Further Biases in Using Dollar-Weighted Returns to Infer Investment Timing Effects

Simon Hayley, Cass Business School*

This version: 22 March 2019

Draft – comments welcome

Abstract

The IRR (dollar-weighted return) reflects the periodic addition or withdrawal of funds by investors, and the difference between IRR and geometric mean is widely used to indicate the impact that the timing of these flows has had on investor returns. This is a biased measure, since it is also affected by investment flows which "chase" previous strong returns. A method has previously been derived for separating this bias from genuine timing effects. This paper demonstrates that using in-sample mean returns for this decomposition causes an additional bias which again misleadingly suggests bad investor timing. This paper quantifies this bias, allowing unbiased investor timing effects to be estimated. A proper understanding of these biases is of significant practical importance, since investors are often presented with biased timing indicators based on IRRs.

JEL Classification: G11
Keywords: performance measurement; internal rate of return; dollar-weighted return, hindsight effect.

*Cass Business School, 106 Bunhill Row, London EC1Y 8TZ, UK.
E-mail: simon.hayley.1@city.ac.uk. Tel +44 20 7040 0230. I am very grateful for comments from Ian Marsh, Richard Payne, Tim Perkins and participants at a seminar at Cass Business School. The usual disclaimer applies

Further Biases in Using Dollar-Weighted Returns to Infer Investment Timing Effects

1. Introduction

The internal rate of return (IRR) reflects the timing of investment flows. As a result, it is widely argued that the "performance gap" between the IRR and geometric mean (GM) return measures the extent to which bad timing of these flows affects returns. Using this method a large number of papers have concluded that the bad timing of investments into the equity market has substantially reduced the overall return to the average investor (Dichev (2007), Friesen & Sapp (2007), Dichev and Yu (2011), Chieh-Tse Hou (2012), Muňoz (2015), Navone and Pagani (2015), Cornell et al. (2016)).

This method has been shown to be biased. Hayley (2014) demonstrated that the IRR can be affected by investment flows *after* a period of unusually high or low returns just as much as it can be affected by investment flows before this period. Indeed, there is a strong case for expecting the first effect (the "hindsight effect") to have a significant negative impact on IRRs, since there is ample evidence that investors "chase returns" by investing more following periods of high returns (e.g Siri and Tufano (1998), Phillips et al. (2012)).

Hayley (2014) also derived a method for decomposing the performance gap into the hindsight effect and genuine timing effects. The results of this decomposition were sensitive to the initial assumption chosen for the mean return, but over a range of different assumptions, the observed negative performance gap in US equities could be explained by the hindsight effect, with any timing effect close to zero.

Despite this, subsequent research has continued to interpret the performance gap between IRRs and GM returns as an indicator of bad timing (e.g. Muňoz (2015), Navone and Pagani (2015), Cornell et al. (2017)). Commercial databases such as Morningstar continue to publish mutual fund IRRs in addition to their GM returns.

Vicente and Muňoz (2018) uses an approach based on Hayley (2014) to decompose the performance gaps on US mutual funds (1990-2016) and finds that although the majority of the performance gap can be attributed to the hindsight effect, there still appears to be significant genuine bad timing of investment flows into all the major categories of mutual fund (e.g. a "corrected" timing effect of 0.71% out of a total performance gap of 1.80% for all US domestic equity mutual funds). However, these decompositions were derived by initially assuming for each fund that the return in each period is equal to the GM return observed for that fund over the entire investment period. It is demonstrated below that this assumption consistently biases the decomposition.

The intuition behind this additional bias is that using the sample mean implies that above-average returns prior to any given period must by construction be followed by subsequent below-average returns. Given that return-chasing by investors leads to additional investment inflows following periods with above-average returns, these inflows will by construction come ahead of periods of below-average returns. This is an entirely spurious correlation. It will tend to push the

DWRR below the GM return, and this is likely to be misinterpreted as evidence that investors timed their investment flows badly.

Section 3 (below) derives an expression for the expected size of this bias. This is close to the size of the timing effects claimed by Vicente and Muňoz (2018), suggesting that their results are not evidence of consistent bad timing by investors. More importantly, identifying and quantifying this bias offers a methodological improvement which will allow unbiased estimates of any genuine timing effects to be estimated. A proper understanding of these biases is of significant practical importance, since investors are often presented with biased timing indicators based on IRRs.

2. Approximating the Hindsight Effect in the IRR

The IRR is usually defined as the discount rate that sets the NPV of investment cashflows to zero:

$$K_0 + \sum_{t=1}^{T} \frac{a_t}{(1+IRR)^t} - \frac{K_T}{(1+IRR)^T} = 0 \tag{1}$$

where K_t is the portfolio value at the end of period t, and a_t is the flow of new cash invested into this portfolio each period. This flow could be positive or negative, and is assumed to take place at the end of each period.[1] Remaining assets K_T are treated as if liquidated in the final period.

[1] Hayley (2014) shows that shifting to assuming that flows come at the start of each month has negligible impact on the resulting decompositions.

The change in portfolio value each period is a function of the return r_t on the portfolio during this period and any additional cash a_t invested: $K_t=K_{t-1}(1+r_t)+a_t$. Substituting this relationship into equation (1) and rearranging shows that the IRR can also be expressed as a weighted average of the individual period returns r_t, where the weight used is determined by the present value of the investment at the start of each period (see Dichev and Yu, 2011). Hence the IRR is also referred to as the dollar-weighted rate of return:

$$IRR = \sum_{t=1}^{T} \left(r_t \frac{K_{t-1}}{(1+IRR)^{t-1}} \Big/ \sum_{t=1}^{T} \frac{K_{t-1}}{(1+IRR)^{t-1}} \right) \quad (2)$$

The IRR is a complex polynomial function of the periodic returns r_t over the course of the investment horizon. This complexity means that approximations are generally required in order to derive useful analytic results. The simulations presented later in this paper allow us to assess the accuracy of these assumptions. We assume that the returns r_t are serially independent, which can be interpreted as weak form market efficiency. If inflows and outflows after period 0 and the standard deviation of r_t are all fairly small, K_{t-1} grows at an approximately uniform rate which is roughly equal to the IRR, implying that each r_t will be given an approximately equal weight in the IRR calculation, so the IRR will be close to the mean:

$$IRR \approx \frac{1}{T} \sum_{1}^{T} r_t \quad (3)$$

From this simple starting condition, we introduce a single investment inflow a at the end of period n (a is expressed here as a percentage of the portfolio value at that time, and a negative a indicates an outflow as investors withdraw funds). This inflow increases K_t in all subsequent periods in the investment horizon. Instead of all weights being *1/T*, the weights given to r_t will now be approximately as shown in equation 4:

$$IRR \approx \frac{1}{\mathrm{n+(1+a)(T-n)}} \sum_1^n r_t + \frac{1+\mathrm{a}}{\mathrm{n+(1+a)(T-n)}} \sum_{n+1}^T r_t \tag{4}$$

$$\approx \frac{1}{\mathrm{n+(1+a)(T-n)}} \sum_1^{n-1} r_t + \frac{\mu}{\mathrm{n+(1+a)(T-n)}} + \frac{(r_n-\mu)}{\mathrm{n+(1+a)(T-n)}} + \frac{1+\mathrm{a}}{\mathrm{n+(1+a)(T-n)}} \sum_{n+1}^T r_t \tag{5}$$

Prior research has suggested that the cash inflow *a* is strongly related to the immediately preceding return (Sirri/Tufano (1998), and Phillips et al. (2012)). This is commonly referred to as "return chasing" by investors, and is presumed to be due to strong returns in one period resulting in investors becoming more optimistic about future returns. For simplicity we assume that *a* is a function of r_n, but not of earlier returns. This, and the serial independence of the returns, means that the summations in equation (5) have no correlation with their multiplicands, and so have expected value μ in each period. Only r_n is correlated with *a* in the denominator, thus:

$$\mathrm{E}[IRR] \approx \frac{(\mathrm{n}-1)\mu+\mu+(1+\mathrm{a})(\mathrm{T-n})\mu}{\mathrm{n+(1+a)(T-n)}} + \mathrm{E}\left[\frac{(\mathrm{r_n}-\mu)}{\mathrm{n+(1+a)(T-n)}}\right] \tag{6}$$

$$\approx \mu + \mathrm{E}\left[\frac{(\mathrm{r_n}-\mu)}{\mathrm{n+(1+a)(T-n)}}\right] \tag{7}$$

For small a, $\frac{1}{\mathrm{n+(1+a)(T-n)}} \approx \frac{1}{\mathrm{T}}$ and we approximate using 1/(1+x)≈1-x for small x:

$$\mathrm{E}[IRR] \approx \mu + \left(\frac{1}{T}\right) E[(r_n-\mu)\left(2-\frac{1}{T}\left(\mathrm{n}+(1+\mathrm{a})(\mathrm{T}-\mathrm{n})\right)\right)] \tag{8}$$

Assuming the relationship between flows and returns to be linear ($a= w(r_n-\mu)$, where w is a constant), and noting that $\mathrm{E}[r_n-\mu]=0$, $\mathrm{E}[(r_n-\mu)^2]=\sigma^2$:

$$E[IRR] \approx \mu - w\frac{\sigma^2}{T^2}(T-n) \tag{9}$$

The derivation above shows how the hindsight effect identified in Hayley (2014) affects the performance gap (IRR-μ). Return-chasing by investors leads to a significant investment inflow after an above-average return r_n in the immediately preceding period. Thus the numerator $(r_n$-$\mu)$ in equation (7) is positively correlated with the denominator. For example, a high return r_n would generate a positive investment inflow a which increases the relative weight given to future investment returns. But the weights given to all the periods in our investment horizon must sum to one, so this inflow a also reduces the weight given to prior returns, including r_n. Conversely, a below-average return r_n leads to an outflow of funds which reduces the weight given to future returns and increases the relative weight given to r_n. Thus above-average returns are given a reduced weight, and below-average returns an increased weight, consistently reducing the IRR.

It would be a mistake to interpret this as bad investment timing – it does not have any effect on investors' expected terminal wealth since it is a retrospective adjustment of the relative weights in the IRR calculation after the period has passed because inflows tend to be correlated with past returns. Instead, genuine timing effects need to adjust investors' exposures *before* the return concerned (i.e. inflows correlated with future returns).

Equation (9) gives us the effect on the IRR from a specific investment flow immediately after period n. In order to generate the total effect on the IRR, we sum this effect over each specific period from $n=1$ to T, giving us a result equivalent to that in Perkins (2018):

$$E[IRR - \mu] \approx -\, w\frac{\sigma^2}{2}\left(\frac{\mathrm{T}-1}{\mathrm{T}}\right) \tag{10}$$

For large T, this approximates to $-\frac{w\sigma^2}{2}$ so this bias is not a small sample effect. The impact that each individual r_t has on the IRR declines with T as it becomes a smaller part of the sample. But the total effect on the IRR is then the sum of a correspondingly increased number of such individual effects.[2]

The above derivation considered the effect of each individual flow a in isolation, ignoring any interaction between the effects of flows in different periods. Given the underlying assumptions that returns r_t are independently and identically distributed and that flows are a stable function of these returns, the effect of interactions between these periodic flows in the IRR calculation are likely to be of second order. The simulations (Table 1, below) confirm that equation (10) is an adequate approximation for investment horizons in the range used by Vicente and Muňoz (2018).

[2] Equation (10) gives an estimate of the performance gap as normally defined: IRR-geometric mean (GM). This may seem contradicted by the fact that μ was defined as the expected periodic return E[r_t], i.e. the arithmetic mean return. However, there is another effect at work even if there are no periodic cashflows (a_t=0): an above-average r_t increases the portfolio value K_t, increasing the relative weight given in the IRR calculation to subsequent returns, and hence reducing the relative weight given to r_t itself (a below-average r_t will correspondingly increase its own weight). For simplicity suppose that r_t=μ for all t, except for r_n>μ. The weight given to r_t in the IRR calculation is a function of the present value of K_{t-1}. The value of K_{t-1} is unaffected by r_t, but the discount rate will have increased (by around $\frac{r_n-\mu}{T}$, as a first approximation), reducing the present value by approximately $\left(\frac{1}{(1+(\mathrm{r_n}-\mu)/\mathrm{T})}\right)^{n-1}$. The weights given to r_1 through r_{n-1} will also have been reduced, but the weights for r_{n+1} through r_T will have been increased because of the increased portfolio values K_t to K_T. If the r_t are serially independent then the only systematic effect will be that each large r_t reduces its own weight in the IRR calculation (and vice versa). Taking a Taylor expansion of the $r_n\left(\frac{1}{(1+(\mathrm{r_n}-\mu)/\mathrm{T})}\right)^{n-1}$ term, taking expectations and averaging over all t gives $\mu - \frac{\sigma^2}{2}$, i.e. even before introducing any intermediate cashflows a_t, the expected value of the IRR is the GM, not the AM. Equation (10) shows the *additional* effect which pulls the IRR below the GM as we introduce periodic cashflows which are positively correlated with prior returns.

Equation (10) gives us an expression for the expected size of the theoretical performance gap $E[IRR - \mu]$ when investment flows chase returns (according to $a_n = w(r_{n-1} - \mu)$).We might be concerned that in practice because μ is not known, the empirical performance gap $IRR - \bar{r}$ is used, which compares the IRR to the sample mean. However, whilst equation (7) requires the existence of a well-defined distribution mean μ, it does not require us to know its value. Our corresponding estimate of μ as the sample mean would inevitably include some noise, but it is unbiased, so the expected value of the empirical performance gap $(IRR - \bar{r})$ is the same as our expression for the true gap in equation (10). By contrast, the following section shows that using the sample mean results in a substantial bias in the *decomposition* of this gap into an estimated hindsight effect and genuine timing effect.

3. Decomposing the Performance Gap Using the Sample Mean Return

This section investigates whether using the sample mean rather than the true mean μ affects the decomposition of $(IRR - \bar{r})$ into timing and hindsight effects. For this purpose, we again assume a set of serially uncorrelated returns r_t. These can be thought of as generated by a pure random walk process, removing the possibility of consistently good or bad timing. We then derive an expression for the size of the timing effect that we would estimate when we decompose this performance gap.

The decomposition process derived by Hayley (2014) works by initially setting all periodic returns r_t equal to an assumed mean μ, and all periodic flows a_t to zero. At this point the IRR for this investment will by construction be equal to μ. The first observed return r_1 is then substituted in, and the IRR recalculated. Then the first observed investment inflow a_1 is substituted in, and the IRR again recalculated. These substitutions are repeated for each successive period. Following all these substitutions, all actual data are in place, so the final calculation gives the IRR observed in the actual data. Because these substitutions are made in chronological order, substituting in the returns has a non-zero net effect on the IRR to the extent that each r_t is correlated with prior investment flows (this represents genuine good/bad timing of these flows which will affect investor wealth). By contrast, substituting in the flows a_t affects the IRR to the extent that the flows a_t are correlated with prior returns the hindsight effect).

We derived equation (10), giving the expected total performance gap, by initially considering each flow a_t in isolation. By contrast, now that we are estimating the apparent "timing effect" we need explicitly to include the investment flows prior to period t in order to estimate the degree to which these are correlated with subsequent returns. We continue to assume that each inflow (measured as a percentage of the fund's size at the time) is $a_t = w(r_t - \mu)$.

We will first consider what happens when we substitute in a single r_n. We then repeat this process for n=1 to T to generate an aggregate overall effect for $1 \leq n \leq T$, replicating the process by which the total timing effect is estimated. As above, the IRR calculation immediately before substituting in r_n gives a relative weight to each of the prior returns r_1 to r_{n-1} which depends on the cumulative investment inflows ahead of each period (a_0 is normalised to zero, and we set the initial portfolio value $K_0=1$). All future returns r_n to r_T are given an equal relative weight determined by the cumulative inflows so far $(1+\sum_{1}^{n-1} a_{t-1})$, with these future returns at this stage all assumed equal to μ:

$$\text{IRR} \approx \frac{\sum_{t=1}^{n-1}\left(1+\sum_{j=1}^{t} a_{j-1}\right) r_t + \left(1+\sum_{t=1}^{n-1} a_t\right) \sum_{n}^{T} \mu}{\sum_{t=1}^{n-1}(1+\sum_{j=1}^{t} a_{j-1}) + \left(1+\sum_{t=1}^{n-1} a_t\right)(\text{T}-\text{n}+1)} \tag{11}$$

From this starting point we now substitute in the actual value of r_n:

$$\text{IRR} \approx \frac{\sum_{t=1}^{n-1}\left(1+\sum_{j=1}^{t} a_{j-1}\right) r_t + (1+\sum_{t=1}^{n-1} a_t)(r_n-\mu) + \left(1+\sum_{t=1}^{n-1} a_t\right) \sum_{n}^{T} \mu}{\sum_{t=1}^{n-1}(1+\sum_{j=1}^{t} a_{j-1}) + \left(1+\sum_{t=1}^{n-1} a_t\right)(\text{T}-\text{n}+1)} \tag{12}$$

The denominator does not change, since we have not yet substituted in a_t, so the change in IRR resulting from substituting in r_n is the difference between equations 11 and 12:

$$dIRR \approx \frac{(1+\sum_{t=1}^{n-1} a_t)(r_n-\mu)}{\sum_{t=1}^{n-1}(1+\sum_{j=1}^{t} a_{j-1}) + \left(1+\sum_{t=1}^{n-1} a_t\right)(\text{T}-\text{n}+1)} \tag{13}$$

E[dIRR] in equation 13 is zero, since $(r_n-\mu)$ is independent of prior a_t and $E[r_n-\mu]=0$. This correctly reflects the fact that if the r_t follow a random walk, and hence are genuinely unforecastable, then we should expect no timing effect. But in practice we do not know μ with certainty. Hayley (2014) responded to this by considering a range of different values for μ. Vicente and Muňoz (2018) instead estimate the timing and hindsight effects for different US funds just by initially setting all r_t equal to the observed mean return $\bar{r}$, giving:

$$dIRR \approx \frac{(1+\sum_{t=1}^{n-1} a_t)(r_n-\bar{r})}{\sum_{t=1}^{n-1}(1+\sum_{j=1}^{t} a_{j-1}) + \left(1+\sum_{t=1}^{n-1} a_t\right)(\text{T}-\text{n}+1)} \tag{14}$$

This means that E[*dIRR*] is no longer zero, since $(\mathrm{r_n} - \bar{\mathrm{r}})$ is correlated with $\sum_{t=1}^{n-1} a_t$, since $a_t = w(r_t - \mu)$. Approximating, using *1/(1+x)≈1-x* and noting that the denominator is approximately equal to *T*:

$$dIRR \approx \frac{(\mathrm{r_n} - \bar{\mathrm{r}})}{\mathrm{T}}\left(1 + \sum_{t=1}^{n-1} a_t\right)\left(2 - \frac{1}{\mathrm{T}}\left(\sum_{t=1}^{n-1}\left(1 + \sum_{j=1}^{t} a_{j-1}\right) + \left(1 + \sum_{t=1}^{n-1} a_t\right)(\mathrm{T} - \mathrm{n} + 1)\right)\right) \tag{15}$$

$$dIRR \approx \frac{(\mathrm{r_n} - \bar{\mathrm{r}})}{\mathrm{T}}\left(1 + \sum_{t=1}^{n-1} a_t\right)\left(1 - \frac{1}{\mathrm{T}}\left(\sum_{t=1}^{n-1}\sum_{j=1}^{t} a_{j-1} + (\mathrm{T} - \mathrm{n} + 1)\sum_{t=1}^{n-1} a_t\right)\right) \tag{16}$$

$$\begin{aligned} dIRR \approx{} & \frac{(\mathrm{r_n} - \bar{\mathrm{r}})}{\mathrm{T}} + \frac{(\mathrm{r_n} - \bar{\mathrm{r}})}{\mathrm{T}}\left(\sum_{t=1}^{n-1} a_t - \frac{1}{\mathrm{T}}\left(\sum_{t=1}^{n-1}\sum_{j=1}^{t} a_{j-1} + (\mathrm{T} - \mathrm{n} + 1)\sum_{t=1}^{n-1} a_t\right)\right) \\ & - \frac{(\mathrm{r_n} - \bar{\mathrm{r}})}{\mathrm{T}^2}\left(\sum_{t=1}^{n-1} a_t\left(\sum_{t=1}^{n-1}\sum_{j=1}^{t} a_{j-1} + (\mathrm{T} - \mathrm{n} + 1)\sum_{t=1}^{n-1} a_t\right)\right) \end{aligned} \tag{17}$$

The first term is simply the increase in the recorded mean at this stage in the decomposition as a result of substituting in the actual value of r_n in place of $\bar{r}$. This has an expected value of zero. Recalling that $a_n = w(r_t - \mu)$, the last term is a collection of third order terms in the deviation of r_t from either the sample mean or the true mean. The expected value of these third order terms is likely to be small, and we omit them to focusing instead on the second-order terms. The second term has expectation:

$$E[dIRR] \approx -E\left[\frac{(\mathrm{r_n} - \bar{\mathrm{r}})}{\mathrm{T}^2}\left(\sum_{t=1}^{n-1}\sum_{j=1}^{t} a_{j-1} - (\mathrm{n} - 1)\sum_{t=1}^{n-1} a_t\right)\right] \tag{18}$$

$$E[dIRR] \approx -\frac{\mathrm{w}}{\mathrm{T}^2} E\left[\sum_{t=1}^{n-1} \sum_{j=1}^{t} (\mathrm{r_n} - \bar{\mathrm{r}})(\mathrm{r}_{j-1} - \mu) \ - (\mathrm{n} - 1) \sum_{t=1}^{n-1} (\mathrm{r_n} - \bar{\mathrm{r}})(\mathrm{r}_t - \mu)\right] \tag{19}$$

$$E[dIRR] \approx -\frac{\mathrm{w}}{\mathrm{T}^2} E\left[\sum_{t=1}^{n-1} \sum_{j=1}^{t} (\mathrm{r_n r_{j-1}} - \bar{\mathrm{r}}\mathrm{r}_{j-1} + \bar{\mathrm{r}}\mu - \mathrm{r_n}\mu) \ - (\mathrm{n} - 1) \sum_{t=1}^{n-1} (\mathrm{r_n r_{t-1}} - \bar{\mathrm{r}}\mathrm{r}_{t-1} + \bar{\mathrm{r}}\mu - \mathrm{r_n}\mu)\right]$$

Recalling that $\mathrm{E[r_n]} = \mathrm{E}[\bar{\mathrm{r}}] = \mu$, and that serial independence means that $\mathrm{E[r_n r_{n-j}]} = \mu^2$:

$$E[dIRR] \approx -\frac{\mathrm{w}}{\mathrm{T}^2} E\left[\sum_{t=1}^{n-1} \sum_{j=1}^{t} \left(\mu^2 - \frac{r_t^2 + (T-1)\mu^2}{T}\right) - (\mathrm{n} - 1) \sum_{t=1}^{n-1} \left(\mu^2 - \frac{r_t^2 + (T-1)\mu^2}{T}\right)\right] \tag{20}$$

$$E[dIRR] \approx -\frac{\mathrm{w}}{\mathrm{T}^2} \left(\frac{-(n-1)(n-2)\sigma^2}{2T} + \frac{(n-1)^2\sigma^2}{T}\right) \tag{21}$$

$$E[dIRR] \approx -\frac{wn(n-1)\sigma^2}{2T^3} \tag{22}$$

Summing from n=1 to T, the expected value of this estimated timing effect over the whole decomposition process is:

$$\approx -\left(\frac{T-1}{\mathrm{T}}\right)^3 \frac{w\sigma^2}{6} \tag{23}$$

Thus when substituting in the r_t, there is an expected reduction in the IRR even if, as assumed in the derivation above, there is no correlation between successive returns or between flows and subsequent returns. This reduced IRR should not be interpreted as the effect of bad investment timing. Instead it is entirely due to the use of the sample mean return rather than the actual mean μ, which introduces a spurious negative autocorrelation of returns, since above-sample-mean cumulative returns must by construction be followed on average by below-sample-mean returns.

Referring to equation (14), $\mathrm{E}[(1+\sum_{t=1}^{n-1} a_t)(r_n-\bar{r})]<0$ because even though $E[(r_n-\bar{r})]=E[a_t]=0$, these two components are negatively correlated because of the return-chasing investor behaviour which leads to a large (small) $\sum_{t=1}^{n-1} a_t$ if prior returns have been above (below) average, and hence future returns will by construction be below (above) average. By contrast, if we were to use the actual mean μ, $E[(1+\sum_{t=1}^{n-1} a_t)(r_n-\mu)]=0$ because the two component terms are independent.

In effect, this is a form of the spurious negative autocorrelation effect identified by Kendall (1954). That was a small sample effect which disappears as the length of the horizon increases. In the derivation above, the effect of each individual return r_n does indeed become negligible for large T (equation 22), but the total effect does not disappear when we sum over all periods (equation 23). Thus the effect of this spurious autocorrelation does not disappear for large horizons, implying that decomposting the performance gap using the sample mean is likely to result in a misleading negative "timing" effect of $-w\sigma^2/6$ even when, as assumed in our derivation above, r_t follows a random walk.

One way around this is, as in Hayley (2014), to use a range of plausible alternative assumptions for the true mean of the return distribution. Failing this, we can use equation (23) to construct a ready-reckoner for the size of the likely bias, and only apparent timing effects significantly different from this should be construed as evidence of genuine timing effects.

We saw earlier that the total expected effect of return-chasing on the IRR when returns follow a random walk is $\frac{-w\sigma^2}{2}$. Of this, we have found that $\frac{-w\sigma^2}{6}$ is likely to be falsely attributed to bad timing, because of the spurious autocorrelation of $(IRR - \bar{r})$. The remaining $\frac{-w\sigma^2}{3}$ will then be correctly attributed to the hindsight effect (as confirmed in the annex). Thus, for large T, one third of the recorded "performance gap" (IRR-GM) is likely to be interpreted as being due to bad timing, even if there is no such effect. The remaining two thirds will be correctly attributed to the hindsight effect caused by return chasing.

Vicente and Muňoz (2018) found that the total performance gap of 1.80% per annum for all US domestic equity mutual funds decomposed into a hindsight effect of 1.09% and bad timing of 0.71%. They interpreted this as evidence that investment flows were indeed badly timed, although the scale of this effect had been substantially reduced by taking account of the hindsight effect. However, their decomposition was based on the assumption that the mean return for each fund represented the true mean of the underlying return distribution. As demonstrated above, this assumption introduces a bias which should be expected to lead to around one third of the performance gap being misclassified as bad timing. This bias would account for 0.60% of their 0.71% estimated bad timing effect, implying very little evidence of actual bad timing.

4. Simulation Evidence

We made a number of approximations in the derivations above, but in this section we find that simulation evidence is consistent with these results.

These simulations generated monthly returns and then the corresponding monthy investment inflow/outflow using $a_t = w(r_t - \mu)$. The GM and IRR returns were calculated for these simulated series. The resulting performance gap (IRR-GM) was then decomposed using the Hayley (2014) method, but with the monthly returns always initially assumed equal to the observed sample GM. Where this decomposition records a significant timing effect we know that this is entirely spurious, since monthly returns were generated as a random walk which cannot be consistently timed.

The simulations in Table 1 below were conducted using lognormal monthly returns over a wide range of different investment horizons. These results use $w=3$ to determine the size of the return-chasing effect in the investment flows, since this generates performance gaps of the same sort of magnitude as are observed in practice.[3] The monthly volatility of returns r_t is set to 4% to represent the volatility of a well diversified portfolio (annual standard deviation of 14%) and 8% to represent periods of exceptional volatility when the results derived above might be expected to be less accurate approximations.

[3] Using a smaller value of w would scale down both the simulated and theoretical effects proportionately, leaving their comparative values unchanged. Sirri and Tuffano (1998) and Philips et al. (2012) found that on average the effect is roughly $w \approx 0.3$, although it can be much larger for the best performing funds. Darendeli (2017) also argues that the return-chasing effect can be much larger than 0.3 for some funds (where the appropriate return metric is included in the monthly factsheet). Our derivation above shows that the average bias observed will be heavily influenced by outliers, so the net effect is likely to be much larger than would be generated using $w=0.3$, so for this reason we use a much larger figure ($w=3$) in order to determine whether our mathematical derivation is acceptably accurate when we generate hindsight effects of around the size found in empirical studies.

Table 1 – Simulations: apparent timing effects as percentage of total performance gap

Monthly s.d. of returns		Investment horizon (years) 1	2	4	8	16
4%	Performance Gap	-0.17%	-0.20%	-0.22%	-0.23%	-0.24%
	Timing (% of gap)	42.1%	37.4%	35.1%	33.5%	31.0%
8%	Performance Gap	-0.71%	-0.80%	-0.86%	-0.92%	-0.95%
	Timing (% of gap)	42.1%	36.4%	34.0%	29.9%	28.2%

The simulations do indeed generate a consistently negative performance gap. The size of this gap is not of central interest here, since it is driven by our choice of *w*. Instead, our key interest is in the accuracy of our result that when the decomposition by initially sets monthly returns equal to the sample mean return it will mistakenly attribute around one third of the performance gap to bad investor timing.

For investment horizons of only one or two years rather more than one third of the performance gap is attributed to bad timing, but any performance measures estimated over such a short time horizons are likely to be very noisy. The horizons considered by typical investors are likely to be longer than this, and for these (between 4 and 16 years) we find that our conclusion that around one third of the observed performance gap will be spuriously attributed to bad timing is a fairly accurate estimate. Even in exceptionally volatile markets (monthly s.d.=8%) one third is likely to be an acceptable ready reckoner. For very long horizons second order effects appear to become more significant, implying that the spurious timing effect will account for less than one third of the measured performance gap.

These simulations are consistent with the result derived in Section 3: that decomposing the performance gap using the sample mean gives rise to a spurious timing effect equal to around one third of the performance gap.

5. Concluding Remarks

A large number of papers have used the “performance gap” between the IRR and GM returns as an estimate of the effect that the timing of investment flows has had on the returns earned by investors. These suggest that bad timing has substantially reduced investor returns.

Academic studies are typically been skeptical of claims that it is easy to generate consistent outperformance by predicting market returns. By contrast, studies based on performance gaps have been surprisingly willing to conclude that large groups of investors have consistently generated very substantial negative alpha entirely by accident (since investors were presumably not trying to time their investment flows badly).

Hayley (2014) demonstrated that a misleading hindsight effect is inherent in the performance gap when investors chase returns. Vicente and Muňoz (2018) investigated the performance gap for US domestic equity mutual funds and argued that even after correcting for this hindsight effect, bad timing had significantly reduced investor returns. However, as demonstrated above, the assumptions used in their decomposition should be expected to lead to around one third of the performance gap being misclassified as bad timing. This would account for the large majority of their estimated bad timing effect, leaving very little evidence of actual bad timing. This is an important result, since the apparent bad timing effect was economically significant.

However, the contribution of this paper is not just to correct this particular finding, but also to offer a methodological improvement: a more robust method for interpreting performance gap data, taking appropriate account of the hindsight effect and also the bias resulting from using the sample GM return in the decomposition. This will allow future studies to derive more accurate estimates of the genuine effects of the timing of investment flows.

We know that investor flows can be driven by entirely spurious data (e.g. Phillips et al., 2012) showed that investor flows into mutual funds increase when the latest annual return rises purely because a significant monthly loss one year ago had dropped out of this calculation of this annual return. This is entirely spurious, since this effect is predictable and gives no new information on the skill of the managers concerned. This suggests that we should be concerned that investors will be similarly responsive to entirely spurious data on the "performance gap" (defined as IRR-GM) that similarly has no relevance to the fund manager's skill. It is likely to be very difficult to remove predictable behavioural biases from investors allocation of their savings, but we should at least aim to stop investors from being presented with misleading performance indicators.

References

Chieh-Tse Hou, T. (2012). Return persistence and investment timing decisions in Taiwanese domestic equity mutual funds. *Managerial Finance*, *38*(9), 873-891.

Cornell, B., Hsu, J. and Nanigian, D. "Does Past Performance Matter in Investment Manager Selection?" *Journal of Portfolio Management* 43.4 (2017): 33-43.

Darendeli, A., (2017) "Do Retail Investors Respond to Summary Disclosures? Evidence from Mutual Fund Factsheets", working paper.

Dichev, I. D. (2007). What are stock investors' actual historical returns? Evidence from dollar-weighted returns. *American Economic Review*, 97(1), 386-401.

Dichev, I. D., & Yu, G. (2011). Higher risk, lower returns: What hedge fund investors really earn. *Journal of Financial Economics*, 100(2), 248-263.

Friesen, G. C., & Sapp, T. R. (2007). Mutual fund flows and investor returns: An empirical examination of fund investor timing ability. *Journal of Banking & Finance*, 31(9), 2796-2816.

Hayley, S. (2014). "Hindsight Effects in Dollar-Weighted Returns." *Journal of Financial and Quantitative Analysis*, 49(1), 249-269.

Hsu, J., Myers, B. W., & Whitby, R. (2016). Timing Poorly: A Guide to Generating Poor Returns While Investing in Successful Strategies. *Journal of Portfolio Management*, 42(2), 90-98.

Muñoz, F. (2016). Cash flow timing skills of socially responsible mutual fund investors. *International Review of Financial Analysis*, *48*, 110-124.

Muñoz, F., & Vicente, R. (2018). Hindsight effect: What are the actual cash flow timing skills of mutual fund investors? *Journal of Empirical Finance*, 45, 181-193.

Navone, M., & Pagani, M. (2015). Brothers from Different Mothers How Distribution Fees Change Investment Behavior. *Journal of Banking & Finance*, *51*, 12-25.

Perkins, T (2018), "Expected Dollar-Weighted Rate of Return When Investment Growth Rate Is a Linear Function of Return", Working paper.

Phillips, B., Pukthuanthong, K., & Rau, P. R. (2012, September). Do Mutual Fund Investors Chase False Returns? *The 2012 Annual Meeting of the Academy of Behavioral Finance & Economics (p. 111)*.

Sirri, E. R., & Tufano, P. (1998). Costly Search and Mutual Fund Flows. *Journal of Finance* 53(5), 1589-1622.

Annex: Estimating the Hindsight Effect

Above we considered substituting in r_n during the decomposition of the performance gap to replicate the process of identifying the timing component of the gap. Now we look at the effect of substituting in a_n to derive the corresponding hindsight effect. We start with a version of equation (12), with r_n already set to its actual value, and future values set to $\bar{r}$:

$$\mathrm{IRR} \approx \frac{\sum_{t=1}^{n}\left(1+\sum_{j=1}^{t} a_{j-1}\right) r_t + \left(1+\sum_{t=1}^{n-1} a_t\right) \sum_{n+1}^{T} \bar{r}}{\sum_{t=1}^{n-1}\left(1+\sum_{j=1}^{t} a_{j-1}\right) + \left(1+\sum_{t=1}^{n-1} a_t\right)(\mathrm{T}-\mathrm{n}+1)}$$

$$\mathrm{IRR} \approx \frac{1}{T}\left(\sum_{t=1}^{n}\left(1+\sum_{j=1}^{t} a_{j-1}\right) r_t + \left(1+\sum_{t=1}^{n-1} a_t\right)\sum_{n+1}^{T} \bar{r}\right)\left(2-\frac{1}{T}\left(\sum_{t=1}^{n-1}\left(1+\sum_{j=1}^{t} a_{j-1}\right) + \left(1+\sum_{t=1}^{n-1} a_t\right)(\mathrm{T}-\mathrm{n}+1)\right)\right)$$

We then substitute in a_n. The numerator changes by $a_n(T-n)\bar{r}$ and the denominator changes by $a_n(T-n)$:

$$\mathrm{IRR} \approx \frac{\sum_{t=1}^{n}\left(1+\sum_{j=1}^{t} a_{j-1}\right) r_t + \left(1+\sum_{t=1}^{n-1} a_t\right) \sum_{n+1}^{T} \bar{r} + a_n(T-n)\bar{r}}{\sum_{t=1}^{n-1}\left(1+\sum_{j=1}^{t} a_{j-1}\right) + \left(1+\sum_{t=1}^{n-1} a_t\right)(\mathrm{T}-\mathrm{n}+1) + a_n(T-n)}$$

$$\mathrm{IRR} \approx \frac{1}{T}\left(\sum_{t=1}^{n}\left(1+\sum_{j=1}^{t} a_{j-1}\right) r_t + \left(1+\sum_{t=1}^{n-1} a_t\right)\sum_{n+1}^{T} \bar{r} + a_n(T-n)\bar{r}\right)\left(2 - \frac{1}{T}\left(\sum_{t=1}^{n-1}\left(1+\sum_{j=1}^{t} a_{j-1}\right) + \left(1+\sum_{t=1}^{n-1} a_t\right)(\mathrm{T}-\mathrm{n}+1) + a_n(T-n)\right)\right)$$

$$\text{dIRR} \approx \frac{a_n}{T}(T-n)\bar{r}\left(2-\frac{1}{T}\left(\sum_{t=1}^{n-1}(1+\sum_{j=1}^{t}a_{j-1})+\left(1+\sum_{t=1}^{n-1}a_t\right)(\mathrm{T-n+1})\right)\right)$$

$$-\frac{a_n(T-n)}{T^2}\left(\sum_{t=1}^{n}\left(1+\sum_{j=1}^{t}a_{j-1}\right)r_t+\left(1+\sum_{t=1}^{n-1}a_t\right)(T-n)\bar{r}+a_n(T-n)\bar{r}\right)$$

Noting that $E[a_n a_{n-j}]=0$:

$$\text{E[dIRR]} \approx \text{E}\left[\frac{a_n}{T}(T-n)\bar{r}-\frac{a_n(T-n)}{T^2}\left(\sum_{t=1}^{n}r_t+(1+a_n)(T-n)\bar{r}\right)\right]$$

$$\approx \text{E}\left[\frac{a_n r_n}{T^2}(T-n)-\frac{a_n(T-n)}{T^2}\left(r_n+(1+a_n)\frac{r_n}{T}(T-n)\right)\right]$$

$$\approx \text{E}\left[-\frac{a_n r_n(T-n)^2}{T^3}(1+a_n)\right]$$

Substituting $a_n=w(r_n-\mu)$, summing over n=1 to T, and approximating for large T, small E[a_n]

$$\text{E[dIRR]} \approx -\frac{w\sigma^2(T-1)^2}{3T^3}(1+E[a_n]) \approx -\frac{w\sigma^2}{3}$$

=> $\text{E[dIRR]} \approx \frac{\mathrm{w}\sigma^2}{3}$ for large T, consistent with equations (10) and (23) for the overall performance gap and timing effect respectively.