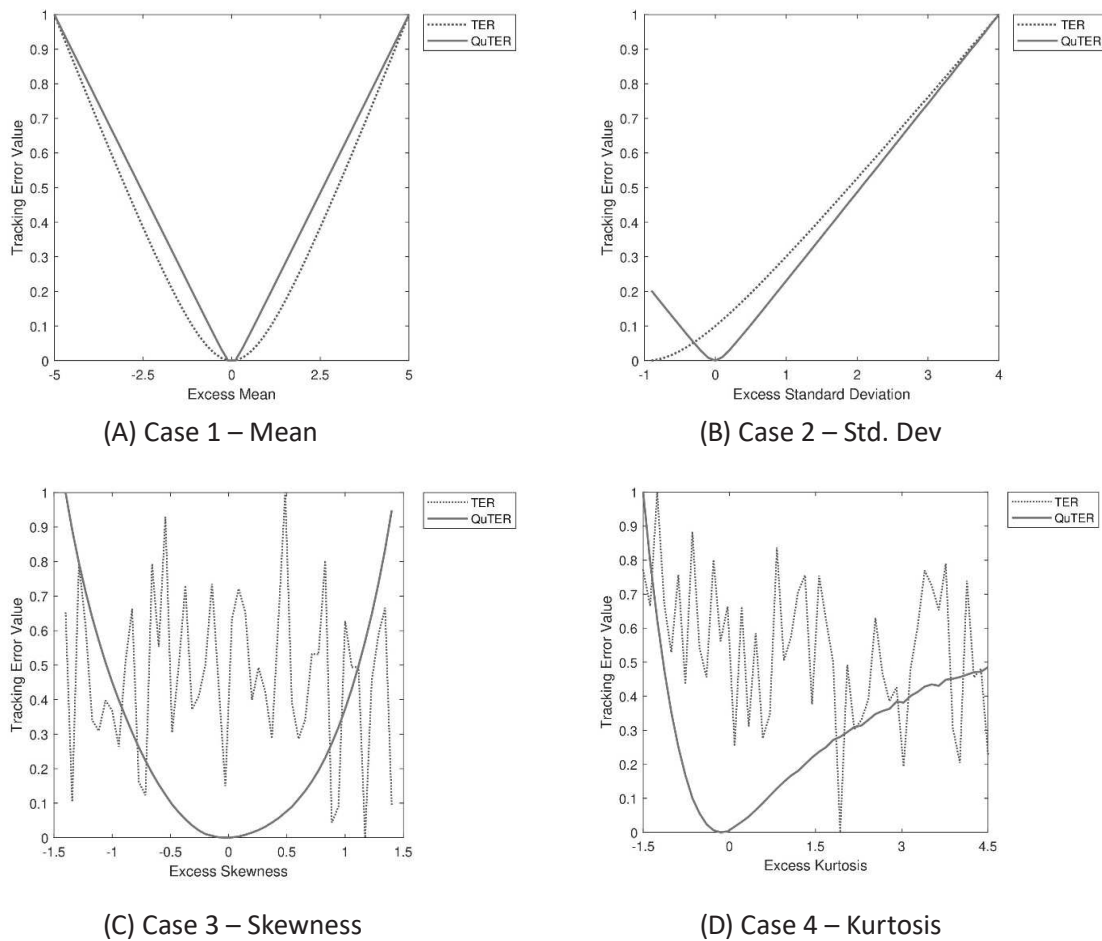


Figure 2. Scaled TER and QuTER sensitivity plots

The regression is specified as follows, $Y_i = \alpha + \beta_1 Em_i + \beta_2 D_{QuTER_i} + \beta_3 D_{QuTER_i} Em_i + e_i$, with typical assumptions on the error term. The object of interest is testing if $\beta_3 = 0$, which would imply that the tracking error measures behave similarly as the excess moment rises. Further, the directionality can be gauged from the sign of the estimated coefficient. For instance, a positive β_3 from the skewness regression (Case 3) would imply that QuTER is more sensitive than TER to variations in skewness between the tracking portfolio and the benchmark.

Regression results of the standardized QuTER and TER upon absolute excess statistical moments across each of the 100 percentiles. Case 1 captures excess mean as $|mean_i^P - mean_i^B|$. Case 2 captures excess standard deviation as $|stddev_i^P - stddev_i^B|$. Case 3 captures excess skewness as $|skew_i^P - skew_i^B|$. Case 4 captures excess kurtosis as $|kurtosis_i^P - kurtosis_i^B|$. Table entries refer to the slope estimate averaged across 10,000 paths.

Table 3 demonstrates that the estimated β_3 is positive and statistically significant for Cases 3 and 4.

Table 3. QuTER and TER regression

Case	β_1			β_3		
	Estimate	SE	t-stat	Estimate	SE	t-stat
Case 1	0.669	0.007	89.756***	0.004	0.011	0.376
Case 2	0.793	0.018	44.787***	0.019	0.025	0.746
Case 3	0.396	0.260	1.520	1.956	0.368	5.310***
Case 4	-0.379	0.128	-2.961***	1.074	0.181	5.935***

Note: *** significant at $p < 0.001$.

Specifically, the QuTER statistic is roughly six times more sensitive than TER to deviations in skewness and three times more sensitive than TER to deviations in kurtosis. This finding aligns with Figure 2, where QuTER appears to detect changes in the third and fourth moment, while TER is unable to do so.

2.3. Robustness to quantile grid granularity

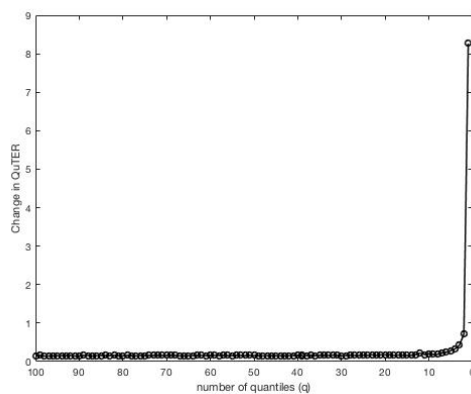
This subsection explores whether the granularity of the quantile grid for the QuTE statistics impacts their ability to detect differences in the distributions of the tracking portfolio and the benchmark.

The exercise of Section 2.2 is repeated by simulating the benchmark returns as simple Gaussian noise and then varying the tracking portfolio in four ways, Case 1 alters the mean, Case 2 alters the variance, Case 3 alters the skewness, and Case 4 alters the kurtosis. Figure 3 depicts the percent-

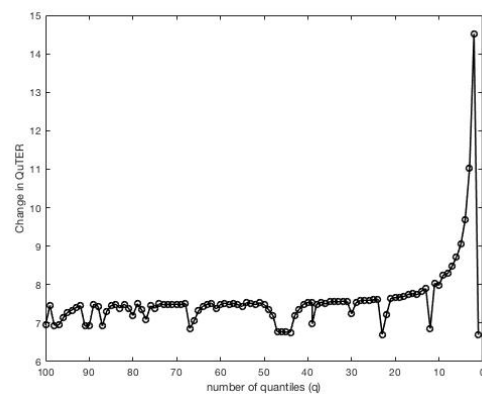
age change in the QuTER statistic in a given Case relative to Case 0. The x-axis varies the size of the quantile grid (\mathcal{T}). The reported values are the median across 10,000 simulated paths.

The percentage change in the QuTER statistic falls as the number of quantiles in the grid rises. The relationship appears to plateau near 10 quantiles, indicating that the QuTER measure is robust to the choice of quantile grid.

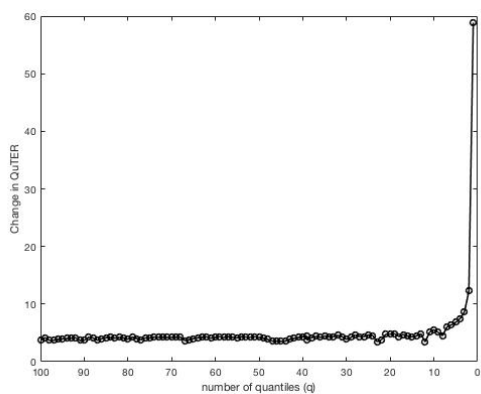
Figure 3 plots the percent change in QuTER over the base case as decreasing the number of quantiles from 100 evenly spaced quantiles (percentiles) to 1 quantile (median). Panel A depicts the change in QuTER over the base case when alters the mean. Panel B depicts the change in QuTER over the base case when alters the standard deviation. Panel C depicts the change in QuTER over the base case when alters the skewness. Panel D depicts the change in QuTER over the base case when alters the kurtosis.



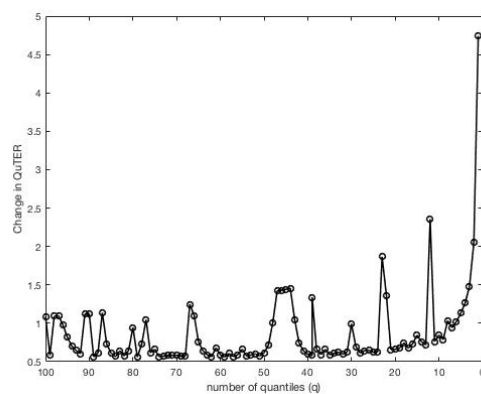
(A) Case 1 – Mean



(B) Case 2 – Std. Dev



(C) Case 3 – Skewness



(D) Case 4 – Kurtosis

Figure 3. Granularity of grid for the QuTER statistic

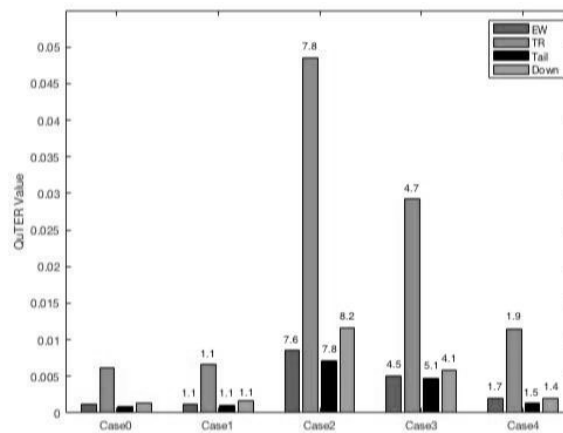


Figure 4. Effect of varying weights on QuTER

2.4. Impact of varying quantile weights

This section explores whether variations in the quantile weighting scheme affect QuTE's ability to detect deviations between the distributions of the tracking portfolio and benchmark.

Blitz and Hottinga (2001) illustrate how to compare various investment strategies via a Tracking Error framework. In a similar vein, various quantiles are weighted by whatever criterion is most important to the investor. Four different weighting schemes are considered: equal weight, tail risk weight, downside risk weight, and total return attribution.

For the equal weight scheme, each quantile has equal importance. For the tail risk weighting scheme, set $\lambda = 0$ for quantiles 1-5% and quantiles 95-100% and $\lambda = 1/90$ for all other quantiles. For the downside risk weighting scheme, set λ equal among all quantiles with downside deviations. This scheme is inspired by loss aversion ala Kahneman and Tversky (1979), and is closely connected to the Semi-Standard Deviation based (quantile) tracking errors. Finally, consider a total return attribution weighting scheme, wherein each quantile is weighted according to its contribution to the portfolio's total return. Specifically, the relative frequency of return observations that fall within that bin is computed. Then, take the

average bin return times relative frequency and divide by the total portfolio return¹¹ to compute the attribution of any given bin. By design, these attributions sum to 1, and thus are viable choices for quantile weights λ .

Figure 4 illustrates how the QuTER objective function varies with the four aforementioned weighting schemes using the structure from Section 2.2. The height of each bar is the associated QuTER averaged over 10,000 paths. The number above each bar is the gross change of that average QuTER statistic relative to Case 0. For instance, the 1.1 above the first bar in Case 1 implies that the QuTER value for the equal weight scheme in Case 1 is 1.1 times as large as the equal weighting scheme QuTER statistic for Case 0. The legend can be read as follows: EW = Equal Weight, TR = Total Return Attribution, Tail = Tail Risk, and Down = Downside Risk.

Figure 4 reports the value of QuTER in five cases: Case0 – tracking and benchmark portfolio come from the same distribution; Case1 – means differ; Case2 – variances differ; Case3 – skewness differs; and Case4 – kurtosis differs. See Section 2.2 for details. The height of each bar marks the QuTER value averaged over 10,000 simulated paths. The number on top of each bar represents the gross change of that QuTER value relative to Case0.

A quantile weighting scheme of equal weight or total return attribution is robust to a wide array of

¹¹ More precisely, divide by the sum of the average bin returns times relative frequencies. Due to the averaging across the bins, this value may not be equal to the actual portfolio return in any given dataset but will approach that value as the distance between the grid points approach 0.

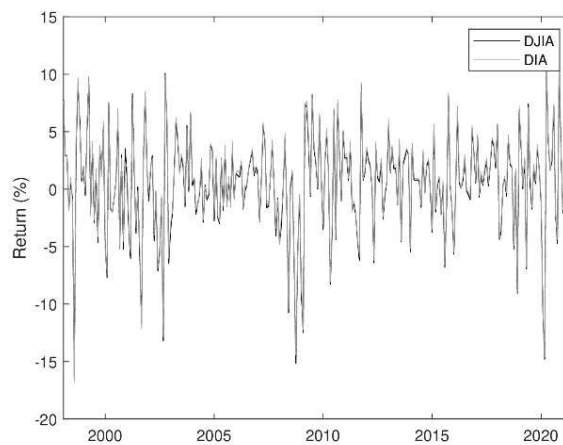


Figure 5. DIA and DJIA returns

differences in the underlying return distributions of the benchmark and tracking portfolio¹².

3. EMPIRICAL RESULTS AND DISCUSSION

In this section, two small case studies are conducted to illustrate the behavior of QuTE alongside a traditional measure of tracking error. The first case regards tracking the DJIA, while the second focuses on tracking the MSCI Emerging Markets Index. QuTER and TER measures are applied in both an unconditional and conditional setting.

3.1. Tracking the DJIA

The Dow Jones Industrial Average (DJIA) is the benchmark and the DIA SPDR ETF is the tracking portfolio. The DJIA is a leading index of equity market returns in the USA, launched on May 26, 1896, and with approximately 1,876.70 dollars indexed to its performance. The DIA is among the largest of the DJIA ETF tracking portfolios, with an average of 6,912,000 USD in daily volume since the inception date. It is also one of the oldest ETFs

to track the DJIA portfolio, with an inception date of January 13, 1998¹³.

The dataset contains monthly log returns for both the DJIA (benchmark) and the DIA (tracking portfolio) over the period January 1998 to June 2021. Figure 5 depicts the time variation of these two return series overlaid upon one another. Simple visual inspection suggests they are quite similar. In fact, the correlation between these two return series is almost 1.

Figure 5 displays the relationship between the DIA and the DJIA monthly returns from January 1998 to June 2021.

Table 4 contains basic descriptive statistics such as mean, standard deviation, skewness, and kurtosis, as well as select quantiles of these two series.

Table 4 reports the statistical comparison of the DJIA benchmark and the DIA tracking portfolio. Monthly log returns from January 1998 through June 2021 are used. PVal is the p-value from tests of equal moments and quantiles (Harrel & Davis, 1982). The KS test is displayed

Table 4. Statistical comparison of DJIA and DIA

Asset	Mean	Std. Dev	Skewness	Kurtosis	q5	q10	q25	q50	q75	q90	q95
DIA	0.70	4.38	-0.75	4.72	-6.70	-4.86	-1.50	1.04	3.13	5.87	7.50
DJIA	0.52	4.39	-0.75	4.69	-6.85	-5.04	-1.59	0.82	3.04	5.62	7.43
PVal	0.63	0.98		0.92	0.87	0.75	0.79	0.54	0.76	0.76	0.82

12 These findings are similar for AQuTE and AAQuTE.

13 The adjusted close prices are from Yahoo Finance.

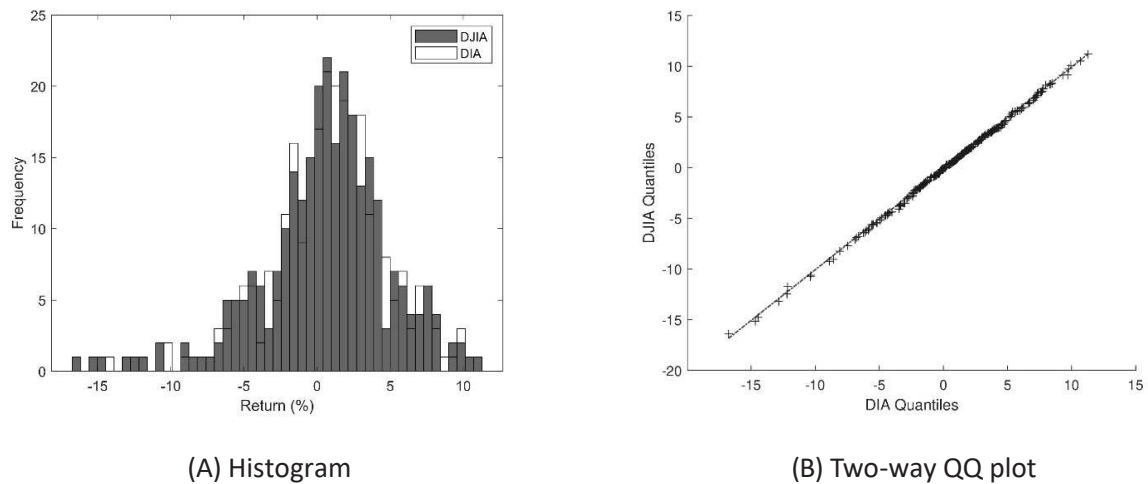


Figure 6. Comparing the benchmark and tracking portfolio

for the skewness and kurtosis columns and the general quantile comparison test (Wilcox et al., 2014) is used for the quantiles.

Figure 6 complements the comparisons in Table 4 by overlaying histograms of the tracking portfolio and benchmark in Panel A, and presenting a two-way QQ plot in Panel B. In addition, Table 5 presents various measures of (quantile) tracking errors. Note that the TE and QuTE values are not directly comparable, given the different scaling of each measure.

Figure 6 compares the statistical distributions of the DIA and DJIA. Panel A plots the histogram of DIA and DJIA returns. Panel B plots a two-way QQ plot of the DIA and DJIA returns. The sample period is from January 1998 through June 2021.

Table 5. (Quantile) Tracking errors

Tracking Type	TE	Value	QuTE	Value
Average	ATE	0.1789	AQuTE	0.1712
Risk	TER	0.2916	QuTER	0.2006
Volatility	TEV	0.2307	NA	–
Root Mean Square	RMSTE	0.2919	NA	–
Avg. Absolute	AATE	0.2370	AAQuTE	0.1796
Absolute Risk	ATR	0.2916	AQuTER	0.2006
Absolute Volatility	ATV	0.4015	NA	–

Table 5 reports the (Quantile) tracking errors for the DJIA (benchmark) and the DIA (tracking portfolio). Monthly simple returns from January 1998 through June 2021 are used.

Taken together, the above results reveal that the DIA has distributional properties that are remarkably similar to the DJIA, thereby supporting the visual inspection. Each of the moments and quan-

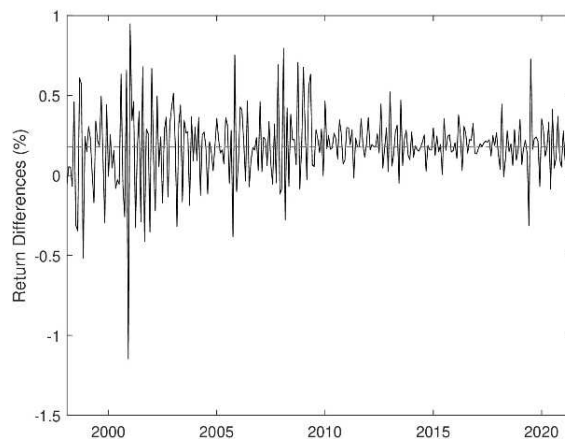


Figure 7. Return differences (TE) of DIA and DJIA

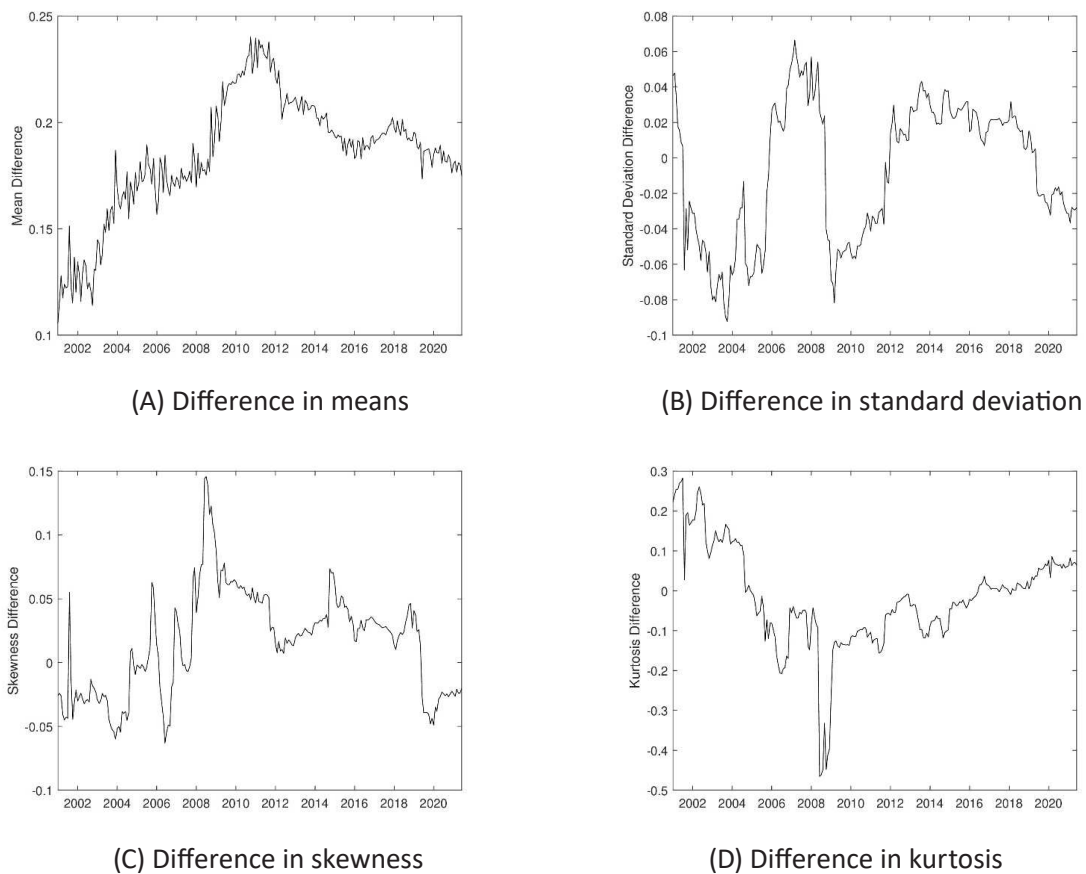


Figure 8. Difference in 3-year moments rolled through time for DIA and DJIA

tiles examined are statistically identical across the two portfolios.

Nonetheless, the two series can differ in ways that are important to portfolio managers and investors. Figure 7 charts the difference in returns (TE) for each month. Deviations between the two series are particularly visible during the aftermath of the Dot Com bubble in 2001 and during the Great Recession of 2008–2010. Of particular note is the variability in the TE over time. Figure 8 depicts the time variation in the difference in the first four moments of the tracking portfolio and benchmark.

Figure 7 plots the relationship of the DIA (tracking portfolio) and the DJIA (benchmark) return differences (TE) over time. Monthly returns from January 1998 to June 2021 are used.

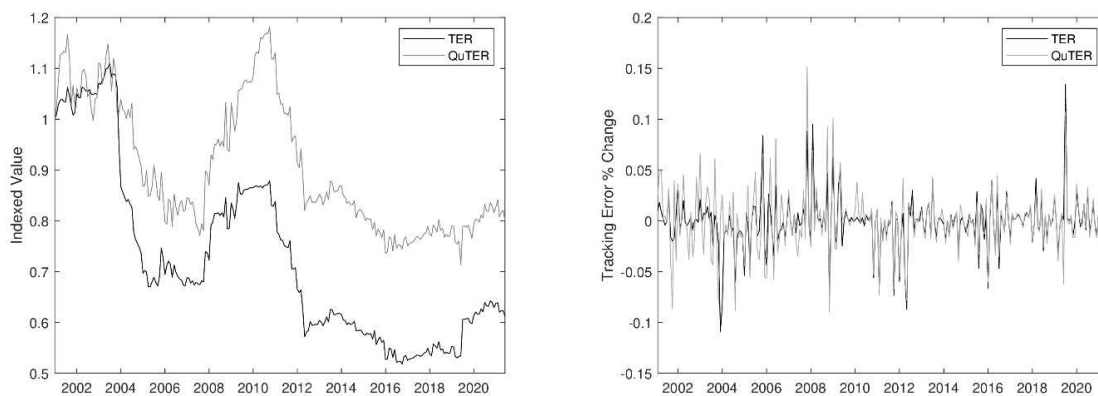
Figure 8 shows the differences in the first four moments of the DIA and DJIA. Panel A plots the mean differences. Panel B plots the standard deviation differences. Panel C plots the skewness

differences. Panel D plots the kurtosis differences. All differences in moments were computed over 3-year rolling windows. The sample period is from January 1998 to June 2021.

In a similar fashion, TER and QuTER statistics are computed between the benchmark and tracking portfolio. Panel A of Figure 9 depicts the rolling tracking measures computed over rolling three-year windows, while Panel B depicts the month-to-month percent change in each tracking measure.

Figure 9 plots the TER and QuTER rolling tracking measures. Panel A plots the indexed value of the TER measure and the indexed value of the QuTER measure. Panel B plots the monthly percent change of the calculated rolling TER and QuTER measures. All differences in moments were computed over 3-year rolling windows. The sample period is from January 1998 to June 2021.

According to Figure 7, there is a large spike in the TE during 2001, followed by small fluctua-



(A) Indexed TER and QuTER

(B) TER and QuTER monthly percent change

Figure 9. TER and QuTER (3-year rolling window)

tions of the TE before 2003. Panel A of Figure 8 shows that the differences in mean returns between DJIA and DIA were relatively small and steady in 2002, while Panel D shows big differences and fluctuations in kurtosis. The TER is steady near 1.05 during this period, while the QuTER rises from 1.02 to 1.10 in the first half of 2002, then falls back down to 1 by October 2002. These movements in the QuTER reflect its sensitivity to differences in return distributions that were not detected by TER.

Another episode of interest is the Great Recession. The TE swings wildly from 0.72 to 0.83 over the period 2008 to 2010. The mean return differences, as depicted in Panel A of Figure 8, vary between 0.17 and 0.24, and with it TER varies between 0.72 to 0.88. Notice that skewness changed from 0.04 to 0.05 and kurtosis from -0.08 to -0.11 over that period¹⁴. QuTER captured these movements, by increasing by almost 29 percent over that period, rising from 0.87 to almost 1.12, outpacing the roughly 15% change in TER.

3.2. Tracking the MSCI Emerging Markets Index

In this case study, the MSCI Emerging Markets Index (MSCI-EM) is used as the benchmark and the iShares MSCI Emerging Markets ETF (EEM) is used as the tracking portfolio. A recent episode

is carefully investigated to exemplify the differences between TER and QuTER. The dataset consists of monthly simple returns over the period January 2013 through June 2021¹⁵.

The correlation between these two return series is 0.97 during this sample period. As depicted in Figure 10, the empirical distributions are similar. Nonetheless, as depicted in Figure 11, there are differences between the two series. Analogous to Figure 8, Figure 12 illustrates the time variation in the differences of the first four empirical moments of the benchmark and tracking portfolio. Panels B, C, and D show stark time variation in the differences of standard deviation, skewness, and kurtosis.

Figure 10 displays the histogram of MSCI-EM Index and EEM iShares ETF monthly return distributions. The sample period is from January 2013 through June 2021.

Figure 11 plots the relationship of the EEM iShares ETF (tracking portfolio), and the MSCI-EM Index (benchmark) return differences (TE) over time. Monthly returns from January 2013 through June 2021 are used.

Figure 12 shows the differences in the first four moments of the EEM iShares ETF and the MSCI-EM Index. Panel A plots the mean differ-

14 During this period, the difference in skewness and kurtosis reached a high of 0.15 and -0.47 in mid-2008, respectively.

15 The data is obtained from <https://finance.yahoo.com>

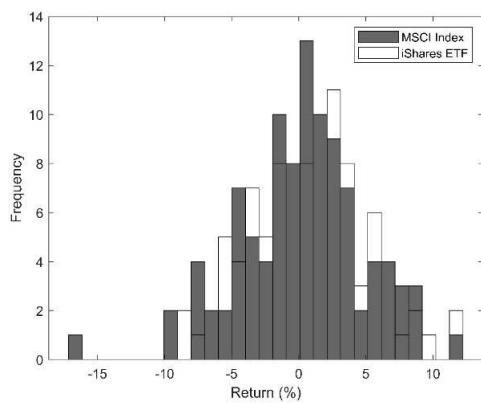


Figure 10. Return distribution of MSCI-EM and EEM

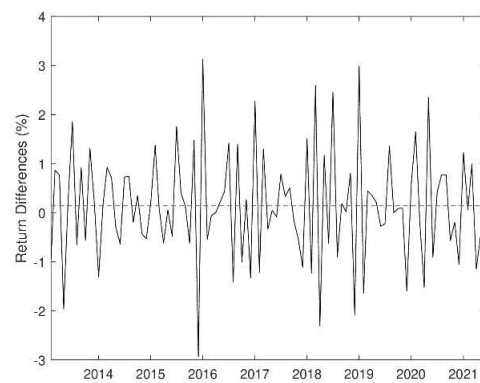
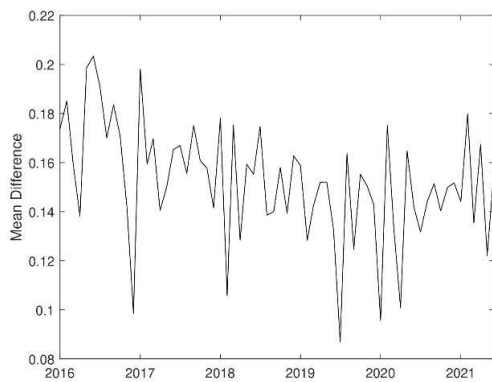
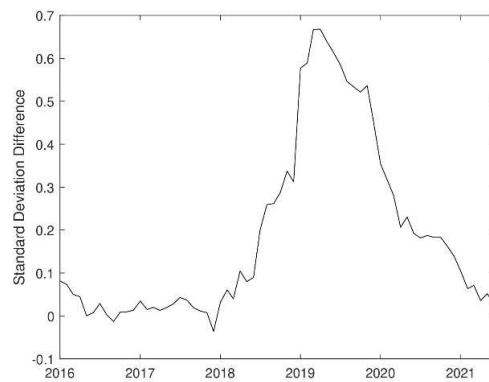


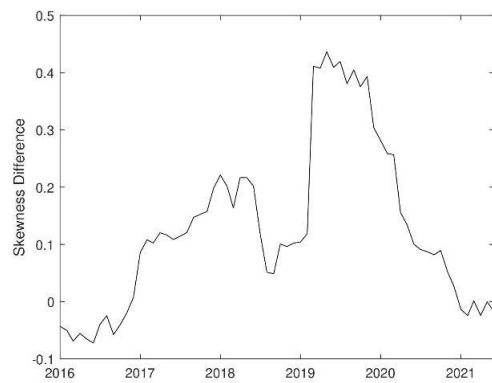
Figure 11. Return differences (TE) of EEM and MSCI-EM



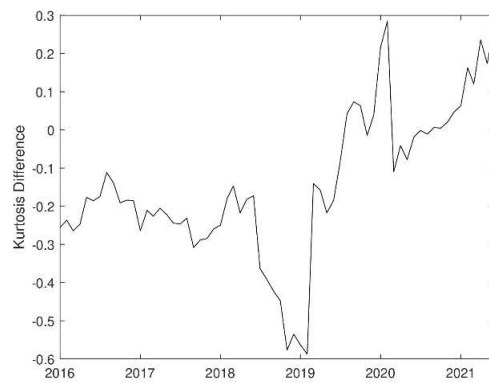
(A) Difference in means



(B) Difference in standard deviation



(C) Difference in skewness



(D) Difference in kurtosis

Figure 12. Difference in 3-year moments rolled through time for EEM and MSCI-EM

ences. Panel B plots the standard deviation differences. Panel C plots the skewness differences. Panel D plots the kurtosis differences. All differences in moments were computed over 3-year rolling windows. The sample period is from January 2013 to June 2021.

TER is little changed during this period, as seen in Figure 13, ranging from approximately 0.9 to 1.2. Meanwhile, QuTER is able to detect these variations in the series, ranging between 0.8 and 1.7. The relative sensitivity of QuTER is even more stark in Panel B of Figure 13.

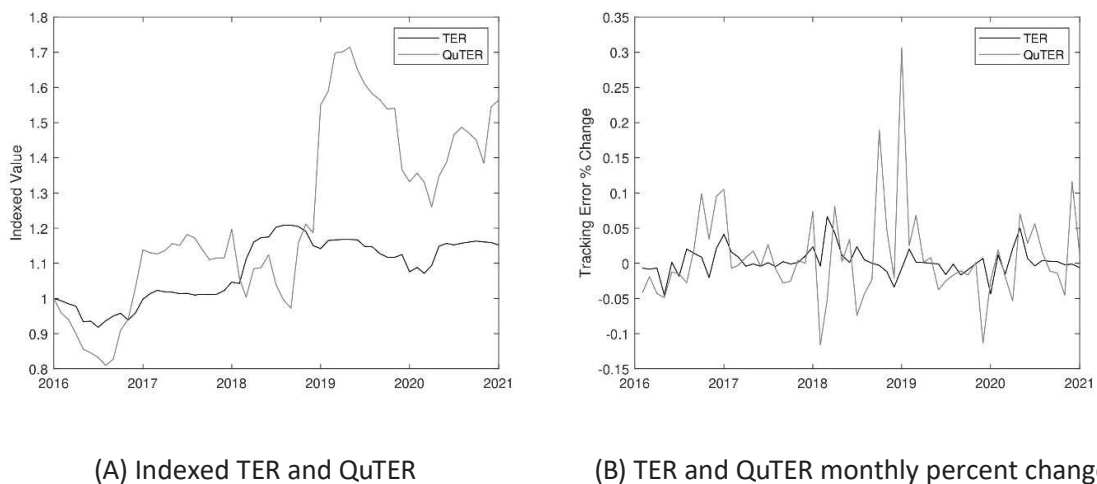


Figure 13. QuTER and TER (3-year rolling window)

Figure 13 plots the TER and QuTER rolling tracking measures. Panel A plots the indexed value of the TER measure and the indexed value of the QuTER measure. Panel B plots the monthly percent change of the calculated rolling TER and QuTER measures. All differences in moments were computed over 3-year rolling windows. The sample period is from January 2013 to June 2021.

In summary, as shown in the two case studies, being able to detect these deviations in higher order moments is important for creating robust

tracking portfolios. Although the DIA appears to be a great tracking portfolio for the DJIA, it does deviate from its benchmark at times of turbulence, like the Dot Com bubble and the Great Recession. Traditional tracking measures like TER were not able to respond quickly to the fat tails of 2002. Moreover, the QuTER was twice as sensitive as TER to the skewness of the Great Recession. The results of the Emerging Markets case study further illustrate the robustness of building tracking portfolios with quantile-based measures.

CONCLUSION

The purpose of this study is to develop better tracking portfolios. A key ingredient is having a robust measure of the differences between a candidate portfolio and its benchmark. Traditional tracking error measures like TEV and TER are insufficient. The QuTE class of tracking measures introduced in this study can detect important differences between two portfolios that are seemingly identical.

The simulations suggest that tracking performance relative to a benchmark with QuTER is statistically more powerful than using traditional measures. The QuTER statistic is robust to various calibrations, such as the choice of quantiles to match. Moreover, the quantiles chosen for matching can be weighted to reflect directions of deviation that are most important to the investor. The case studies illustrate this power in Emerging market and Developed market equities during the turbulent episodes of the Dot Com crash and the Great Recession.

Performance measurement and portfolio evaluation might benefit from including quantile-based measures alongside traditional tracking errors. Moreover, given the success exhibited by the case studies, managers of index and tracking portfolios should consider leveraging the QuTE class for portfolio construction.

AUTHOR CONTRIBUTIONS

Conceptualization: Mike Aguilar, Anessa Custovic.

Data curation: Anessa Custovic, Ruyang Chengan, Ziming Huang.

Formal analysis: Mike Aguilar, Anessa Custovic, Ziming Huang.

Investigation: Mike Aguilar, Anessa Custovic, Ruyang Chengan, Ziming Huang.

Methodology: Mike Aguilar, Anessa Custovic, Ziming Huang.

Project administration: Anessa Custovic.

Resources: Mike Aguilar.

Software: Anessa Custovic, Ruyang Chengan, Ziming Huang.

Supervision: Mike Aguilar.

Validation: Ruyang Chengan.

Visualization: Anessa Custovic, Ruyang Chengan, Ziming Huang.

Writing – original draft: Mike Aguilar, Anessa Custovic, Ruyang Chengan.

Writing – review & editing: Mike Aguilar, Ziming Huang.

REFERENCES

1. Ammann, M., & Tobler, J. (2000). *Measurement and decomposition of tracking error variance* (Working paper). University of St. Gallen.
2. Barro, D., & Canestrelli, E. (2009). Tracking error: a multistage portfolio model. *Annals of Operations Research*, 165(1), 47-66. Retrieved from <https://ideas.repec.org/p/wpa/wuwpge/0510012.html>
3. Beasley, J. E., Meade, N., & Chang, T. J. (2003). An evolutionary heuristic for the index tracking problem. *European Journal of Operational Research*, 148(3), 621-643. [https://doi.org/10.1016/S0377-2217\(02\)00425-3](https://doi.org/10.1016/S0377-2217(02)00425-3)
4. Blitz, D., & Hottinga, J. (2001). Tracking error allocation. *The Journal of Portfolio Management*, 27.
5. Blume, M., & Edelen, R. (2004). S&P 500 indexers, tracking error, and liquidity. *The Journal of Portfolio Management*, 30, 37-46. <https://doi.org/10.3905/jpm.2004.412317>
6. Chincarini, L., & Kim, D. (2006). *Quantitative equity portfolio management: an active approach to portfolio construction and management*. McGraw-Hill.
7. Chung, Y. P., Johnson, H., & Schill, M. J. (2006). Asset pricing when returns are non-normal: Fama-French factors versus higher-order systematic co-moments. *The Journal of Business*, 79(2), 923-940. <http://dx.doi.org/10.1086/499143>
8. Doroc'akov'a, M. (2017). Comparison of ETFs performance related to the tracking error. *Journal of International Studies*, 10, 154-165. <https://doi.org/10.14254/2071-8330.2017/10-4/12>
9. Follmer, H., & Leukert, P. (1999). Quantile hedging. *Finance and Stochastics*, 3(3), 251-273.
10. Franks, E. C. (1992). Targeting excess-of-benchmark returns. *The Journal of Portfolio Management*, 18(4), 6-12.
11. Gaivoronski, A., & Pflug, G. (2005). Value-at-risk in portfolio optimization: properties and computational approach. *Journal of Risk*, 7(2), 1-31. <http://dx.doi.org/10.21314/JOR.2005.106>
12. Giovannetti, B. C. (2013). Asset pricing under quantile utility maximization. *Review of Financial Economics*, 22(4), 169-179. Retrieved from <https://ideas.repec.org/p/spa/wpaper/2012wpecon16.html>
13. Harrell, F. E., & Davis, C. E. (1982). A new distribution-free quantile estimator. *Biometrika*, 69(3), 635-640. <https://doi.org/10.2307/2335999>
14. Jorion, P. (2004). Portfolio optimization with tracking-error constraints. *Financial Analysts Journal*, 59. <https://doi.org/10.2469/faj.v59.n5.2565>
15. Kahneman, D., & Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica*, 47(2), 263-291. <https://doi.org/10.2307/1914185>
16. Kritzman, M. P. (1987). Incentive fees: some problems and some solutions. *Financial Analysts Journal*, 43(1), 21-26. <https://doi.org/10.2469/faj.v43.n1.21>
17. Ma, L., Tang, Y., & Gomez, J. (2019). Portfolio manager compensation in the U.S. mutual fund industry. *The Journal of Finance*, 74(2), 587-638.
18. Mills, T. C. (1995). Modelling skewness and kurtosis in the London stock exchange FT-SE index return distributions. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 44(3), 323-332. <https://doi.org/10.2307/2348703>
19. Pope, P. F., & Yadav, P. K. (1994). Discovering errors in tracking error. *The Journal of Portfolio Management*, 20(2), 2732. <https://doi.org/10.3905/jpm.1994.409471>
20. Roll, R. (1992). A mean/variance analysis of tracking error. *The Journal of Portfolio Management*, 18(4), 13-22. <https://doi.org/10.3905/jpm.1992.701922>

21. Rostek, M. (2010). Quantile maximization in decision theory. *The Review of Economic Studies*, 77(1), 339-371. <https://doi.org/10.1111/j.1467-937X.2009.00564.x>
22. Rudolf, M., Wolter, H.-J., & Zimmermann, H. (1999). A linear model for tracking error minimization. *Journal of Banking & Finance*, 23(1), 85-103. [https://doi.org/10.1016/S0378-4266\(98\)00076-4](https://doi.org/10.1016/S0378-4266(98)00076-4)
23. Wilcox, R., Erceg-Hurn, D., Clark, F., & Carlson, M. (2014). Comparing two independent groups via the lower and upper quantiles. *Journal of Statistical Computation and Simulation*, 84(7), 1543-1551. <https://doi.org/10.1080/00949655.2012.754026>
24. Yamai, Y., & Yoshida, T. (2002). On the validity of value-at-risk: comparative analyses with expected shortfall. *Monetary and Economic Studies*, 20(1), 57-85.