How Much Can a Risk Metric Lie?

Measuring the Manipulability of Ten Standard Risk and Performance Statistics

Rohan Goel | Quanta Student Capital | Paper No. 2

Abstract. Risk and performance statistics decide how funds are ranked, how capital is allocated and how strategies are marketed. Almost all of them are computed from reported returns, which means they inherit whatever the reporting process does to those returns. This paper investigates how far each of ten widely used statistics can be pushed by someone who wants to look good on that statistic but is constrained to deliver no genuine economic improvement. We define a single comparable measure, the manipulation ratio, and evaluate ten metrics against five attacks: return smoothing, tail-risk selling, path reordering, reporting-period selection and option-based payoff reshaping. Two results follow analytically and require no data at all. An exponential smoother with parameter θ scales reported volatility by $\sqrt{(1-\theta)/(1+\theta)}$, which inflates the Sharpe ratio by a factor of 1.53 at $\theta = 0.4$ while leaving the underlying holdings completely untouched. As terminal wealth is invariant to the order in which returns arrive, a fixed set of returns ending at an identical final value can report a maximum drawdown anywhere between 3.20% and 17.38% in our worked example. This was a factor of 5.43 from reordering alone. We found the worst-case manipulation ratio across all five attacks was 10.00 for Volatility and 1.00 for Skewness.

1. Research Question

Fund performance is often reduced to a handful of familiar metrics. Factsheets highlight measures such as the Sharpe ratio, maximum drawdown, Sortino ratio and Calmar ratio, which are then used in screening tools, allocator shortlists and platform rankings. However these metrics are rarely accompanied by any indication of how robust they are to manipulation, or how easily they can be improved without meaningfully improving the underlying investment strategy.

That is the question this paper takes up. Rather than asking what each statistic measures, we ask what happens when someone attacks it:

If a manager wanted to score well on a given risk metric without generating any real economic edge, how far could they push it – and which metrics resist this the most?

Most financial metrics are evaluated by how well they describe typical market behavior. Here, the question is different, and asks - how do those same metrics behave when someone is actively trying to exploit their weaknesses? This way of thinking is more common in fields such as cybersecurity, where a system is tested against deliberate attempts to break it rather than ordinary use. Applying that perspective to investment performance metrics is the central idea of this work.

It is important to distinguish clearly between what is already established and what is new here. The manipulability of performance statistics is not itself a new result. Lo (2002) showed the statistical fragility of the Sharpe ratio. Getmansky, Lo and Makarov (2004) demonstrated how illiquid assets can produce artificially smooth reported returns. Goetzmann, Ingersoll, Spiegel and Welch (2007) showed that the Sharpe ratio can be manipulated. These studies each focus on a particular statistic or mechanism. What we believe is still missing is a framework that tests multiple performance metrics against multiple forms of manipulation, while making the underlying machinery available for others to reproduce and extend.

2. The Metric Panel

Ten statistics were selected to cover the main metric families an allocator would typically encounter, alongside one control. Throughout this paper, r_t denotes the periodic return, k the number of periods per year, μ and σ the periodic mean and standard deviation, and W_t the cumulative wealth index.

The family of metrics are chosen to reflect what investors are likely to encounter in practice. Priority is therefore given to statistics that commonly appear on fund factsheets and screening platforms, rather than measures that may be theoretically preferable but are rarely reported. The set includes two dispersion measures, two tail-risk measures and two path-dependent measures, allowing the results to be compared across broader metric families rather than treated in isolation. Three ratio measures are included because

they are widely used to rank and screen funds. Skewness is included because it is commonly used to describe the asymmetry of a return distribution and is therefore a natural target for manipulation.

Metric	Family	Definition
Volatility σ_a	Dispersion	$\sigma\sqrt{k}$
Downside deviation σ_d	Dispersion	$\sqrt{k\,\mathbb{E}[\min(r_t-\tau,0)^2]}$
Value at Risk VaR_{95}	Tail	$-Q_r(0.05)$, the empirical 5th percentile
Expected shortfall ES_{95}	Tail	$-\mathbb{E}[r_t \mid r_t \le Q_r(0.05)]$
Maximum drawdown $D_{\max}$	Path	$\max_t\,(1-W_t/\max_{s\le t} W_s)$
Ulcer index U	Path	$\sqrt{\frac{1}{T}\sum_t d_t^2}$, d_t the drawdown at t
Skewness γ_1	Shape	$\mathbb{E}\left[(r_t-\mu)^3\right]/\sigma^3$
Sharpe ratio S	Ratio	$(\mu-r_f)\sqrt{k}\,/\,\sigma$
Sortino ratio S_o	Ratio	$(\mu-\tau)\sqrt{k}\,/\,\sigma_d$
Calmar ratio C	Ratio	CAGR / $D_{\max}$
MPPM Θ (control)	Utility	see Equation (1)

Goetzmann et all. (2007) developed a manipulation-proof measure. In this paper we call this the control, and it is designed so that it cannot be improved through trading straegies that do not also increase the investor's expected utility:

$$\Theta = \frac{1}{(1-\rho)\,\Delta t}\,\ln\left[\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1+r_t}{1+r_{f,t}}\right)^{1-\rho}\right], \tag{1}$$

where ρ is the coefficient of relative risk aversion and Δt is the length of one period in years. We take $\rho = 3$. If the framework is working correctly, every attack should move the ten metrics and leave Θ essentially untouched. This provides a built-in validity check, allowing the results to be assessed against a measure specifically designed to resist manipulation.

3. Defining the Manipulation Ratio

Before beginning, there are two things to clear up. The first is the conventional direction used for each metric. For some metrics a higher value looks better (Sharpe, Sortino, Calmar), and for others a lower value looks better (volatility, VaR, drawdown). We attached a direction indicator $d_M = +1$ when higher is better and $d_M = -1$ when lower is better, so that a ratio above one always means the attack made the fund look better.

The more important issue is that a metric can be improved trivially if the underlying return distribution is also allowed to improve. To isolate manipulation from genuine investment skill, any attack must therefore leave the investor no better off. We impose this constraint using the manipulation-proof performance measure in Equation (1), requiring that an attack does not increase the investor's expected utility.

For a metric M, an honest return stream $\mathbf{r}$ and an attack A, define

$$\mathrm{MR}(M,A) \;=\; \left[\frac{M\big(A(\mathbf{r})\big)}{M(\mathbf{r})}\right]^{d_M}, \qquad \text{subject to} \quad \Theta\big(A(\mathbf{r})\big) \;\le\; \Theta(\mathbf{r}). \tag{2}$$

A manipulation ratio of 1.00 means the metric was completely unmoved. A ratio of 2.00 means the attacker doubled the apparent quality of a fund that is, by construction, no better than it was. As a real adversary chooses whichever attack works best, we report the worst case across attacks,

$$\mathrm{MR}^*(M) \;=\; \max_{A\in\mathcal{A}} \mathrm{MR}(M,A), \tag{3}$$

as the headline robustness score for each metric, rather than an average.

4. The Attack Library

In this section we investigate five attacks on the metrics. The five attacks fall into two categories. Reporting and Measurement attacks change the return series observed by the evaluator, without improving the underlying investment outcome. On the other hand, Payoff and Strategy attacks change the 'return-generating' process itself, but are still constrained so that investor welfare has not improved. These attacks allow us to test whether a metric can be distorted through its presentation, and whether the metric can be improved through economically unproductive trading.

4.1 Reporting and Measurement Attacks

Attack 1 — Return smoothing. Reported returns are passed through a smoother,

$$\tilde{r}_t \;=\; \theta\,\tilde{r}_{t-1} \;+\; (1-\theta)\,r_t, \qquad 0 \le \theta < 1. \tag{4}$$

This is the mechanism Getmansky, Lo and Makarov identified in hedge fund returns, where illiquid or stale-marked positions produce the same effect without anybody intending it. The mean overall is unchanged, but for independent returns the reported variance is

$$\mathrm{Var}(\tilde{r}) \;=\; \mathrm{Var}(r)\,\frac{(1-\theta)^2}{1-\theta^2} \;=\; \mathrm{Var}(r)\,\frac{1-\theta}{1+\theta}, \qquad \text{so} \qquad \frac{\tilde{\sigma}}{\sigma} = \sqrt{\frac{1-\theta}{1+\theta}}. \tag{5}$$

By smoothing, we reduce the volatility. As the mean stays unchanged, the reported Sharpe ratio is inflated. At $\theta = 0.4$, Sharpe is increased by $1.53\times$ and at $\theta = 0.5$, $1.73\times$. In this attack, the portfolio is not changed in any way. The reporting is only changed by smoothing. This approach is particularly relevant to hedge funds and other businesses or firms that hold illiquid assets. As illiquid holdings aren't always aligned to fresh market prices, their reported returns can look much smoother than their economic returns.

Attack 2 — Path reordering. This attack keeps the same return series, but changes the order of each return. As terminal wealth, $W_T = \prod_t (1 + r_t)$, only depends on the product of the returns, reordering does not change the final wealth. However path-dependent statistics like maximum drawdown, Ulcer index and Calmar ratio are dependent on the order of the returns. This attack therefore tests how strongly the metrics are influenced by changing the order of returns in a given series.

Attack 3 — Reporting-period selection. This attack is probably the simplest conceptually. The full return series is unchanged, however one may cherry pick a certain window of returns that makes their target metric look better. This attack therefore tests how sensitive a metric is to a select start and end date, rather than making a genuine improvement in the underlying investment.

4.2 Payoff and Strategy Attacks

Attack 4 — Tail-risk selling. Suppose a model of returns that earns small positive returns, p most of the time, but ocassionally suffers a large loss, L:

$$r_t = \begin{cases} p & \text{with probability } 1-q, \\ p - L & \text{with probability } q. \end{cases} \tag{6}$$

By setting $p = qL$, we see that the expected returns are zero, demonstrating no positive return edge. If the case of the loss is rare, it is unlikely it will be seen in the reports, making the metrics look strong. However, the model is still sketchy. Supposing $q = 0.01$ per month, the probability that in five years there is no event of the loss is $0.99^{60} = 54.7\%$. Hence, more than half the records will look strong, despite the model being flawed.

Attack 5 — Payoff reshaping. This attack involves using an option overlay to alter the shape of the returns. For example, a return series with a few negative returns and a few positive returns can be shaped as one negative return and everything else positive. This allows us to test whether a metric can be made to look stronger, purely by redistributing where gains and losses occur.

5. Results

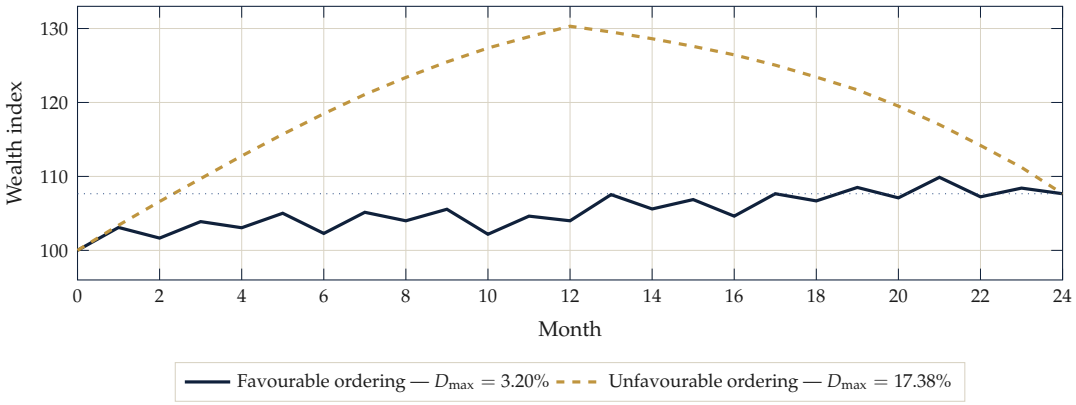


Figure 1: Here, we took an example return series, and ordered it more faouvrably in Python, over 24 months. It is not real data. Both paths start and finish at the same Wealth index of 107.66. The reported maximum drawdown was found to differ by a factor of 5.43

	Smoothing	Tail-sell	Reorder	Period	Payoff
Calmar	3.57	0.55	1.92	4.49	1.75
Sortino	6.36	0.39	1.00	3.18	1.56
Sharpe	3.96	0.61	1.00	2.28	1.43
Skewness	1.00	1.00	1.00	1.00	1.00
Ulcer	3.19	0.63	1.77	3.38	2.08
Max drawdown	3.06	0.98	1.92	2.87	2.09
ES (95%)	4.99	10.00	1.00	1.79	2.88
VaR (95%)	8.54	1.36	1.00	1.93	2.23
Downside dev.	5.93	1.22	1.00	2.08	1.94
Volatility	3.69	10.00	1.00	1.54	1.78

Figure 2: These are the results for each attack on each metric, displayed as a heatmap A value of 1.00 means the metric was unmoved by the attack.

The heatmap shows that different attacks have substantially different effects on the different metrics. Overall, Return smoothing had the greatest broader effect, as it produces large vales for each metric. Tail-risk selling is seen to be much more concentrated. Most of its effect was on ES(95%) and volatility. The other metrics with a value of less than 1.00 show that this attack makes some metrics less favorable. Path reordering mainly affected the path-dependent metrics, leaving the distribution-based metrics largely unchanged. Reporting-period selection and Payoff reshaping affected a wider range of metrics, but by smaller amounts than Smoothing. Interestingly, Skewness remained unchanged in every attack.

Table 2 — Worst-case manipulation ratio by metric

Metric	MR*	Worst attack	Interpretation
Volatility	10.00	Tail-sell	Highly vulnerable to hidden tail risk
ES (95%)	10.00	Tail-sell	Extremely distorted by tail losses
VaR (95%)	8.54	Smoothing	Sensitive to return smoothing
Sortino	6.36	Smoothing	Strongly inflated by smoothing downside variation
Downside dev.	5.93	Smoothing	Highly sensitive to smoothing
Calmar	4.49	Period	Strongly dependent on reporting window
Sharpe	3.96	Smoothing	Inflated by smoothing volatility
Ulcer	3.38	Period	Dependent on reporting window
Max drawdown	3.06	Smoothing	Sensitive to return smoothing
Skewness	1.00	none	Robust against all attacks
MPPM Θ (control)			Should remain near 1.00

For the payoff and strategy attacks, the MPPM control ensures that any improvement in the the target metric does not follow with improvement in investor welfare.

There is one clear expectation that was contradicted with these results. We expected tail-risk selling to inflate Sharpe and Sortino, as a sample containing small gains and no tail event should appear very stable, with low volatility. However the attack actually reduced both measures, with manipulation ratios of 0.61 and 0.39 respectively. This suggests that the benefit of hiding tail risk is more metric-specific.

6. Application to Real Fund Returns

It is worth investigating how these results and ideas actually show up in practice. Return smoothing gives us a way to describe how. For the exponential smoother of Equation (4), the first order autocorrelation of reported returns is exactly the smoothing parameter,

$$\rho_1(\tilde{r}) = \theta, \tag{7}$$

so θ can be estimated straight from a published return series as $\hat{\theta} = \hat{\rho}_1$. Inverting Equation (5) then recovers an implied 'de-smoothed' Sharpe ratio,

$$S_{\text{implied}} = S_{\text{reported}} \sqrt{\frac{1-\hat{\theta}}{1+\hat{\theta}}}. \tag{8}$$

We can now apply this to different funds and ETFs to see if there is any underlying smoothness.

Table 3 — Implied smoothing in exchange-traded funds

Fund	Exposure	Reported S	$\hat{\theta}$	Implied S	Cut
Highly liquid underlying holdings					
SPY	US large-cap equity	0.99	0.00	0.99	0%
QQQ	US technology equity	1.08	0.00	1.08	0%
IEF	US Treasuries, 7–10 year	0.17	0.03	0.17	3%
Less liquid underlying holdings					
IWM	US small-cap equity	0.54	0.00	0.54	0%
HYG	US high-yield credit	0.61	0.00	0.61	0%
VNQ	US real estate	0.38	0.00	0.38	0%
EMB	EM sovereign debt	0.34	0.00	0.34	0%

Here we see a null result. No fund in this sample shows smoothing, and de-smoothing changes Sharpe ratios by no more than 3%. This suggests that the smoothing effect depends more on how the assets are

priced and report, than on their illiquidity. As ETFs are continuously traded, stale valuations are limited. The smoothing effect identified by Getmansky, Lo and Makarov is therefore more relevant to self-reported or infrequently valued vehicles such as hedge funds and private funds.

7. Conclusion

Across the five attacks, the worst-case manipulation ratios were highest for volatility and expected shortfall, followed by VaR, Sortino, downside deviation, Calmar, Sharpe, Ulcer index and maximum drawdown, while skewness remained unchanged in our tests. Return smoothing was the most broadly effective attack, materially improving almost every metric, although tail-risk selling produced the largest individual distortions. The main implication for an allocator is that no performance statistic should be interpreted in isolation. A strong Sharpe ratio, low volatility or limited drawdown can reflect the structure or reporting of returns rather than genuine economic improvement, so metrics should be considered alongside the full return history, payoff structure and valuation process.

We found a few key limitations in this investigation. The first limitation is that these attacks are not really observed. This paper is purely a test on how one could manipulate the figures. If one were to manipulate these figures, it is likely it would be with illiquid assets, however we do not claim that funds have actually used these attacks. Secondly, the economic constraint depends on the assumed coefficient of relative risk aversion. All of our results use $\rho = 3$, and the manipulation-proof measure is only manipulation-proof relative to the preferences it represents. Someone with different risk preferences could therefore draw a different line between manipulation and improvement.

References

Getmansky, M., Lo, A. W., & Makarov, I. (2004). An econometric model of serial correlation and illiquidity in hedge fund returns. *Journal of Financial Economics*, 74(3), 529–609.

Goetzmann, W., Ingersoll, J., Spiegel, M., & Welch, I. (2007). Portfolio Performance Manipulation and Manipulation-proof Performance Measures. *Review of Financial Studies*, 20(5), 1503–1546.

Lo, A. W. (2002). The Statistics of Sharpe Ratios. *Financial Analysts Journal*, 58(4), 36–52.

Disclaimer. Quanta Student Capital is a student-led research initiative. This paper is produced for educational and research purposes only. Nothing in it constitutes investment advice, a recommendation, or an offer to buy or sell any security.