COGNITIVE ERROR IN THE MEASUREMENT OF INVESTMENT RETURNS

Simon Hayley

Thesis submitted for the award of PhD in Finance, Cass Business School, City University London, comprising research conducted in the Faculty of Finance, Cass Business School.
April 2015

Table of Contents

List of Tables and Figures

Chapter 2: Dollar Cost Averaging: the Role of Cognitive Error

Chapter 3: Dynamic Strategy Bias of IRR and Modified IRR – The Case of Value Averaging

Chapter 4: Measuring Investors' Historical Returns: Hindsight Bias In Dollar-Weighted Returns

Declaration

I hereby grant powers of discretion to the University Librarian to allow this thesis to be copied in whole or in part without further reference to me. This permission covers only single copies made for study purposes, subject to normal conditions of acknowledgement.

Abstract

This thesis identifies and quantifies the impact of cognitive errors in certain aspects of investor decision-making. One error is that investors are unaware that the Internal Rate of Return (IRR) is a biased indicator of expected terminal wealth for any dynamic strategy where the amount invested is systematically related to the returns made to date. This error leads investors to use Value Averaging (VA). This thesis demonstrates that this is an inefficient strategy, since alternative strategies can generate identical outturns with lower initial capital. Investors also wrongly assume that the lower average purchase cost which is achieved by Dollar Cost Averaging (DCA) results in higher expected returns. DCA is a similarly inefficient strategy.

Investors also adopt strategies such as Volatility Pumping, which appears to benefit from high asset volatility and large rebalancing trades. This thesis demonstrates that any increase in the expected geometric mean associated with rebalancing is likely to be due to reduced volatility drag, and that simpler strategies involving lower transactions costs are likely to be more profitable. Academic papers in highly-ranked journals similarly misinterpret the reduction in volatility drag achieved by rebalanced portfolios, mistakenly claiming that it results from the rebalancing trades "buying low and selling high".

The previously unidentified bias in the IRR has also affected an increasing number of academic studies, leading to misleadingly low estimates of the equity risk premium and exaggerated estimates of the losses resulting from bad investment timing. This thesis also derives a method for decomposing the differential between the GM return and the IRR into (i) the effects of this retrospective bias, and (ii) genuine effects of investor timing. Using this method I find that the low IRR on US equities is almost entirely due to this bias, and so should not lead us to revise down our estimates of the equity risk premium. This method has wider applications in fields where IRRs are used (e.g. mutual fund performance and project evaluation).

In identifying these errors this thesis makes a contribution: (i) to the academic literature by correcting previous misleading results and improving research methods; (ii) to investment practitioners by identifying avoidable errors in investor decision-making. It also makes a contribution to the field of behavioural finance by altering the range of investor behaviour which should be seen as resulting from cognitive error rather than the pursuit of different objectives.

Chapter 1

Summary and Motivation

This chapter summarises the key findings of this thesis, and sets out the relationship between these results and existing financial research within this field.

This thesis identifies and quantifies the impact of cognitive errors in certain aspects of investor decision-making. These errors relate primarily to the behaviour of retail investors, although as will be discussed below, institutional investors and academics are not immune from these errors. They are:

(1) Investors wrongly assume that the lower average purchase cost which is achieved by *Dollar Cost Averaging* (DCA) results in higher expected returns. Far from improving returns, DCA is a demonstrably inefficient strategy (chapter 2).

(2) Investors are unaware that the Internal Rate of Return (IRR) is a biased indicator of expected terminal wealth for any dynamic strategy where the amount invested is systematically related to the returns made to date. Specifically, this error leads investors to follow *Value Averaging* (VA), a strategy which generates attractive IRRs. Chapter 3 demonstrates that these high IRRs are misleading, and that VA is an inefficient strategy, since alternative strategies can generate identical outturns with lower initial capital.

(3) Investors are misled by the maths of rebalanced portfolios, leading them to adopt strategies such as *Volatility Pumping* which aim to increase the scale of rebalancing trades. I demonstrate instead that any increase in the expected geometric mean associated with rebalancing is likely to be due to reduced

volatility drag, and simpler strategies involving lower transactions costs are likely to be more profitable. This is covered in chapter 5.

Academic studies in the highest ranked journals have also been affected by errors (2) and (3). The bias in the IRR has not previously been identified, so the difference between historical geometric mean (GM) and IRR figures has wrongly been attributed to bad timing by investors. This has led to excessively low estimates of the equity risk premium (Dichev, 2007) and exaggerated estimates of the losses resulting from bad investment timing by mutual funds (Friesen and Sapp, 2007, Clare and Motson, 2010) and hedge funds (Dichev and Yu, 2011). Similarly, the increased expected geometric mean associated with rebalanced portfolios has been wrongly attributed to rebalancing trades "buying on downticks and selling on upticks" (e.g. Ferhnolz and Shay (1982), Luenberger (1997), Willenbrock (2011)).

In identifying these errors this thesis makes a contribution: (i) to the academic literature by correcting previous misleading results and improving research methods; (ii) to investment practitioners by identifying avoidable errors in investor decision-making. It also makes a contribution to the interpretation of behavioural finance. The expansion of behavioural finance over recent years has seen investor behaviour which would previously have been regarded as misguided instead being seen as a sensible way of pursuing a much richer set of possible underlying motives. By contrast, this thesis identifies a number of areas where it is reasonable to regard investors as making cognitive errors which lead them into investment strategies which are inefficient means of pursuing their underlying objectives.

This thesis contains four substantive chapters. In brief:

(1) Dollar Cost Averaging - The Role of Cognitive Error

Dollar Cost Averaging (DCA) is an investment strategy which remains very popular even though previous research has long since established that it is mean-variance inefficient. More recent research has focused on identifying alternative investor objectives which can explain why the strategy nevertheless remains popular. In this chapter I demonstrate that some of these explanations (eg. loss aversion/prospect theory) must be rejected, since DCA is an inefficient strategy for investing available funds regardless of the risk weighting that is applied to the probability distribution of possible terminal wealth outturns. Plausible explanations for DCA's popularity also need to address the arguments that are commonly used by those who recommend this strategy. Regret avoidance and the need for investor discipline are sometimes alluded to, but the argument that is almost invariably made is that DCA achieves an average purchase cost which is always lower than the average price of the securities concerned over the investment period. It seems intuitively obvious that a lower average purchase cost must raise expected returns, and previous research has failed to identify the error in this argument. This chapter explicitly identifies the error in this argument, and hence suggests that cognitive error rather than behavioural motives is the most plausible explanation for DCA’s continued popularity.

(2) Dynamic Strategy Bias of IRR and Modified IRR – the Case of Value Averaging

This chapter demonstrates that the internal rate of return (IRR) is a biased measure when the scale of new investment is correlated with the return recorded to date. The IRR is normally defined as the root of a polynomial equation representing investor cash flows, but it is equivalent to a weighted average of individual period returns where the weight given to each period is proportional to the present value of the portfolio value at the start of the period. The intuition behind this bias is that an additional investment increases the weight given to future returns and reduces that given to past returns. For example, investing ahead of above-average returns is good timing which will boost expected terminal wealth. However, investing after unusually low returns reduces the weight given to these low returns in the IRR calculation. This will boost the expected IRR, but without increasing expected terminal wealth. This is the source of the bias.

Value averaging (VA) is a strategy which systematically invests more following poor returns, thus keeping the portfolio value at a pre-determined target level. This strategy implies an inherent correlation between returns to date and net investment flows. This chapter demonstrates that unless there is systematic autocorrelation of returns the high IRRs recorded by VA are entirely due to this bias. I demonstrate that VA is, like DCA, an inefficient strategy regardless of investors risk preferences and I quantify the resulting efficiency losses. I discuss whether behavioural finance factors could explain VA popularity, but again find that cognitive error is a more plausible explanation.

(3) Measuring Investors’ Historical Returns: Hindsight Bias In Dollar-Weighted Returns

A growing number of papers, often in leading journals, use the IRR as an indication of investors’ skill (or lack of it) in timing their investments. An IRR lower than the corresponding geometric mean return (GM) is taken to indicate that on average investor cash flows were badly timed. The scale of this effect is substantial: mainstream US equity returns are found to be 1.3% lower using the IRR, which is taken as evidence of consistently bad investor timing (Dichev, 2007), and would imply that the equity risk premium must be correspondingly lower than previously thought. Other studies use the same technique to conclude that investors have shown consistently bad timing in their investments into mutual funds and hedge funds. In this chapter I demonstrate that the method used by these studies is subject to the same bias as Value Averaging, since return chasing (increased investment following unusually strong periodic returns) biases the IRR downwards, and that this is mistakenly interpreted as bad investment timing.

I derive a method for decomposing the differential between the GM and IRR into (i) the effects of this retrospective bias, and (ii) genuine effects of investor timing. Using this method I find that the low IRR on US equities is almost entirely due to this bias (due to return-chasing behaviour by investors), with negligible impact from bad timing. Thus low IRRs should not be taken to imply a correspondingly low figure for the equity risk premium. This is an important contribution to the literature, because the equity risk premium is one of the key parameters of finance theory, and because it implies a re-evaluation of an expanding field of the existing literature. The method derived here for decomposing the return differential also has wider applications in other fields where IRRs are used (e.g. mutual fund performance measurement and project evaluation). Furthermore, the fact that academic research has been subject to the same unidentified bias supports the hypothesis that the popularity of Value Averaging is due to investors making the same cognitive error.

A version of this paper has been published under the title 'Hindsight Effects in Dollar-Weighted Returns' in the *Journal of Financial and Quantitative Analysis*, Vol. 49, No. 1, February 2014.

(4) Diversification Returns, Rebalancing Returns and Volatility Pumping

It is widely claimed by both academics and practitioners that periodic rebalancing of portfolios to keep asset weights constant will directly boost geometric returns by buying on downticks and selling on upticks. This chapter refutes this claim by showing that comparable improvements arise even without rebalancing. Volatility pumping is a strategy which appears to benefit from high asset volatility and large rebalancing trades. The popularity of this strategy again appears to be due to cognitive error which leads investors and academics to misinterpret a reduction in "volatility drag" as the profits from this active strategy.

This chapter makes a contribution in both academic and practical terms. It refutes the claim made in a number of papers in highly-ranked journals that rebalancing strategies are inherently "buying on downticks and selling on upticks". This also has important practical implications, since rebalancing strategies that claim to benefit from this "rebalancing return" are widely recommended to investors, and are implemented by large fund managers. This chapter demonstrates that the real source of return on such strategies is an implied risk premium on the assets used, and that strategies motivated by "rebalancing returns" are likely to lead to inefficient portfolio allocations and high transactions costs.

The overall theme of this thesis it that these four chapters identify a number of errors in the methods commonly used to evaluate investment performance and suggest more meaningful alternative methods. This has important implications both for investment practitioners (since it shows that some popular investment strategies are likely to be counterproductive) and for the academic literature (since some of the same errors have resulted in misleading conclusions being drawn in previous published research).

Chapter 2

Literature Review

This chapter considers the overarching themes of this thesis, and places them within the context of the wider literature. Each subsequent chapter contains a more specific literature review which considers the research which is relevant to the chapter concerned. The overall theme of this thesis is cognitive errors which lead investors to time their investments badly. The relevant literature thus falls naturally into two sections: (i) issues of investor cognitive error; (ii) evidence of bad timing by investors. We consider each in turn.

Preferences and Beliefs

This thesis identifies a number of forms of investor behaviour that it argues should be seen as cognitive errors. However, we need to exercise great care in separating cognitive errors from the results of other beliefs and preferences.

Finance theory has seen dramatic changes over recent years as it became increasingly clear that investor behaviour was often not adequately explained by the rational pursuit of risk-adjusted expected wealth. Behavioural finance has expanded to offer alternative explanations. Barberis and Thaler (2003) make two distinctions which will be useful when we consider a topic as broad as behavioural finance: between limits to arbitrage and psychology and, within psychology, the distinction between beliefs and preferences.

Some behavioural finance effects clearly reflect alternative preferences which substitute for the utility functions of traditional finance theory. For example, *prospect theory* (Kahneman and Tversky, 1979; Shefrin and Statman, 1985) incorporates *loss* *aversion, framing* (where these losses are assessed relative to a starting point which may seem arbitrary in terms of the investor's stated long-term goals) and *excess sensitivity* to small probabilities (Kahneman and Tversky, 1972). Other examples are:

- *Regret aversion* (Loomes and Sugden, 1982), which suggests that investors aim to reduce the degree of ex post unhappiness they associate with their choices.
- *Ambiguity aversion* (Ellsberg, 1961, Einhorn and Hogarth, 1986), which suggests that investors are much more averse to situations where risks cannot be estimated.
- *Myopic loss aversion* whereby investors look at high short term possible risk, rather than the lower long-term risks that are likely to be more consistent with their underlying objectives. This can help explain why the equity risk premium seems to be much larger than would be consistent with plausible consumer preference parameters (Benartzi and Thaler, 1995, Haigh and List, 2005).
- By contrast to the exponential discounting of finance textbooks, the *hyperbolic discounting* that has been observed in practice (e.g. Laibson, 1997) leads to time inconsistencies that are hard to reconcile with consistent longer-term preferences.

Other behavioural finance effects are more ambiguous, since they may reflect either preferences or fallacious beliefs. The *disposition effect* (e.g. Shefrin and Statman, 1985) describes an unwillingness to realise losses. Odean (1998) notes that it can be regarded as a preference if motivated by the desire to avoid the pain of realising a loss, or a cognitive error if based on a misplaced belief (gamblers fallacy) that returns are inherently likely to show negative autocorrelation. *Mental accounting* (Tversky and Kahneman, 1981, Thaler, 1985) can be seen as another explanation for the disposition effect, as investors regard realised and unrealised losses as being different in kind even though one can always be converted into the other, subject to transaction costs. The *house money effect* whereby individuals are more willing to take risks with "winnings" can be seen as another result of mental accounting, as investors regard winnings as different in kind from other forms of wealth. However this effect could in some instances also represent a pure preference implicit in a greater risk appetite after gains have been made.

The explosion of research interest into behavioural finance has led to an ever-increasing number of identified "biases". *Overconfidence* (Fischoff, Slovic and Lichtenstein, 1977; Barber and Odean, 2001; Gervais and Odean, 2001) is a fallacious belief which might be useful in an evolutionary cost-benefit analysis, but which appears to be very costly when applied to finance. Barber and Odean (2000) use discount broker accounts to show that retail investors lose out by trading too much. This could be a preference if trading was enjoyable, but Barber and Odean regard it as a fallacious belief resulting from overconfidence and *attribution bias* (Larson, 1977) as investors interpret profitable trades as reflecting their own skill rather than luck, leading to an incorrect belief that this trading adds value.

Indeed, fallacious beliefs might in principal be regarded as the results of cognitive error, since it could be argued that evidence exists which would contradict these beliefs, if only the individuals concerned correctly assessed this evidence. However, this would be an unconvincingly broad definition of cognitive error, since adherence to fallacious beliefs can also very plausibly be seen as the result of strong emotional attachment to these beliefs, even in the light of contradictory evidence. Specifically, the theory of *cognitive dissonance* (Festinger, 1957) suggests that avoiding an objective assessment of the evidence would be an effective way to avoid the psychological distress that would be caused when the evidence conflicts with the investor's pre-existing beliefs.

Furthermore, as we saw above, the wide range of different types of behavioural finance effects that have been identified means that it is hard to distinguish *preferences* from *mistakes*. Some behavioural finance effects which express alternative preferences are likely to conflict with investors' stated goals, such as maximising risk-adjusted wealth or risk-adjusted lifetime consumption. For example mental accounting, loss aversion and prospect theory more generally apply arbitrary frames to decision-making and consider profits and losses relative to reference points which may have very little relevance to investors' stated long-term financial goals. But even in such cases it would be misleading to consider this contradiction as a cognitive error. Instead we merely have a conflict between investors' stated preferences and the preferences which are revealed by these behavioural finance effects. In such a situation investor behaviour may represent an entirely sensible trade-off between these goals, rather than a cognitive error.

Nor can we have any presumption that decisions that are influenced by emotion are more likely to involve cognitive error. The distinction between rational *homo economicus* and his emotional real-world human counterparts has been blurred as emotional responses have been shown to be a vital part of decision-making (Becharia et al, 1994). Monitoring of physiological data for professional traders also suggests that emotional stimulation plays a part in the processing of real time financial risk (Lo and Repin, 2002). More specifically, there is some evidence that there is an optimal degree of emotional involvement, with traders who show an unusually high correlation between their emotional state and their daily performance tending to generate lower overall levels of profitability (Lo, Repin, and Steenbarger, 2005). Coates and Herbert (2008), find that traders generally make greater profits on days when they have higher than average levels of testosterone, although the causation here is inevitably hard to pin down. Such links between trader behaviour and emotional state have also been suggested as the underlying explanation of why in aggregate equity market prices have been found to be related to the amount of sunshine (Hirshleifer and Shumway, 2003) and reduced sunlight during winter (Kamstra et al, 2003).

Bounded rationality, heuristics and acceptable levels of error

The concept of *bounded rationality* (Simon, 1955) also opens up another challenge to our ability to attribute some types of investor behaviour to cognitive error. Full optimisation of portfolio choices would require a massive amount of information and computation. Obtaining and processing all this information would be a huge task. Given that investors' time and cognitive ability is inevitably limited, full optimisation is clearly an unrealistic benchmark.

Moreover, if time and cognitive ability are scarce resources then standard microeconomic theory suggests that the limited amount that is available should be deployed only to the most productive uses, recognising its high opportunity cost. Thus not only is full optimisation unrealistic, it would also be a major error for investors to aim for anything approaching full optimisation. Instead, it may be entirely optimal to use *heuristics* (simple rules which are designed to give adequately good decision-making whilst avoiding cognitive overload). More bluntly: investors should aim merely for an acceptably low level of avoidable investment errors ─ aiming for zero errors would itself be a much greater error.

Lo (2012) formalises this intuition into the Adaptive Markets Hypothesis (in deliberate contrast to the Efficient Markets Hypothesis). This assumes that, given their inevitable cognitive limitations, investors are boundedly rational and make use of simple heuristics in their decision making. These heuristics evolve as investors abandon those heuristics which have been shown by events to be inappropriate. Investors who fail to adapt their heuristics cease to be competitive. Outdated heuristics appear to researchers as behavioural biases. The underlying dynamic evolution process can be seen as an efficient means of coping with limited cognitive abilities. It can also help explain booms and crashes as investors systematically alter the heuristics that they use.

The widespread use of heuristics thus suggests that we must be cautious in regarding investor behaviour as resulting from cognitive error, since even though the use of an individual heuristic in a specific situation might be shown to be a mistake, use of such heuristics more generally may nevertheless be optimal when cognitive effort is scarce. Investors' goal should be an adequately low, but non-zero, level of cognitive error. Examples of such heuristics which might be justified on this criterion are:

- *Naive diversification* (e.g. the 1/n rule is found to be a good alternative to more complex theoretical mean-variance optimisation rules, DeMiguel et al, 2009).
- The *representativeness* and *availability* heuristics (see below).
- *Conservatism* (Barberis et al., 1998), which can be seen as a means of avoiding cognitive effort by sticking to existing beliefs.
- *Anchoring*, whereby decisions are strongly influenced by an early piece of information that may be of very limited relevance (Tversky and Kahneman, 1974).
- *Herd behaviour*, resulting from information free riding, which may be entirely rational for an individual investor, even if it could give rise to inefficient collective outturns (Banerjee, 1992, Bikhchandani and Hirshleifer, 1992)

Heuristics are, of course, at work much more widely than just in financial decisions. Kahneman (2011) ─ following Stanovich and West (1998) ─ labels two alternative modes of cognition in everyday decision-making: "system 1" (quick, intuitive and requiring little cognitive effort) and "system 2" (slow and considered). Goel et al. (2000) presents evidence that these two systems correspond to activity in different parts of the brain. In many everyday decisions rapid "system 1" decision-making is entirely adequate. Indeed, in an evolutionary context it is straightforward to argue that for some life-threatening situations such rapid decisions are absolutely vital (for example in the triggering of the "fight or flight" reflex in response to the very earliest perception of a threat), and that the occasional mistakes made when system 1 reaches an inappropriate conclusion are a small price to pay, since such mistakes are generally unlikely to be life-threatening.

One useful application of this distinction is the *representativeness* heuristic, which classes new objects (or events) as belonging to the group of already-known objects that most resemble them. Probably the most famous example of an error resulting from the use of this heuristic is in Tversky and Kahneman (1983) where, responding to a description of the fictional "Linda", participants in the study judged that she was more likely to be a "feminist and bank teller" than simply "a bank teller". The more detailed category seemed more representative of the description that they had been given, but this is a clear error, since the probability of her being in the larger group "bank teller" must exceed the probability of her being in a subset of this group. This is clearly an error of logic (the conjuction fallacy) rather than a matter of preference. It can also be seen as an example of "base rate neglect", where the more representative-seeming category was chosen, ignoring the relative size of the populations involved. However, this leaves open the important question of whether in a wider context the use of this heuristic could nonetheless be optimal behaviour, since the cost of occasional errors is outweighed by the costs of the huge cognitive effort and correspondingly slow response times that would be required in using deliberative logic (system 2) rather than a simple heuristic (system 1).

Many similarly error-prone heuristics have been identified. For example, the availability heuristic classifies new objects into the group with the representative example which can most easily be recalled (Tversky and Kahneman, 1973). This implies that recently-experienced examples can exercise an excessive influence on our judgement. Consistent with this, Barber and Odean (2006) find that retail investors tend to buy "attention grabbing" stocks, e.g. those that have recently featured in the media. Again, this could be regarded as a cognitive error, or a rational attempt to cope with information overload.

However, even though the general use of imperfect heuristics can be defended as optimal, we can still consider changing the heuristics used where these can be shown to be inappropriate, even if these work at a subconscious level. Outside the field of finance, the idea that ingrained behavioural habits can be unlearned has been gaining ground. Cognitive Behavioural Therapy (CBT) is predicated on the observation that patients persist in reacting to situations in ways which are counterproductive to their stated long-term goals, and that this behaviour can be modified by helping patients to examine their reactions and identify situations in which they tend to interpret and respond to situations in ways which are unrealistic or simply not helpful. CBT remains controversial, in particular because the direct involvement of the patient in the treatment process means that double blind trials are not possible. However belief in the efficacy of CBT has been gradually becoming more mainstream. For example, the UK government has adopted an objective of increasing state-funded provision of therapies such as CBT in the National Health Service (NHS, 2013), based on evidence that CBT is particularly effective for problems such as anxiety, depression, eating disorders, post-traumatic stress disorder and drug misuse (NHS Direct, 2014).

This thesis will argue that the investor behaviours identified in it can validly be interpreted as resulting from cognitive errors. Their use could in principle be defended as representing heuristics that are generally valid (strategies which buy at a lower average cost or generate a higher IRR generally will increase expected profits, although in the specific context of the DCA and VA strategies they do not). Even so, I demonstrate that modification of such heuristics could be argued to improve investor welfare, just as it does in CBT.

Cognition and IQ

A massive amount of research has considered investor preferences, and in particular how these differ from the assumptions of the expected utility hypothesis. Comparatively little research has directly addressed issues of cognition.

Some has focused on the decline in cognition associated with increased age, and the consequent impact on decision-making (e.g. Agarwal et al, 2009). This runs into the problem of the collinearity of age and levels of experience, and studies which attempt to separate these two explanations then face the problem of observationally equivalent cohort-specific effects. However identification is made much easier because the decline in cognition is found to be highly non-linear, accelerating sharply at advanced ages. For example, Korniotis and Kumar (2011) find that older/more experienced investors are more likely to use simple heuristics that reflect their greater knowledge, but that the advantages of their greater experience are more than outweighed by the adverse effects of ageing on cognition. Older investors suffer less from (i) the disposition effect (Dhar and Zhu, 2006), (ii) under-diversification (Goetzmann and Kumar, 2005) and (iii) overconfidence (Barber and Odean 2001). Korniotis and Kumar (2007) conclude that more experienced investors have preferences which are closer to rationality (stronger preference for diversification, less frequent trading and greater tendency for year-end tax loss selling), but are less skilful in implementing these preferences successfully. More generally, the impact of age on investor behaviour has also been explained in terms of reduced cognitive ability leading to reduced ability to control the inherent emotional responses to decisions.

Other studies link financial behaviour directly to IQ. This might appear to be useful in identifying cognitive errors, since we would anticipate that these errors will be less frequent among investors with higher IQ. For example, Grinblatt et al. (2012) find that high IQ investors are less affected by the disposition effect and show superior market timing, stockpicking and trade execution. However, interpreting this behaviour is made more difficult by a negative correlation between cognitive ability and susceptibility to emotional biases. Questions such as the following are widely used as tests of cognitive ability: "a bat and a ball cost $1.10. The bat costs $1 more than the ball. How much does the ball cost?". Relating the answers to the results of separate tests shows that those who wrongly answer 10c tend to be significantly less patient (i.e. unwilling to accept delayed rewards) than those who correctly answer 5c. Indeed, Frederick (2005) notes that "the relation is sometimes so strong that the preferences themselves effectively function as expressions of cognitive ability". Hence he interprets questions such as the bat and ball problem as measuring *cognitive reflection* which he defines as "the ability or disposition to resist reporting the response that first comes to mind" — in other words, the willingness to engage "system 2" rather than just "system 1".

For the purposes of this thesis, the bat and ball and similar test questions may be testing something which is of limited relevance to investors choosing the trading strategies analysed here. Questions set in a test environment must be answered within a limited time, so the correlation with impatience is perhaps unsurprising. By contrast investors have as long as they wish to consider the effectiveness of an investment strategy. However something that these strategies have in common with the bat and ball question is that both have an easy and intuitive answer, whereas questioning this intuitive answer is likely require much more effort.

Many studies have also found that higher cognitive skills are correlated with greater patience (lower short-term discounting) and less risk aversion (at least over small risks), for example Benjamin et al. (2013), Dohmen et al. (2010), Burks et al (2009), Beauchamp et al. (2011), Shamosh and Gray (2008). IQ has also been linked to lower levels of loss aversion and other behavioural biases (Bucher-Koenen & Ziegelmeyer, 2011, Grinblatt, Keloharju, and Linnainmaa, 2012). There are also gender differences in this association. As Frederick (2005) puts it: “expressed loosely, being smart makes women patient and makes men take more risks”.

Two key explanations have been put forward for this correlation. First, as we saw above, “two systems” theories are interpreted as reflecting a conflict within the brain between short-term emotionally-based decision heuristics and more dispassionate cognition which can in some circumstances override inappropriate emotional responses. Evidence for this internal conflict is provided by the fact that brain imaging studies show that choices involving immediate rewards consistently activate some areas of the brain more than decisions which only involved future rewards (Cohen, 2005). Specifically, the prefrontal cortex has been identified as being involved in the more dispassionate cognitive processes. Kuhnen and Knutson (2005) further supports the distinction between system 1 and system 2 cognitive processes by finding that different areas of the brain are activated in (i) risky and risk seeking behaviour; (ii) risk averse behaviour and riskless choices. Breiter et al. (2001) famously find that the parts of the brain activated by monetary rewards are the same as those activated in cocaine addicts.

Empirical investigation of the relationship between investor behaviour and IQ is also made harder by the fact that those with greater cognitive ability tend to have higher levels of participation in the stock market (Christelis, Jappelli, & Padula, 2010) and financial markets more generally (Cole & Shastry, 2009). Thus investor IQs are likely to exhibit a smaller range of variation than the general population. For all these reasons, comparing IQ and behaviour generally does not allow us to identify cognitive errors. However, in some specific financial decisions can nevertheless be identified as clear-cut mistakes, as discussed in the following section.

Lusardi and Mitchell (2014) model financial literacy as an item of human capital which requires significant investment. They document low levels of financial literacy even though individuals are now being asked to make more complex financial decisions than in the past, as a result of the increasing availability of mortgages and credit cards, the reduced cost of securities transactions and the pressure for individuals to make their own pension provision rather than relying on employer-funded or government-funded schemes.

This approach regards financial decision-making skill as an acquired rather than an innate skill which reflects underlying cognitive ability. However, this distinction does nothing to undermine the view that some financial decisions can be regarded as cognitive errors. The crucial distinction in this thesis is that some financial decisions reflect preferences which are more complex than those of traditional finance theory, whilst other decisions can validly be regarded as cognitive errors. The acid test of this distinction is whether additional education could be expected to alter investor behaviour. Where this behaviour is motivated by preferences such as regret aversion or loss aversion it should not be expected to change. Indeed, where investors’ current behaviour accurately reflects these preferences, altering their behaviour could be regarded as welfare-reducing (when we define welfare to include wider psychological effects as well as narrowly-defined effects on risk-adjusted wealth).

Evidence that financial decision-making can be improved by suitable education actually supports the view that investors sometimes make clearly defined errors. This thesis identifies a number of cognitive errors that have not been previously identified, so if appropriate education improves decision-making in general then we might have more hope that identification of these errors may help reduce their prevalence in future, to the extent that understanding of these errors percolates through to investors. Surveys have shown many borrowers to be unaware of the interests rates that are charged on their mortgages and credit cards (Lusardi 2011, Disney and Gathergood, 2012). There is, of course, likely to be a substantial amount of endogeneity in the relationship between financial literacy and other relevant factors such as IQ. Van Rooij et al (2011) seek to avoid this problem by controlling for the (exogenous) financial experiences of parents and siblings. Lusardi and Mitchell (2009) instrument using the different degrees to which financial education has been mandated in different US states. Both studies find a separate effect of financial education in increasing equity market participation. The panel-based model of Alessie et al (2011) confirms that financial literacy has a positive effect on retirement planning.

Identified Cognitive Errors

We saw above that some apparently flawed investor behaviours can be interpreted as resulting from investor preferences being very different from those assumed by standard finance theory. Other sub-optimal behaviour can be regarded as resulting from the sensible use of acceptably imperfect heuristics. However, other behaviours clearly represent cognitive errors:

- Individuals taking out payday loans are often confused about the implied annualized interest rate they will end up paying. Providing more transparent information about the costs of these loans is found to reduce take-up to some extent, suggesting that at least in some instances this helps the individuals concerned to avoid mistakes they would otherwise have made (Bertrand and Morse, 2011).

- Similary, Agarwal & Mazumder (2013) find that some consumers make mistakes which lead to unnecessarily high credit card or mortgage costs. It is hard to envisage preferences which would make these outturns seem desirable, since these behaviours result in unambiguously high interest costs. Linking these results to military aptitude tests shows that those with better scores in maths are less likely to make such mistakes.
- French and Poterba (1991) find very high levels of home bias among equity investors that it is very difficult to regard such poor diversification this as part of a rational portfolio allocation). The greater effort required to invest in overseas equities can easily be avoided by investing in domestic mutual funds which invest overseas, so it is hard to derive preferences which would justify a failure to take advantage of what is often - and with justification - referred to as "the only free lunch in finance". Similarly, it is hard to justify investors holding a significant proportion of their wealth in the equity of their employers (Benartzi, 2001). It seems reasonable to regard such behaviour as unambiguously mistaken.
- Slow re-financing of adjustable rate mortgages by US households has also been argued to be a clear-cut mistake which is hard to rationalise as a preference (Campbell, 2006).

One very useful acid test of whether a particular behaviour is due to a cognitive error is whether investors change their behaviour when they understand their error. Agarwal & Mazumder (2013) find that in their use of credit cards or in mortgage applications, "some borrowers may not initially identify the optimal strategy, but...then experience what we refer to as a "eureka" moment, after which they will implement the optimal strategy".

Investors' behaviour thus clearly suggests that they identify their own previous behaviour as having been subject to cognitive error.

Two final categories of investor behaviour should also be mentioned. First, there have been plenty of reports of investors choosing the wrong ticker for share trades. These are clearly mistakes which cannot be rationalised by factors such as alternative preferences, but to consider them as cognitive errors would be misleading: they are most plausibly interpreted as due to inattention, rather than investors having reached erroneous conclusions. These inattentive mistakes come in stark contrast to the examples identified in this thesis, which involve investors deliberately choosing investment strategies based on erroneous logic.

Lastly we have the wide range of potentially profitable anomalies which have been identified in market prices of securities (eg. the size and value premia in equities (Banz, 1981, Fama and French, 1992); momentum (Jegadeesh and Titman, 1993, shows that equities which outperformed over the previous six-months tend to also outperform over subsequent six months) and long-term reversals (De Bondt and Thaler, 1985, find that losers calculated over prior three years returns tends to outperform). Such anomalies need to be interpreted with caution. Some have been found not to be persist once they have been identified (eg. Maclean and Pontiff, 2013, document a widespread reduction in the size of predictable anomalies after then have been identified). Those that do persist (most famously the size and value premia, and momentum) can be interpreted as reflecting investor preferences, specifically some form of risk premium (e.g. premium to compensate for "peso risk" in small firms).

More specific anomalies are harder to explain in this way. For example, Houge and Loughran (2000) confirm the continued existence of the "accruals anomaly" and show that a strategy based on cash flows can generate significant excess returns. They interpret this as investors committing a cognitive error by failing to fully realise the degree to which the difference between accruals and cash flows contains useful information. Post-earnings announcement drift (first identified by Ball and Brown, 1968) has also been shown to be widespread and persistent. The failure by sufficient numbers of investors to take advantage of these anomalies, and in doing so remove them, could be due to other limits to arbitrage. Most obviously, trades designed to take advantage of these anomalies need to be profitable enough to compensate for the transactions costs, administrative effort, capital and residual risk involved (the latter including model risk involved in the identification of the anomaly, which will inevitably be hard to quantify, and noise trader risk (DeLong et al., 1990)).

This would suggest that we should not necessarily regard weak or semi-strong market inefficiency in market prices as evidence of widespread cognitive error. Moreover, even when the scale of these anomalies appears too large to be explained by these factors, we have something that remains very different in kind to the types of cognitive error that are identified in this thesis. At worst, failure to participate in profitable trades to take advantage of anomalies represents an *error of omission* (and even this may be rational given the cognitive and administrative effort that must be expended in starting to take advantage of them). By contrast, the errors considered in this thesis are *errors of commission*, where investors have gone to some effort to investigate and execute trading strategies which are based on demonstrable errors of logic.

Assessment: is it really cognitive error?

The expanding field of behavioural finance tells us that we should be cautious in concluding that specific investor behaviours result from cognitive error. Behavioural finance has given us an ever-expanding range of different preferences which can explain behaviours that would previously have been regarded as irrational exceptions to the normative behaviours implied by standard finance theory. Furthermore, the distinction between system 1 and system 2 cognition and the Adaptive Markets Hypothesis suggests that the use of flawed heuristics may be entirely optimal in a wider context as a means of avoiding cognitive overload at an acceptable cost.

However, as discussed above:

1. Some investment decisions are very hard to interpret as preferences (eg. poor decisions which increase mortgage interest and credit card costs).
2. The strategies that are considered in this thesis are not chosen by investors under time pressure, so use of heuristics should not be relevant. Investors make a deliberate decision to use these strategies, and have as long as they like to cogitate (system 2) on their choice.
3. Changes in behaviour sometimes suggest that investors perceive their own previous cognitive errors (Agarwal & Mazumder, 2013).
4. Unlike most other areas of investor behaviour, we can observe the justifications that are given to particular strategies, by looking at how these strategies are described in the media by their proponents. These contain explicit errors of logic, notably that DCA's lower average costs and VA's high IRRs are argued to increase expected profits.
5. Some academic studies have arrived at the same mistaken conclusions even though academics might be expected to have a very different set of motivations in their studies (compared to investors), and again should clearly be expected to be using system 2 cognition.

This thesis identifies three investment strategies for which proponents make plausible-seeming claims which on investigation turn out to be misleading. Are these strategies popular because of cognitive error, or because they appeal to investor preferences? This is more than just a semantic point, since a cognitive error has a very different impact on investor welfare.

First, we can rule out risk preference as an explanation for the popularity of DCA and VA, since chapters 2 and 3 respectively show that these are inefficient strategies regardless of the form taken by the investor's risk preferences. These chapters also discuss the possibility of other motives which might explain why investors choose these strategies. Furthermore, volatility pumping (chapter 5) tends to encourage investors to hold high volatility assets since this increases the scale of the rebalancing trades that are believed to raise profits. Thus risk aversion would deter investors from this strategy.

Dollar cost averaging, value averaging and volatility pumping all appear to offer increased returns, but this thesis shows that in each case this is misleading. This is prima facie evidence of cognitive error. However, behavioural finance offers a very wide range of alternative motives for investors' actions, so it is not hard to identify emotional biases which might be associated with these investment strategies. They are likely to reduce perceived *ambiguity* (by appearing to offer improved portfolio returns regardless of the distribution of asset returns in the market) and *regret* (since each of these strategies requires completely non-judgemental investment/rebalancing transactions at regular intervals, so once an investor has committed himself to follow such a strategy these regular transactions involve little conscious decision-making, suggesting limited scope for regret).

However, even if these psychological benefits are important, cognitive error would be a necessary part of the explanation, since a reduction in *ambiguity* or potential for *regret* only comes about if the investor genuinely believes that these strategies generate increased returns. This belief implies that he must make the future transactions that these strategies require, or else forsake the apparent greater returns of the strategy. This commitment leaves the investor believing that he subsequently has no real discretion over his the timing of his further investments, and with no choices there is no scope for subsequently regretting these choices. Without the belief that the strategy increase returns, the investor would know that he could abandon it at any moment. This would imply that he still has this choice, so he can also suffer regret if this choice turns out badly. Similarly, these strategies appear to unambiguously raise returns. This would make them very attractive in terms of ambiguity aversion, but again only if the investor believes that they generate higher returns.

This suggests that even if it is not the only explanation, cognitive error a necessary part of any explanation of the continued popularity of these investment strategies. It remains plausible that the effects of this error are magnified and cemented by emotional biases, but the role of these emotional biases is then subject to Occam's Razor. As cognitive error is a vital part of any explanation, but behavioural finance factors are not, so the popularity of these investment strategies offers no independent confirmation that these behavioural finance effects are at work. Their popularity can be more simply explained simply as the result of cognitive error. As Freud reputedly said (although this appears to be apocryphal): "sometimes a cigar is just a cigar".

Further evidence of cognitive error comes from the fact we can observe the arguments which are used by proponents of these strategies, and these focus on the claim that they are likely to make the investor richer than alternative strategies. We can also observe that in two of these three examples academic research has been affected by the same cognitive errors as investors. Failure to identify the bias in the IRR has resulted in misleading conclusions being drawn in research on the effects of bad investor timing and the belief that rebalanced strategies generate "rebalancing returns" in addition to the benefits that maintaining diversification has in reducing volatility drag is similarly misleading. Both errors are found in papers published in the very highest ranked journals. We can speculate that in academic research we are all at risk of emotional biases just as investors are (in particular, confirmation bias and attribution bias) since sticking to our previous conclusions will reduce cognitive dissonance as well as avoiding the embarrassment of admitting that we are mistaken. But, it is hard to construct a scenario in which anything other than genuine cognitive error (a mistaken belief that our research conclusions are correct) would convince authors, editors or referees to write and publish such research.

Timing

The other theme running through this thesis is the timing of investments, since (a) the DCA and VA strategies alter the timing of users' investment flows, and (b) the bias in the IRR affects the methods used by academics to assess whether investment flows are in aggregate well or badly timed.

The literature on timing is smaller and more self-contained than that on cognitive error, and it mainly focuses on analysis of data on firms' issuance and buyback of shares and on mutual fund flows. Ritter (1991) shows that the returns on shares of firms which have recently had IPOs (initial public offerings of shares) tend to underperform those of matched firms (1975-1984). Baker and Wurgler (2000) also find that the share of equity as a proportion of total new US equity and debt issues tends to rise ahead of periods of low equity returns. Ikenberry, Lakonishok and Vermaelen (1994) find that firms which conduct open market repurchases of their shares subsequently outperform (1980-1990), with this outperformance particularly large among value stocks. Similarly, Loughran and Ritter (1995) find that firms which issue shares (IPOs or seasoned offerings) substantially underperform size-matched firms over the period 1970 to 1990. This differential persists for portfolios of issuing and non-issuing firms which are assessed against the three Fama/French factors. However, some subsequent studies have questioned the applicability of the method used in these studies (Schultz, 2003, Brav and Gompers, 1997, Gompers and Lerner, 2003).

This is evidence that investors' aggregate net investment flows have been badly timed (only issues and buybacks represent net investment flows – other equity transactions are merely transfers between investors). Firms' timing of issues and buybacks of shares has been correspondingly good, lowering their measured cost of capital. This effect appears to be due to issues being concentrated at the end of equity market booms, just ahead of corresponding busts: firms which issue shares during years with high volumes of issuance significantly underperform those which issue during quieter periods.

Analysis of investment flows into and out of mutual funds, has found evidence that investment inflows tend to come ahead of periods of relatively low returns. These flows tend to be related to previous above-average returns (Ippolito, 1992, Sirri and Tufano, 1998), suggesting that investors attempt to "chase" returns and in doing so end up with disappointing subsequent returns. Gruber (1996) and Zheng (1999) find that institutional funds which experience inflows have significantly higher returns than those that have outflows. By contrast, Frazzini and Lamont (2008) find that above-average inflows tend to be followed by below-average returns, suggesting that these inflows are best thought of as "dumb money". Statman (1995) describes this as resulting from a cognitive error (effectively excessive belief in the autocorrelation of investment returns).

As I discuss in chapter 2, he argues that a key advantage of dollar cost averaging is that by imposing a discipline on the timing of investment flows, the strategy prevents investors from chasing returns and that the benefits of this discipline offset the inefficiency of dollar cost averaging in mean/variance terms.

There is widespread evidence that investor timing has been bad. This thesis does not seek to contest the statistical significance of these effects. However, a growing number of studies use the difference between the geometric mean and dollar-weighted (DW) returns as an estimate of the degree to which the bad timing of aggregate investment flows has reduced investor returns (Dichev, 2007, Friesen and Sapp, 2007, Clare and Motson, 2010, Dichev and Yu, 2011). In Chapter 4 I demonstrate that this method is misleading, since the GM-DW return differential is affected by return-chasing regardless of whether investor flows are well or badly timed compared to future returns. When I remove this effect, the 1.3% estimate that Dichev derives for the effect of bad timing in mainstream US equities falls to almost exactly zero.

This thesis also demonstrates another source of bad timing. DCA and VA determine the timing of the additional investments made by investors who follow these strategies. Specifically, they force them to invest gradually, rather than immediately invest all available savings in an appropriate portfolio (i.e. a lump-sum strategy, which I demonstrate in chapters 2 and 3 would be more efficient). Thus these cashflows can be considered badly timed, since they shift individual investors' strategic asset allocations in ways which make them less efficient. However, this is very different from the conclusion that investors' aggregate new investment flows tend to come ahead of periods of relatively low returns. Such poor timing can in a sense be considered to be a zero sum game: reducing returns to investors, but correspondingly reducing the cost of capital to firms. By contrast, the inefficiency of DCA and VA is that by forcing investors to invest gradually, they impose inefficient portfolio allocations. This brings no corresponding benefit to firms. Indeed, by worsening investors expected risk-adjusted returns, it could be argued to increase the cost of capital demanded from firms.

References

Agarwal, S, Gabaix, X, Laibson, D, Driscoll, J (2009) “The age of reason: Financial decisions over the life cycle and implications for regulation” Brookings Papers on Economic Activity, 51.

Agarwal, S & Mazumder, B (2013). "Cognitive Abilities and Household Financial Decision Making," *Applied Economics* 5(1), pages 193-207.

Ball, R., Brown, P. (1968). “An Empirical Evaluation of Accounting Income Numbers.” *Journal of Accounting Research* 6, 159–177.

Banerjee, A. V. (1992). "A Simple Model of Herd Behavior". *Quarterly Journal of Economics* 107 (3): 797–817.

Banz, R. W. (1981). “The relationship between return and market value of common stocks.” *Journal of Financial Economics*, 9(1), 3-18.

Baker, M and Wurgler, J (2000) "The Equity Share In New Issues And Aggregate Stock Returns." *Journal of Finance*, Vol. 55, No 5.

Barber, B. M. and T. Odean, (2000) “Trading Is Hazardous to Your Wealth: The Common Stock Investment Performance of Individual Investors", 55 (2), 773-806.

Barber and Odean (2001) "Boys Will Be Boys: Gender, Overconfidence, And Common Stock Investment," Quarterly Journal of Economics 116, 261-292.

Barber, B.M. and T. Odean (2006) “All that Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors". SSRN Working Paper.

Barber, B.M., Y. Lee, Y. Liu, and T. Odean (2007) “Just How Much Do Individual Investors Lose By Trading?" SSRN Working Paper

Barberis, N., A. Shleifer, and R. Vishny (1998). “A Model of Investor Sentiment”. *Journal of Financial Economics* 49, 307-343.

Beauchamp JP, D Cesarini & M Johannesson (2011). “The Psychometric Properties of Measures of Economic Risk Preferences.” (Working paper)

Becharia A, Damasio AR, Damasio H, Anderson S. (1994). "Insensitivity to future consequences following damage to human prefrontal cortex". *Cognition* 50: 7–15

Benartzi, Shlomo. (2001) "Excessive Extrapolation and the Allocation of 401(k) Accounts to Company Stock." *Journal of Finance* 56(5): 1747–1764.

Benartzi, S & Thaler, R.H., (1995). "Myopic Loss Aversion and the Equity Premium Puzzle," The *Quarterly Journal of Economics*, vol. 110(1), pages 73-92.

Benjamin, D. J., Brown, S. A., & Shapiro, J. M. (2013). "Who is 'behavioral'? Cognitive ability and anomalous preferences." *Journal of the European Economic Association*, 11(6), 1231-1255.

Bertrand, M, and Morse, A (2011) "Information Disclosure, Cognitive Biases and Payday Borrowing", *Journal of Finance*, 66(6), pp. 1865-1893.

Bikhchandani, S, Hirshleifer, D and Welch, I. (1992). "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades." *Journal of Political Economy* 100 (5): 992–1026.

Brav, A., & Gompers, P. A. (1997). "Myth or Reality? The Long-Run Underperformance of Initial Public Offerings: Evidence from Venture and Nonventure Capital-Backed Companies." *Journal of Finance*, *52*(5), 1791-1821.

Breiter, H. C., Aharon, I., Kahneman, D., Dale, A., & Shizgal, P. (2001). "Functional imaging of neural responses to expectancy and experience of monetary gains and losses." *Neuron*, *30*(2), 619-639.

Brennan, T. and A. Lo, 2012, "An Evolutionary Model of Bounded Rationality and Intelligence", PLOS ONE 7: e34569. doi:10.1371/journal.pone.0050310.

Bucher-Koenen, T., & Ziegelmeyer, M. (2011). "Who lost the most? Financial literacy, cognitive abilities, and the financial crisis" ECB working paper series no 1299

Burks, S, Carpenter JP, Goette L, and Rustichini, A (2009) "Cognitive skills affect economic preferences, strategic behavior, and job attachment" PNAS 2009 106 (19) 7745-7750.

Campbell, J. Y. (2006) "Household Finance", *Journal of Finance*, 61 (4), 1553-1604.

Christelis, D., Jappelli, T., & Padula, M. (2010). “Cognitive abilities and portfolio choice.” *European Economic Review*, *54*(1), 18-38.

Clare, A., and N. Motson. (2010) “Do UK Investors Buy at the Top and Sell at the Bottom?” Working Paper, Cass Business School, City University London.

Coates, J. and Herbert, J. (2008), “Endogenous Steroids And Financial Risk Taking On A London Trading Floor”, Proceedings of the National Academy of Sciences 105, 6167–6172.

Cohen, J. D. (2005). “The Vulcanization of The Human Brain: A Neural Perspective on Interactions Between Cognition and Emotion.” *Journal of Economic Perspectives*, 19(4), 3-24.

Cole, S., & Shastry, G. (2009). “Smart Money: The Effect of Education, Cognitive Ability, and Financial Literacy on Financial Market Participation.” Harvard Business School Finance Working Paper, (09-071).

DeBondt, W. F., & Thaler, R. (1985). “Does the stock market overreact?” *Journal of Finance*, *40*(3), 793-805.

DeLong, B. J.; Shleifer, A., Summers, L., Waldmann, R. J. (1990). "Noise Trader Risk in Financial Markets". *Journal of Political Economy* 98 (4): 703–738.

DeMiguel, V, L. Garlappi and R. Uppal (2009) “Optimal versus Naive Diversification: How Inefficient Is the 1/N Portfolio Strategy?” *Review of Financial Studies* 22(5), 1915--1953.

Dhar and Zhu (2006) “Up Close and Personal: an Individual Level Analysis of the Disposition Effect", *Management Science*.

Dichev, I. D. (2007) “What Are Stock Investors' Actual Historical Returns? Evidence from Dollar-Weighted Returns.” *American Economic Review*, 97, 386-401.

Dichev, I. D., and G. Yu. (2011) “Higher Risk, Lower Returns: What Hedge Fund Investors Really Earn.” *Journal of Financial Economics*, 100, 248-263.

Disney, R., & Gathergood, J. (2013). “Financial Literacy and Consumer Credit Portfolios.” *Journal of Banking & Finance*, *37*(7), 2246-2254.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., & Wagner, G. G. (2011). “Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences.” *Journal of the European Economic Association*, *9*(3), 522-550.

Einhorn, H. J., & Hogarth, R. M. (1986). “Decision Making Under Ambiguity.” Journal of Business, 59, 225-250.

Ellsberg, D. (1961) “Risk, Ambiguity, and the Savage Axioms.” *Quarterly Journal of Economics* 75: 643–669.

Fama, E. F., & French, K. R. (1992). “The Cross-Section of Expected Stock Returns.” *Journal of Finance*, 47(2), 427-465.

Fernholz, R and Shay, B. (1982) “Stochastic Portfolio Theory and Stock Market Equilibrium.” *Journal of Finance*, 37(2), 615–624.

Festinger, L. (1957). “A Theory of Cognitive Dissonance.” Stanford University Press.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). “Knowing With Certainty: The Appropriateness of Extreme Confidence”. *Journal of Experimental Psychology: Human perception and performance*, *3*(4), 552.

Frazzini, A., & Lamont, O. A. (2008). “Dumb Money: Mutual Fund Flows and the Cross Section of Stock Returns.” *Journal of Financial Economics*, *88*(2), 299-322.

Frederick, S. (2005). “Cognitive Reflection and Decision Making.” *The Journal of Economic Perspectives*, *19*(4), 25-42.

French, K. R. And J. M. Poterba (1991) "Investor Diversification and International Equity Markets.” *American Economic Review*, v81(2), 222-226.

Friesen, G. C., and T. R. A. Sapp. (2007) “Mutual Fund Flows and Investor Returns: An Empirical Examination of Fund Investor Timing Ability.” *Journal of Banking and Finance*, 31, 2796-2816.

Gervais, S., & Odean, T. (2001). “Learning To Be Overconfident.” *Review of Financial studies*, *14*(1), 1-27.

Goel, V., Buchel, C., Frith, C., & Dolan, R. J. (2000). “Dissociation of Mechanisms Underlying Syllogistic Reasoning”. *NeuroImage*, 12(5), 504-514.

Goetzmann, W. N. and A. Kumar, (2005) “Why do individual investors hold underdiversified portfolios", April 2005. University of Texas at Austin and Yale International Centre of finance Working Paper

Gompers, P. A., & Lerner, J. (2003). “The Really Long-Run Performance of Initial Public Offerings: The Pre-NASDAQ Evidence.” *The Journal of Finance*, *58*(4), 1355-1392.

Grinblatt, M., Keloharju, M., & Linnainmaa, J. T. (2012). “IQ, Trading Behavior, and Performance.” *Journal of Financial Economics*, *104*(2), 339-362.

Gruber, M. (1996) “Another Puzzle: The Growth in Actively Managed Mutual Funds.” *Journal of Finance* 51, 783–810.

Haigh, M.S. and John A. List (2005). “Do Professional Traders Exhibit Myopic Loss Aversion? An Experimental Analysis”. *Journal of Finance*, 60(1).

Hirshleifer, D., & Shumway, T. (2003). “Good day sunshine: Stock returns and the weather.” *Journal of Finance*, 58(3), 1009-1032.

Houge, T and Loughran, T (2000) “Cash Flow is King? Cognitive Errors by Investors”, *Journal of Psychology and Financial Markets,* vol. 1, 2000, 161-175.

Ikenberry, D. L., Lakonishok, J.and Vermaelen, T., (1994) “Market Underreaction to Open Market Share Repurchases.” NBER Working Paper No. w4965.

Ippolito, Richard A., (1992). “Consumer Reaction to measures of Poor Quality: Evidence From The Mutual Fund Industry”. *Journal of Law and Economics* 35, 45-70.

Jegadeesh, N. and Titman, S. (1993), “Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency”. *Journal of Finance*, 48: 65–91

Kahneman, D., (2011).”Thinking, Fast and Slow”. Macmillan.

Kahneman, D., & Tversky, A. (1972). “Subjective Probability: A Judgment of Representativeness.” *Cognitive psychology*, *3*(3), 430-454.

Kahneman, D., & Tversky, A. (1979). “Prospect Theory: An analysis of Decision Under Risk.” *Econometrica: Journal of the Econometric Society*, 263-291.

Kahneman, D., & Tversky, A. (1982). “Variants of Uncertainty.” *Cognition*, *11*(2), 143-157.

Kamstra, M., Kramer, L., and Levi, M., (2003), “Winter Blues: Seasonal Affective Disorder (SAD) and Stock Market Returns”, *American Economic Review* 93, 324–343.

Korniotis, G and Kumar, A (2007) “Does Investment Skill Decline Due To Cognitive Aging or Improve With Experience?" Working paper.

Korniotis, G. M., & Kumar, A. (2011). “Do Older Investors Make Better Investment Decisions?” *The Review of Economics and Statistics*, *93*(1), 244-265.

Kuhnen, C. M., & Knutson, B. (2005). “The Neural Basis of Financial Risk Taking.” Neuron, 47(5), 763-770.

Laibson, D (1997). "Golden Eggs and Hyperbolic Discounting". *Quarterly Journal of Economics* 112 (2): 443–477.

Larson, J.R. (1977). “Evidence for a Self-Serving Bias in The Attribution of Causality.” *Journal of Personalit*y 45(3), 430-441.

Lo, A. (2012) “Adaptive Markets and the New World Order, *Financial Analysts Journal* 68, 18–29.

Lo, A. and Repin, D. (2002) “The Psychophysiology of Real-Time Financial Risk Processing”, *Journal of Cognitive Neuroscience* 14, 323–339.

Lo, A., Repin, D. and Steenbarger, B. (2005) “Fear and Greed in Financial Markets: An Online Clinical Study”, *American Economic Review* 95, 352–359.

Loomes, Graham and R. Sugden, (1982) “Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty”. *The Economic Journal*, Vol. 92, No. 368.

Loughran, T. & Ritter, J.R. (1995) "The New Issues Puzzle." *Journal of Finance*, Vol 50, no. 1, March 1995.

Luenberger, David G. (1997, 2nd ed 2013) “Investment Science”, Oxford University Press.

Lusardi, Annamaria (2011) “Americans’ Financial Capability.” NBER Working Paper 17103.

Lusardi, A and O. S. Mitchell (2014) “The Economic Importance of Financial Literacy: Theory and Evidence”. *Journal of Economic Literature*, 52(1), pages 5-44

McLean, R. David and Pontiff, Jeffrey. (2013) “Does Academic Research Destroy Stock Return Predictability?”. Available at SSRN: http://ssrn.com/abstract=2156623.

NHS (2013) “Improving Access to Psychological Therapies”, www.iapt.nhs.uk

NHS Direct (2014) “Cognitive behavioural therapy (CBT)” http://www.nhs.uk/Conditions/Cognitive-behavioural-therapy/Pages/Introduction.aspx

Odean, T., “Are Investors Reluctant to Realize Their Losses?" *Journal of Finance*, 1998, 53 (5), 1775-1798.

Odean, T. (1999) “Do Investors Trade Too Much?" *American Economic Review*, 89 (5), 1279-1298.

Polkovnichenko, V. (2005) “Household Portfolio Diversification: A Case for Rank-Dependent Preferences", *Review of Financial Studies*, 18 (4), 1467.

Ritter, J. R. (1991). “The Long-Run Performance of Initial Public Offerings.” *Journal of Finance*, *46*(1), 3-27.

Schultz, P. (2003). “Pseudo Market Timing and The Long-Run Underperformance of IPOs. “*Journal of Finance*, *58*(2), 483-518.

Shamosh, NA and Gray, JR (2008) “Delay Discounting And Intelligence: A Meta-analysis.” *Intelligence*, 36(4): 289-305.

Shefrin, H. and M. Statman, (1985) “The Disposition to Sell Winners Too Early and Ride Losers Too Long: Theory and Evidence", *Journal of Finance*, 40 (3), 777-790.

Simon, H. A. (1955). “A Behavioral Model of Rational Choice.” *Quarterly Journal of Economics*, 69(1), 99-118.

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

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology* 127(2).

Thaler, R. (1985) “Mental Accounting and Consumer Choice", *Marketing Science*, 4 (3), 199-214.

Tversky, A; Kahneman (1973). "Availability: A Heuristic For Judging Frequency and Probability". *Cognitive Psychology* 5: 207–233.

Tversky, A. & Kahneman, D. (1974). "Judgment Under Uncertainty: Heuristics and Biases". *Science*, 185, 1124–1130.

Tversky, A., Kahneman, D. (1983). “Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgments." *Psychological Review*, 90, 293-315.

Tversky and Kahneman (1981) “The Framing of Decisions and The Psychology of Choice.” *Science*, 211 (4481): 453-58.

van Rooij, M. C., Lusardi, A., & Alessie, R. J. (2011). “Financial literacy and retirement planning in the Netherlands.” *Journal of Economic Psychology*, *32*(4), 593-608.

van Rooij, M., Lusardi, A., & Alessie, R. (2011). “Financial Literacy and Stock Market Participation.” *Journal of Financial Economics*, *101*(2), 449-472.

Willenbrock, Scott (2011) “Diversification Return, Portfolio Rebalancing, and the Commodity Return Puzzle”. *Financial Analysts Journal*, Vol. 67, No. 4: 42–49

Zheng, L. (1999) “Is Money Smart? A Study of mutual Fund Investors’ Fund Selection Ability.” *Journal of Finance* 54, 901–933.

Chapter 2

Dollar Cost Averaging: The Role of Cognitive Error

Abstract

Dollar Cost Averaging (DCA) has long been shown to be mean-variance inefficient, yet it remains a very popular strategy. Recent research has attempted to explain its popularity by assuming more complex risk preferences. This paper rejects such explanations by demonstrating that DCA is sub-optimal regardless of preferences over terminal wealth. DCA's continued popularity is better explained by behavioural finance effects. Specifically, this paper identifies the cognitive error in the argument that is normally put forward in favor of the strategy. This gives us a simpler and more robust explanation for DCA's continued popularity, and one which brings very different welfare implications.

JEL Classification: G11

I am grateful for comments from George Constantinides, Christopher Gilbert, Stewart Hodges, Robert Hull, Aneel Keswani, Karyl Leggio, Burton Malkiel, Ian Marsh, Frank McGroarty, Moshe Milevsky, Nick Motson, Hugh Patience, Giorgio Questa, Marc Rieger, Astrid Salzmann, Meir Statman, Mark Tapley, Steven Thorley, seminar participants at Cass Business School and participants at the June 2010 Behavioural Finance Working Group in London and the 2012 Infiniti conference in Dublin, and anonymous referees. The usual disclaimer applies.

Dollar Cost Averaging: The Role of Cognitive Error

1. Introduction

Dollar cost averaging (DCA) is the strategy of buying assets gradually over time in equal dollar amounts, rather than buying the desired total immediately in one lump sum. The strategy may be used for investing or for procuring materials whose price is volatile and unpredictable. DCA is still widely recommended even though previous research has demonstrated that it is a mean-variance inefficient strategy.

More recent research has focused on explaining why DCA nevertheless remains so popular. Statman (1995) argues that the answer lies in behavioural finance, but subsequent research has tried instead to argue that DCA can be an optimal strategy for the entirely rational agents of standard finance theory. These rationalist explanations are reviewed briefly below, but they are unsatisfactory for three key reasons. First, explanations based on non-variance forms of risk preference must be rejected because DCA is a sub-optimal strategy regardless of preferences over terminal wealth (I demonstrate this result in a later section). Second, other explanations rely on additional (and unverifiable) assumptions about the individuals or the markets involved, even though proponents of DCA generally recommend the strategy to everybody, regardless of their objectives, expectations or preferences. Finally, most of the theories which have been advanced to explain DCA's popularity make no reference to the factor which is usually central to the case made by its proponents: that DCA automatically purchases more securities when the price is lower, and so achieves an average purchase cost which is below the average market price. The rare exceptions which argue that this is a key attraction of the strategy (Thorley 1994; Greenhut 2006) fail to correctly identify the cognitive error that is implicit in it.

The empirical evidence clearly shows that DCA's lower average costs do not lead to higher expected returns, but it has never been made clear why this is so. Previous papers have argued that the fact that DCA buys at an average purchase cost which is below the average price is irrelevant because it is generally not possible to subsequently sell at this average price (e.g. Thorley 1994; Milevsky and Posner 2003), but this misrepresents the case put forward for DCA. If it was possible to sell at the average price then DCA would generate guaranteed short-term profits. This is not what its proponents are claiming. Subsequent prices are uncertain so DCA remains risky, but no matter how prices evolve it appears obvious that buying at a lower average cost must result in higher profits than buying at a higher average cost would have, so DCA seems beneficial. By contrast, this paper identifies the hidden flaw in this argument, and shows that comparing DCA's average cost with the average market price is systematically misleading. This provides the basis for a simpler explanation of DCA's continuing popularity: that those who use the strategy are making a cognitive error in failing to recognize the flaw in the key argument presented by its proponents.

The contribution of this paper is: (i) demonstrating that most of the recent literature on DCA has been chasing a false lead, since alternative risk preferences cannot explain DCA's continued popularity; (ii) identifying a simpler and more robust explanation by analyzing the key argument made by DCA's proponents. This gives us a better understanding of the reasons for DCA's continued popularity, and of the resulting welfare effects.

2. Previous Explanations for DCA's Popularity

Previous research clearly demonstrates that DCA is mean-variance inefficient. Constantinides (1979) shows that as DCA commits the investor to continue making equal periodic investments, it must be dominated by more flexible strategies which allow the investor to make use of any additional information which is available in later periods. A large number of empirical studies also find that investing the whole desired amount in one lump sum generally gives better mean-variance performance than DCA. These include Knight and Mandell (1992/93), Williams and Bacon (1993), Rozeff (1994) and Thorley (1994).

Proponents sometimes claim that DCA improves diversification by making many small purchases but, as Rozeff (1994) notes, by investing gradually DCA leaves overall profits most sensitive to returns in later periods, when the investor is nearly fully invested. Earlier returns have less impact because the investor then holds mainly cash. Better diversification is achieved by investing immediately in one lump sum, and thus being fully exposed to the returns in each period. Milevsky and Posner (2003) extend the analysis into continuous time, and show that it is always possible to construct a constant proportions continuously rebalanced portfolio which will stochastically dominate DCA in a mean-variance framework. They also show that for typical levels of volatility and drift there is a static buy and hold strategy which dominates DCA.

As the evidence became overwhelming that DCA is mean-variance inefficient, research turned to attempts to explain why it nevertheless remains very popular. These fall into three categories, based on: (i) non-variance investor risk preferences; (ii) behavioural finance effects; (iii) investors' forecasting of asset returns. We shall consider each in turn.

First, DCA's mean-variance inefficiency led some to investigate whether DCA outperforms on non-variance measures of risk. Leggio and Lien (2003) consider the Sortino ratio and upside potential ratio. Their results vary between asset classes, but overall they reject claims that DCA is superior. DCA substantially reduces shortfall risk (the risk of falling below a target level of terminal wealth) compared to a lump sum investment (Dubil 2005; Trainor 2005), but even if investors consider this to be worth the associated reduction in expected return, Constantinides' critique remains potent: less rigid strategies should be expected to dominate, for example by allowing investors to increase their exposures if their portfolios are safely above the required minimum value. I demonstrate below a much more general result: that DCA is sub-optimal regardless of investors' risk preferences, since an alternative strategy can always be constructed which generates exactly the same distribution of terminal wealth as DCA but requires less capital. This must be considered preferable under any plausible set of preferences (provided only that more terminal wealth is preferred to less). Thus hypothesizing alternative investor risk preferences is a sterile area of research which cannot explain DCA's continued popularity.

Statman (1995) argues instead that DCA's popularity is explained by various behavioural finance effects. One of these is prospect theory, but Leggio and Lien (2001) and Fruhwirth and Mikula (2008) have subsequently shown that DCA remains an inferior strategy even when investor preferences are consistent with prospect theory. This is confirmed by the more general suboptimality result in section 5 below. However, other explanations within behavioural finance remain attractive. Statman argues that DCA frames investment decisions in a flattering context. Furthermore, by committing investors to continue investing at a constant rate, DCA limits choice in the short term, which may (i) reduce regret; (ii) reduce the impact of investor myopia (which might otherwise lead to long-term underinvestment); and (iii) protect investors from their tendency to time their investments on the basis of naïve extrapolation of recent price trends. I consider these points further in section 6.

Other papers have sought to justify the use of DCA by making alternative assumptions about investors' forecasting of market returns. Milevsky and Posner (2003) show that if an investor has a firm forecast of the value of a security at the end of the horizon, then as long as volatility is sufficiently high the expected return from DCA conditional on this forecast will exceed the corresponding expected return from investing in this security in one lump sum. This explanation assumes that this expected terminal value remains fixed throughout the horizon, and does not change as market prices shift. Thus, for example, a fall in market prices increases the expected future return and so makes DCA's purchase of additional shares at this lower price very attractive. If instead investor expectations tend to shift in line with market prices (either in response to the same underlying news that shifted market prices, or because investor sentiment is directly affected by market price movements) then this property is removed, and DCA becomes unattractive.

Brennan, Li and Torous (2005) investigate whether DCA's use can be explained by weak-form inefficiency in equity returns. They find that the degree of mean reversion in US equity prices (1926-2003) was too small to offset the underlying inefficiency of DCA as a strategy for building up a new portfolio, but that it was large enough to make DCA a beneficial strategy for adding a new stock to an already well-diversified portfolio. However, this is a new result which required detailed econometric study. For this to explain DCA's popularity the authors are forced to assume that this property was already known to investors as part of inherited "folk finance" wisdom.

In sum, recent research has developed progressively more complex theories to try to explain how DCA could remain popular for the rational investors of standard finance theory. The results have been unsatisfactory. This paper shows that DCA's popularity cannot be explained by non-variance investor risk preferences. Other complex explanations depend on unverifiable assumptions such as “folk finance” or constant investor price expectations. Occam’s razor tells us that theories which do not require such assumptions should be preferred. Such assumptions must also be reconciled with the fact that DCA is generally recommended to investors without any detailed consideration of their goals, expectations or risk preferences, or the properties of the market involved. Instead proponents almost invariably stress the fact that DCA always buys at below the average price, suggesting that it increases the expected return for all investors. Explaining why a lower average purchase price does not actually increase expected returns is thus central to understanding DCA’s popularity. The following section derives such an explanation, and in the process provides a simpler explanation for DCA’s popularity: that investors are making a cognitive error in failing to identify the flaw in the key argument which is put forward by its proponents.

3. The Intuition Behind the Cognitive Error

Table 2.1 shows a numerical example typical of those used by proponents of DCA (the alternative ESA strategies are not normally made explicit and will be explained later). A fixed \$60 each period is invested in a specific equity. The price is initially \$3, allowing 20 shares to be purchased. The sharp fall to \$1 allows 60 shares to be purchased for the same dollar outlay in period two, whilst the rebound to \$2 allows 30 units to be bought in the final period. The argument usually made in favor of DCA is that it buys shares at an average cost (\$180/110 = \$1.64) which is lower than the average market price of the shares over the period during which they were accumulated (\$2). This is achieved because DCA automatically buys more shares during periods when they are relatively cheap and fewer when they are more expensive.

Table 2.1: Illustrative Comparison of Strategies as Share Prices Fall

The DCA strategy invests a fixed $60 per period, and is compared to strategies which buy Equal Share Amounts (ESA) of (b) 20 shares, (c) 30 shares per period. Falling prices mean that ESA1 invests a lower dollar total than DCA. ESA2 is the only ESA strategy which invests the same amount as DCA, but choosing the right number of shares in period one requires knowledge of future share prices.

		(a) DCA		(b) ESA1		(c) ESA2	
Period	Share price	Shares purchased	Investment	Shares purchased	Investment	Shares purchased	Investment
1	$3	20	$60	20	$60	30	$90
2	$1	60	$60	20	$20	30	$30
3	$2	30	$60	20	$40	30	$60
Total		110	$180	60	$120	90	$180

Greenhut (2006) takes issue with the particular return assumptions which are often used in such “demonstrations” of the superiority of DCA. However, there is a much more general issue here. The average purchase cost for DCA investors gives greater weight to periods when the price is relatively low, so price fluctuations will always mean that DCA investors buy at less than the average price, regardless of the particular path taken by prices. The difference is particularly large in the example above due to the large price movements, but any price volatility favors DCA. Only when the share price remains unchanged in all periods will the average cost equal the average price. Rather than challenging the particular numbers used, we need to examine why a strategy which buys assets at a lower average cost does not in fact lead to higher expected profits.

Previous studies have found that DCA is mean-variance inefficient compared to investing the whole desired amount immediately in one lump sum, but proponents of DCA are making a different comparison. In noting that the average cost achieved by DCA is less than the average price they are implicitly comparing DCA with a strategy which invests the same amount by buying a constant number of shares each period (thus achieving an average purchase cost equal to the unweighted average price). This is the comparison that we must make here in order to understand why the case in favor of DCA is misleading.

Table 2.1 compares the cashflows under DCA with two alternative strategies which buy a constant number of shares in each period (equal share amounts: ESA1 and ESA2). The difference between these two alternatives may appear to be a trivial matter of scale, but it is in this difference that the false comparison lies.

ESA1 is an attempt to invest the same total amount as DCA over these three periods, but to do so in equal share amounts. With the share price initially at $3, a reasonable approach would be to buy 20 shares, since if prices remain at this level in periods two and three we will end up investing exactly the $180 total that we desire. But our strategy then requires that we buy 20 shares in each of the following periods, and when prices in periods two and three turn out to be substantially lower, we end up investing only $120. It is only with perfect foreknowledge of future share prices that we could have known that the only way of investing $180 in equal share amounts is to buy 30 shares each period, as shown in ESA2.

When proponents of DCA note that it buys shares at an average cost which is below the unweighted average price during this period they are effectively comparing the DCA strategy with a strategy which invests the same dollar total in equal share amounts (i.e. the ESA2 strategy). But ESA2 can only achieve this if we know future share prices – otherwise we will generally end up investing the wrong amount. Furthermore, this foresight is used in a way which systematically reduces profitability. In this example, the ESA2 strategy reacts to the knowledge that prices are about to fall by investing more than it otherwise would in period one. Conversely, it would invest less in period one if prices in subsequent periods were going to be higher. This is the only way to invest the correct amount but, of course, it systematically reduces profits.

Table 2.2 shows the same strategies, but with the share price rising rather than falling. The DCA strategy again invests $60 each period, but as prices rise fewer shares are purchased in the later periods. Once again DCA achieves an average cost ($180/47=$3.83) below the average price ($4) by buying more shares when they are relatively cheap. This effectively compares the DCA strategy with the ESA2 strategy, which invests the same total amount, but buys only 45 shares compared to 47 using DCA.

Table 2.2: Illustrative Comparison of Strategies as Share Prices Rise

DCA invests a fixed $60 per period, compared to buying Equal Share Amounts (ESA) of (b) 20 shares and (c) 15 shares per period. Rising prices mean that ESA1 invests a larger dollar total than DCA. ESA2 is the only ESA strategy which invests the same amount as DCA, but choosing the right number of shares in period one requires knowledge of future share prices.

		(a) DCA		(b) ESA1		(c) ESA2	
Period	Share price	Shares purchased	Investment	Shares purchased	Investment	Shares purchased	Investment
1	$3	20	$60	20	$60	15	$45
2	$4	15	$60	20	$80	15	$60
3	$5	12	$60	20	$100	15	$75
Total		47	$180	60	$240	45	$180

However, as we saw earlier, the real alternative to DCA is ESA1. In practice our best guess would again be to invest one third of our total budget in the first period, since if prices were to stay at this level we would invest the correct amount. But when prices subsequently rise we end up spending substantially more than this ($240). ESA2 invests the correct amount, but it achieves this only by knowing that prices are about to rise and responding to this knowledge by buying fewer shares than ESA1. Again, profits are reduced.

Comparing DCA's average cost with the average price effectively compares DCA with a strategy which uses perfect foresight in a way which systematically reduces profits and increases losses. DCA's proponents almost invariably[1] refer to its lower unit costs, so this cognitive error appears to be a key factor explaining the strategy's continued popularity.

4. The Arithmetic of the Cognitive Error

This section demonstrates more formally that it is only by making a misleading comparison that DCA appears to offer superior profits. We consider investing in an asset over a series of *n* discrete periods. The price of the asset in each period *i* is p_i. The alternative investment strategies differ in the quantity of shares q_i that are purchased in each period. We evaluate profits at a subsequent point, after all investments have been made. If prices are then p_T, the profit made by the strategy is:

$$\Pi = p_T \sum_{i=1}^{n} q_i - \sum_{i=1}^{n} p_i q_i \tag{1}$$

We define DCA as a strategy which invests *b* dollars in each period ($p_i q_i = b$). This gives us the profits that will result from following a DCA strategy:

$$\Pi_{dca} = p_T \sum_{i=1}^{n} \left(\frac{b}{p_i} \right) - nb \tag{2}$$

We assume that investors who use DCA do not believe that they can forecast market prices. In effect they assume that prices follow a random walk. As Brennan et al. (2005) shows, mean reversion could under some limited circumstances lead DCA to outperform, but the case that is normally made for DCA makes no claim that it is

[1] As an indication of this, of 25 non-academic references accessed using an internet search on "dollar cost averaging" 21 were in favour of the strategy and four were against. Some of these noted that DCA reduces risk (although, as section 5 shows, it is an inefficient means of doing so), but every one of the 21 referred either directly to DCA's reduced unit costs or to the benefits of buying fewer shares when the price is high and more when it is low.

exploiting market inefficiency − instead it is portrayed as a strategy which will outperform in any market. Furthermore, DCA commits investors to invest the same amount no matter what price movements they expect in the coming period. Those who (rightly or wrongly) believe that they can forecast short-term price movements are likely to reject DCA and follow other strategies instead.

We also assume that this random walk has zero drift.[2] This assumption is generous to DCA, since upward drift will tend to penalize the strategy for investing gradually. Investors presumably believe that over the medium term their chosen securities will generate an attractive return, but they must also believe that the return over the short term (while they are using DCA to build up their position) is likely to be small. Investors who expect significant returns over the short term would prefer to invest immediately in one lump sum rather than delay their investments by following a DCA strategy. Given these assumptions, investors will assume ex ante that prices will remain flat, with $E[p_T/p_i]=1$ for all i. Substituting this into Equation 2, we see that the ex ante expected profit from the DCA strategy is zero.

Our alternative investment strategy is to buy equal numbers of shares in each period ($q_i=a$). Substituting this into Equation 1 gives us:

[2] The assumption of zero drift need not imply a loss of generality, since drift could be incorporated by defining prices not as absolute market prices, but as prices relative to a numeraire which appreciates at a rate which gives a fair return for the risks inherent in this asset ($p_i*=p_i/(1+r)^i$, where r reflects the cost of capital and an appropriate risk premium). We could then assume that p_i* has zero expected drift since investors who use DCA will not believe that they can forecast short-term relative returns for assets of equal risk (those who do would choose other strategies). The results derived here would continue to hold for p_i*, with profits then defined as excess returns compared to the risk-adjusted cost of capital. This assumes that funds not yet needed can be held in assets with the same expected return as the risky asset, which is clearly generous to DCA. If instead cash is held on deposit at a lower expected return, then DCA's expected return will clearly be reduced by delaying investment.

$$\Pi_{esa1} = anp_T - a\sum_{i=1}^{n} p_i \tag{3}$$

The ex ante expected profit from this ESA strategy is also zero (this can be seen by substituting E[p_i]=E[p_T] for all i, as an equivalent expression of our driftless random walk). Thus DCA does not give superior expected returns.

This is an intuitive result. We can regard the total return as a weighted average of the returns made on the amounts invested in each period. ESA and DCA differ only in giving different relative weights to these individual period returns. But if prices are believed to follow a random walk with zero drift the expected return will be zero for each period and varying the relative wight given to different periods' returns cannot change the expected aggregate return.[3] By contrast, DCA's popular supporters suggest that even when investors have no belief that they can forecast market returns they can nevertheless expect to beat the market by using DCA.

As we saw in the previous section, the total amount invested under ESA1 ($a\sum p_i$) is likely to differ from the amount (nb) invested under DCA. But the comparison that is usually presented by proponents of DCA assumes that the two techniques invest equal total amounts. Thus to duplicate the conventional "proof" of the benefits of DCA, we need to rescale the number of shares bought under ESA1 by the fixed factor ($nb/a\sum p_i$), so

[3] Expected profits can be expressed as $\sum_{i=1}^{n}\left(\mathrm{E}[p_T q_i]\right)-\sum_{i=1}^{n}\left(\mathrm{E}[p_i q_i]\right)$. Our assumption of a random walk implies that future price movements (p_T/p_i) are always independent of past values of p_i and q_i, so this can be re-written as $\sum_{i=1}^{n}\left(\mathrm{E}\left[\frac{p_T}{p_i}\right]\mathrm{E}[p_i q_i]\right)-\sum_{i=1}^{n}\left(\mathrm{E}[p_i q_i]\right)$. But the random walk has zero drift, so E[p_T/p_i]=1 for all i and expected profits are zero regardless of the amount $p_i q_i$ which is invested in each period.

that an exactly equal amount is invested by the two strategies. This gives us the expected profits resulting from strategy ESA2:

$$\Pi_{esa2} = \Pi_{esa1}\left(\frac{nb}{a\sum_{i=1}^{n} p_i}\right) \tag{4}$$

The use of foresight can be seen in the fact that the scaling factor depends on the average share price throughout the investment horizon. Only if this is known at the outset would we be able to buy the correct number of shares so that we end up spending exactly the same amounts under ESA2 and DCA. Substituting from Equation 3:

$$\Pi_{esa2} = \left(anp_T - a\sum_{i=1}^{n} p_i\right)\left(\frac{nb}{a\sum_{i=1}^{n} p_i}\right) \tag{5}$$

$$= \frac{bn^2 p_T}{\sum_{i=1}^{n} p_i} - nb \tag{6}$$

Subtracting Equation (6) from Equation (2) we find:

$$\Pi_{dca} - \Pi_{esa2} = nbp_T\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{p_i}\right) - \frac{n}{\sum_{i=1}^{n} p_i}\right) \tag{7}$$

The term in brackets is non-negative for positive p_i, and strictly positive if they are not all equal. This follows directly from the arithmetic mean-harmonic mean inequality.[4]

[4] The arithmetic-harmonic mean inequality is usually stated as: $(x_1 + ... + x_n)/n \geq n/(1/x_1 + ... + 1/x_n)$ for positive x_i, so we have substituted $x_i = 1/p_i$ This inequality follows directly from Jensen's inequality that $E[f(x)] \geq f(E[x])$ for any convex function $f(.)$, using the function $f(x) = 1/x$.

This achieves our objective. The analysis above shows that the expected profits from a DCA strategy are identical to those from our ESA1 strategy (both give zero expected profits). By contrast, DCA gives higher expected profits than our ESA2 strategy which scales the level of investment so as to spend exactly the same total amount as DCA. However, ESA2 is not a feasible strategy, since it uses perfect foresight to invest in a systematically loss-making fashion. It is only on this biased comparison (DCA vs ESA2) that DCA appears to make greater returns, yet it is exactly this comparison which is implicitly being made when it is noted that DCA buys at an average cost which is lower than the average price.

This biased comparison suggests that DCA makes profits of (P_T – average purchase cost) per share whereas other strategies make (P_T – average price). Thus DCA appears to shift the whole distribution of possible profits upwards by the extent of the difference between these averages. For most investment strategies reducing the average purchase cost really will increase expected returns, so cost minimization is normally a useful heuristic goal for investors. However, DCA reduces its average cost by increasing its purchases of shares after prices have risen (and vice versa). This is a retrospective response to previous price movements, and will boost expected profits only if asset prices systematically tend to mean-revert. Comparing DCA's average cost with the average market price is systematically misleading since it implicitly compares DCA with a strategy where the amount invested depends on foreknowledge of future prices. The error involved in this comparison has not previously been identified[5]. Faced with such

[5] Thorley (1994) rightly argues that DCA is based on a fallacy, but does not correctly identify the nature of the fallacy. He argues that the comparison of average purchase cost with the average price "would be relevant only if the investor could sell shares at the average historical price" but, as noted above, if investors could sell at this average price then DCA would guarantee immediate profits, which is not what its proponents claim. By contrast, buying any given number of shares

apparently obvious benefits, it should perhaps not be surprising that DCA remains so popular.

5. The Sub-Optimality of DCA

A number of previous studies have attempted to explain DCA's popularity by hypothesizing that although inefficient in mean-variance terms, DCA could still be attractive to rational investors whose risk preferences take alternative (non-variance) forms. This section shows that this is an unproductive line of research, since DCA is a sub-optimal strategy regardless of the investor's risk preferences. For this purpose we use Dybvig's Payoff Distribution Pricing Model (Dybvig, 1988a and 1988b).

As a very simple illustration, Figure 2.1 shows a binomial model of a DCA strategy over four periods. The equity element of the portfolio is assumed to double in a good outturn and halve in a bad outturn. At the start of the first period 16 is invested in equities, with 48 in cash[6]. A further 16 of this cash is invested each period. All paths are assumed to be equally likely. The key to this technique is comparing the terminal wealths with their corresponding state price densities (the state price divided by the probability – in this case $16(1/3)^u(2/3)^d$, where *u* and *d* are the number of up and down states in the path concerned[7]). An efficient strategy will generate the highest terminal wealths in the

at a lower average cost would be very relevant, since it would increase expected returns. It is only because of the retrospective adjustment identified above that this comparison is systematically misleading for DCA.

[6] Investors with regular monthly income and expenditure streams may choose to save a fixed dollar amount each month, leaving them following a DCA strategy by default. However, proponents of DCA are not simply arguing that regular saving is desirable – they claim that it is advantageous to invest any available lump sum gradually rather than immediately. Thus we compare DCA with alternative ways of investing a lump sum.

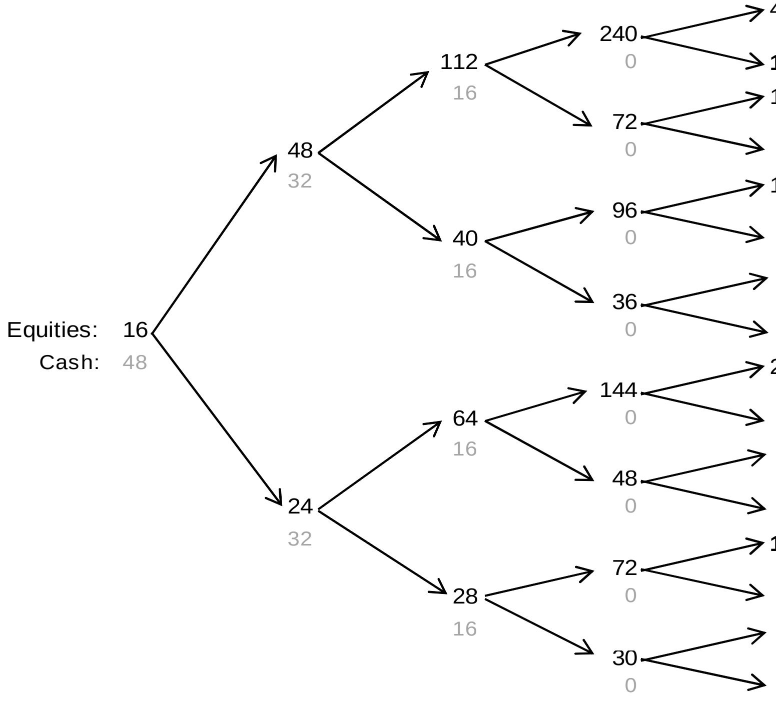
[7] More generally, the state price densities of one period up and down states are $(1/(1+\mathrm{r}\Delta t))\left(1-\left((\mu-\mathrm{r})\Delta t/\sigma\sqrt{\Delta t}\right)\right)$ and $(1/(1+\mathrm{r}\Delta t))\left(1+\left((\mu-\mathrm{r})\Delta t/\sigma\sqrt{\Delta t}\right)\right)$ respectively, where r is the

paths for which these outturns are "cheapest" (i.e. have the lowest state prices). This is generally the case in Figure 2.1, but there are exceptions. DDUU results in a larger terminal wealth than UUUD despite seeing fewer lucky outturns. Similarly, DDDU beats UUDD and UDUD. This can be loosely interpreted as DCA making ineffective use of some comparatively lucky paths.

continuously compounded annual risk-free interest rate and the risky asset has annual expected return μ and standard deviation σ. The corresponding one period risky asset returns are $\left(1+\mu\Delta t+\sigma\sqrt{\Delta t}\right)$ and $\left(1+\mu\Delta t-\sigma\sqrt{\Delta t}\right)$. See Dybvig (1988b).

Figure 2.1: Simple Model of DCA Strategy

This tree shows the value of the investor's equity and cash holdings at the start of each period. Equity values double in a good outturn and halve in a bad outturn (for simplicity cash is assumed to earn no interest and any dividends are assumed immediately reinvested to give the total equity returns shown). Investors start with 16 invested in equities (the upper figure at each node) and 48 in cash (the lower figure). They then invest a further 16 in each subsequent period, leaving zero at the start and end of the final period. All paths are assumed to have equal real world probabilities (the corresponding risk neutral probabilities are 1/3 and 2/3). The sub-optimality of this strategy stems from the fact that in some cases (highlighted) paths with a higher state price density achieve higher terminal wealth than luckier paths with a lower state price density.



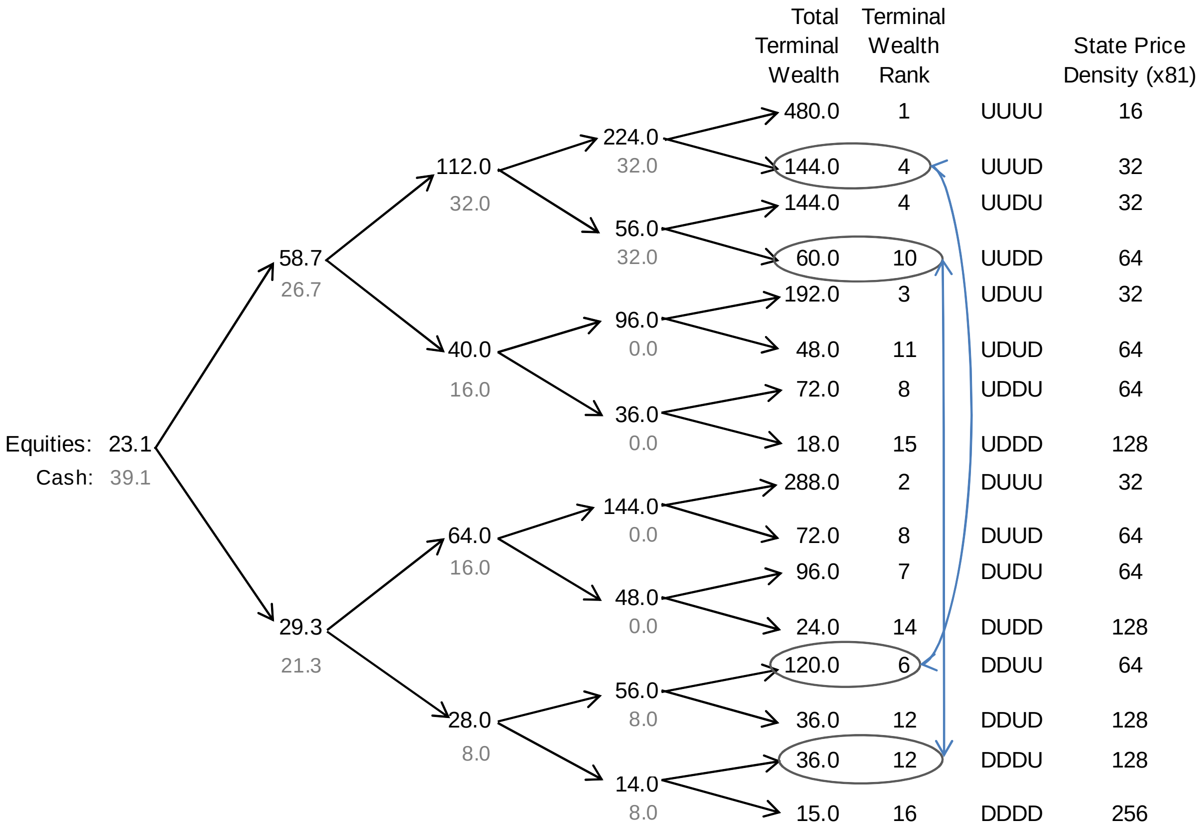
	Terminal Wealth Rank		State Price Density(x81)
480	1	UUUU	16
120	**6**	**UUUD**	**32**
144	4	UUDU	32
36	**12**	**UUDD**	**64**
192	3	UDUU	32
48	**11**	**UDUD**	**64**
72	8	UDDU	64
18	15	UDDD	128
288	2	DUUU	32
72	8	DUUD	64
96	7	DUDU	64
24	14	DUDD	128
144	**4**	**DDUU**	**64**
36	12	DDUD	128
60	**10**	**DDDU**	**128**
15	16	DDDD	256

The inefficiency of DCA can be demonstrated by deriving an alternative strategy which generates exactly the same 16 outturns at lower cost. This is done by changing our strategy to ensure that the best outturns occur in the paths which have the lowest state price densities (i.e. the greatest number of up states), so we switch the outturns for DDUU and UUUD in Figure 1, and those for DDDU and UUDD. The state prices can then be used to determine the value of earlier nodes (thus determining the proportion of the portfolio which must be held in cash at each point in order to duplicate the terminal wealth outturns of a DCA strategy). This in turn determines the initial capital required to generate these outturns. This alternative strategy is shown in Figure 2.2 and requires only 62.2 initial capital (23.1 in equities and 39.1 in cash), compared to 64 above. This improvement is achieved without additional borrowing, merely by making better use of existing capital. This shows the degree to which DCA is inefficient. A key advantage of this method is that it demonstrates DCA's inefficiency without needing to specify the investor's risk preferences, since our alternative strategy generates exactly the same terminal wealth outturns as DCA at a lower initial cost.[8]

[8] Earning an identical set of terminal wealth outturns at a lower initial cost (or, equivalently, greater terminal wealth on all paths at the same initial cost) must be regarded as preferable provided only that what the investor cares about is terminal wealth, and that the investor prefers more terminal wealth to less. These are modest assumptions, although they do rule out effects such as regret which might imply that investor utility also depends on the path by which each terminal wealth outturn was generated. I return to this point in section 6.

Figure 2.2: Optimized Strategy Giving Identical Outturns to DCA

This tree shows the amounts invested in equities and in cash in each period with amount of cash held at the start of each period set to duplicate the outturns in Figure 1, but optimized so that the largest outturns occur in the paths with the lowest state price density (compared with Figure 1, the outturns for UUUD and DDUU have been switched, and the outturns for UUDD and DDDU). Returns on cash and equities are assumed the same as in Figure 1. The lower total capital (62.2) required by this optimized strategy shows the extent to which DCA is an inefficient strategy.



Total Terminal Wealth	Terminal Wealth Rank		State Price Density (x81)
480.0	1	UUUU	16
144.0	4	UUUD	32
144.0	4	UUDU	32
60.0	10	UUDD	64
192.0	3	UDUU	32
48.0	11	UDUD	64
72.0	8	UDDU	64
18.0	15	UDDD	128
288.0	2	DUUU	32
72.0	8	DUUD	64
96.0	7	DUDU	64
24.0	14	DUDD	128
120.0	6	DDUU	64
36.0	12	DDUD	128
36.0	12	DDDU	128
15.0	16	DDDD	256

The amount of cash which must be held at each point is shown below the equity holdings. We can see that the optimized strategy holds considerably less cash than DCA during the first two periods. This supports the interpretation put forward by Rozeff (1994) that DCA is inefficient because it takes too little exposure in the early periods, leaving terminal wealth disproportionately sensitive to returns in later periods.

However, for volatility levels typical of developed equity markets this halving or doubling of equity values at each step of the tree would represent a number of years between each investment. We use this unrealistic assumption merely to allow us to show the dynamic inefficiency in a very simple tree. For more plausible strategies we can consider 12 step trees corresponding to a DCA strategy of equal monthly investments over a one year horizon. Such trees contains 4,096 outturns, and so are not shown here in full, but Table 2.3 shows that a wide range of different assumptions for the market risk premium and volatility all result in efficiency losses. Furthermore, these losses are roughly proportional to the assumed risk premium, which again supports the interpretation that they stem from the returns foregone by holding excessive cash during the early periods. As a robustness check, these calculations were replicated for an 18-step binomial tree, giving 262,144 outturns (the maximum which was computationally practical). The resulting inefficiency estimates were very similar to those in Table 2.3, being larger by a maximum of 0.01%.

Table 2.3: Quantifying the Inefficiency of DCA (% of Initial Capital)

This table uses Dybvig's PDPM model to derive the cost of an optimized strategy which generates the same set of final portfolio values as those achieved by a DCA strategy which invests one twelfth of its initial capital at the start of each month. The table shows the percentage by which the capital required by the DCA strategy is greater than that required by the optimized strategy to generate an identical set of outturns. These figures were derived using a 12 period binomial tree where returns are assumed IID with the binomial steps calibrated to give monthly returns distributed with the annualized risk premia and volatilities shown (almost identical results were found for a 18-step tree with 262,144 outturns). The risk-free rate is assumed to be 5%, but the results are not sensitive to this assumption (adjusting it to 0% or 10% alters these figures by less than 0.005%).

Standard deviation of security (per annum)	Risk premium (per annum)			
	2%	4%	6%	8%
10%	0.12	0.24	0.35	0.46
20%	0.12	0.24	0.36	0.48
30%	0.12	0.24	0.36	0.48

We have assumed that market returns have a binomial distribution, but the fact that the level of volatility in Table 2.3 has very little effect on the size of the inefficiency is a reassuring indication that these results are not sensitive to the particular distribution which is assumed. More importantly, Rieger (2011) demonstrates formally that this inefficiency is not specific to the binomial distribution: he generalizes Dybvig's results, showing that strategies which are path-dependent and generate terminal wealths which have (as in this case) a non-monotonic relationship with market returns are sub-optimal regardless of the distribution of market returns.

Taking the most plausible estimates of the risk premium to be around the middle of the range shown in Table 2.3, the associated efficiency losses are modest, but should nevertheless be regarded as economically significant. For example, sustained return differentials on this scale are likely to be seen as relevant by investors when assessing the performance of competing fund managers. Furthermore, these figures should be regarded as conservative estimates of the actual efficiency losses. In each case the optimized strategy generates the same outturns as DCA but uses less capital. This shows that DCA is inefficient regardless of the form taken by investor risk preferences. This is a powerful result, but there is no reason why in practice investors' preferred option should be to replicate DCA's outturns. Given each investor's specific preferences there are likely to be other strategies which are even more attractive alternatives, so the efficiency losses shown in Table 2.3 must be regarded as lower bounds.

In conclusion: DCA is an inefficient strategy for investing available funds, and this result applies for all plausible forms of investor risk preference. Thus investors' use of DCA cannot be explained as a rational consequence of non-variance risk preferences. The following section considers more plausible explanations for DCA's popularity within behavioural finance.

6. Behavioural Finance Effects

A number of papers have attempted to use standard finance theory to explain why investors might rationally choose DCA despite its mean-variance inefficiency. These explanations have not proved satisfactory. The previous section showed that explanations based on non-variance investor risk preferences must be rejected and, as described above, other explanations require unjustified assumptions about investors' forecasts of asset returns. In this section I instead consider explanations within behavioural finance. Specifically, Statman (1995) sets out four behavioural finance effects which could help explain DCA's popularity: (i) prospect theory and framing effects, (ii) cognitive error, (iii) aversion to regret and (iv) self-control problems. We now consider each of these in turn.

Prospect theory has been rejected as an explanation of DCA's popularity by subsequent empirical studies and section 5 above demonstrates more generally that DCA is a sub-optimal strategy regardless of the form taken by investor risk preferences. However, Statman also suggests that DCA is attractive because it frames investment outturns in a flattering manner, allowing investors to feel that they have already gained by buying at a lower average cost. The cognitive error that I identify above is consistent with this framing effect, and shows exactly why comparing DCA's average cost with the average price is misleading: investors who frame their choice in terms of this comparison are making a specific mathematical error. However, even though this error leads investors to choose a strategy which is demonstrably (and measurably) inefficient in terms of the terminal wealth outturns it generates, Statman's argument suggests that the psychological feelings of wellbeing that this misleading framing creates could in principle offset the direct inefficiency costs of DCA.

Statman's other points are based on indirect benefits to investors resulting from the rigid investment timetable that DCA imposes. First, he identifies another form of cognitive error: investors' misguided belief that using their discretion on investment timing will help boost returns. There is plenty of evidence that investors' market timing has tended to be poor (e.g. Ritter 1991; Loughran and Ritter 1995), so DCA can indirectly increase expected returns by preventing investors from trying to time the market. However, Hayley (2014) shows that for the average US equity investor the effect of this bad timing is much smaller than has been suggested by other recent studies, and smaller than the estimated efficiency losses shown in Table 2.3. Investors with particularly bad timing may still find that the benefits of not trying to time the market outweigh the inefficiencies of DCA, but this does not appear to be the case for the average investor. Statman also notes that the discipline imposed by DCA (i) prevents a myopic desire for greater current consumption from interfering with investors' long-term investment goals; (ii) reduces the feelings of regret resulting from adverse market outcomes since investors feel less responsibility when DCA restricts their choices. The strength of these explanations is that the existence of these behavioural finance effects has been well established in other contexts. This comes in stark contrast to the additional assumptions required by some rationalist explanations for DCA's popularity.

A normative case for using DCA can thus be constructed by weighing the inherent inefficiency of DCA against the combined welfare costs of the investor regret, myopia and bad timing associated with less disciplined investment strategies. However, the current ill-informed analysis in the media gives little reason to suppose that such sensible reasoning is often what leads investors to choose DCA. Instead DCA is generally recommended by its proponents to all investors, with no reference to their specific preferences, objectives or beliefs. Avoidance of regret is sometimes mentioned, but by far the most common rationale given for DCA is that it boosts returns by buying at an average cost which is lower than the average price. A positive explanation for DCA's popularity needs to address this argument. Identifying the cognitive error in this argument thus gives a more straightforward explanation for why many investors still choose DCA.

Identifying this cognitive error also brings very different welfare implications. Previous research has argued that DCA could be welfare-improving (and hence an entirely rational choice) for investors with specific types of non-variance risk preferences. The analysis above rejects this argument. The wider behavioural finance benefits identified by Statman (1995) suggest that use of DCA may nevertheless be beneficial. This possibility is hard to prove or disprove, given that such psychological benefits cannot readily be measured. By contrast, the present paper identifies the cognitive error in the key argument used by DCA's proponents, and argues that in failing to spot this error investors who use DCA may actually be reducing their welfare.

Conclusion

DCA has long been shown to be mean-variance inefficient, so most recent research has focused on explaining why it nevertheless remains so popular. Recent papers have attempted to derive entirely rational explanations for DCA's popularity, but these are not satisfactory. Specifically, this paper demonstrates that DCA is a sub-optimal strategy regardless of investor risk preferences, so explanations based on alternative forms of such preferences must be rejected.

The other contribution made by this paper is to explicitly identify the error involved in comparing DCA's average purchase cost with the average market price. DCA's popularity can now be regarded as resulting from a specific and demonstrable cognitive error. This gives us a better explanation of DCA's popularity since – unlike most other explanations – it addresses the argument that is normally central to the case made by proponents of DCA.

References

Brennan, M. J., F. Li, and W.N. Torous, 2005. Dollar cost averaging. *Review of Finance* 9, 509-535.

Constantinides, G. M., 1979. A note on the suboptimality of dollar-cost averaging as an investment policy. *Journal of Financial and Quantitative Analysis* **14** 443-450.

Dichtl, H., and W. Drobetz, 2011. Dollar-cost averaging and prospect theory investors: an explanation for a popular investment strategy. *The Journal of Behavioral Finance* **12** 41-52.

Dubil, R., 2005. Lifetime dollar-cost averaging: forget cost savings, think risk reduction. *Journal of Financial Planning* **18** 86-90.

Dybvig, P.H., 1988a. Inefficient dynamic portfolio strategies or how to throw away a million dollars in the stock market. *The Review of Financial Studies* **1** 67-88.

Dybvig, P. H., 1988b. Distributional analysis of portfolio choice. Journal of Business, 369-393.

Greenhut, J. G., 2006. Mathematical illusion: why dollar-cost averaging does not work. *Journal of Financial Planning* **19** 76-83.

Hayley, S., 2014. Hindsight effects in dollar-weighted returns. *Journal of Financial and Quantitative Analysis*, **49**(1) 249 – 269.

Knight, J. R., and L. Mandell, 1992/93. Nobody gains from dollar cost averaging: analytical, numerical and empirical results. *Financial Services Review* **2** 51-61.

Fruhwirth, M., and G. Mikula, 2008. Can prospect theory explain the popularity of savings plans? *Working paper, available at SSRN: http://ssrn.com/abstract=1681343.*

Leggio, K., and D. Lien, 2001. Does loss aversion explain dollar-cost averaging? *Financial Services Review* **10** 117-127.

Leggio, K., and D. Lien, 2003. Comparing alternative investment strategies using risk-adjusted performance measures. *Journal of Financial Planning* **16** 82-86.

Loughran, T., and J. R. Ritter, 1995. The New Issues Puzzle. *Journal of Finance* **50** 23-51.

Milevsky, M. A., and S.E. Posner, 2003. A continuous-time re-examination of the inefficiency of dollar-cost averaging. *International Journal of Theoretical & Applied Finance* **6** 173-194.

Rieger, M. O., 2011. Co-monotonicity of optimal investments and the design of structured financial products. *Finance and Stochastics* **15** 27-55.

Ritter, J., 1991. The long-run performance of initial public offerings. *Journal of Finance* **46** 3-27.

Rozeff, M. S., 1994. Lump-sum investing versus dollar-averaging. *Journal of Portfolio Management* **20** 45-50.

Statman, M., 1995. A behavioral framework for dollar-cost averaging. *Journal of Portfolio Management* **22** 70-78.

Thorley, S. R., 1994. The fallacy of dollar cost averaging. *Financial Practice and Education* **4** 138-143.

Trainor, William J Jr., 2005. Within-horizon exposure to loss for dollar cost averaging and lump sum investing. *Financial Services Review* **14** 319-330.

Williams, R. E. and P.W. Bacon, 1993. Lump-sum beats dollar cost averaging. *Journal of Financial Planning* **6** 64–67.

Chapter 3

Dynamic Strategy Bias of IRR and Modified IRR – the Case of Value Averaging

Abstract

This chapter demonstrates that the IRR and modified IRR are biased indicators of expected profits for any dynamic strategy which is based on a target return or profit level, or which takes profits or "doubles down" following losses. Value Averaging is a popular example of such a dynamic strategy, but this chapter shows that it is inefficient under any plausible investor risk preferences and quantifies the resulting welfare losses. Value Averaging appears to be popular because investors mistakenly assume that the strategy's attractive IRR implies greater expected terminal wealth.

JEL Classification: G11

I am grateful to William Bernstein, Ilya Dichev, Maela Giofré, Aneel Keswani, Ionnis Kyriakou, Ian Marsh, Nick Motson, Richard Payne, Giorgio Questa, Nick Ronalds and participants at the 2012 Behavioural Finance Working Group and the May 2012 CFA UK seminar for useful comments, and in particular to Stewart Hodges for the derivation of the results in the appendix. The usual disclaimer applies.

Dynamic Strategy Bias of IRR and Modified IRR – the Case of Value Averaging

1. Introduction

Value averaging (VA) is a popular formula investment strategy which invests available funds gradually over time so as to keep the portfolio growing at a pre-determined target rate. It is recommended to investors because it demonstrably achieves a higher internal rate of return (IRR) than plausible alternative strategies. An online search on "value averaging" and "investment" shows many thousands of references to this strategy. These references, and those in other media, are overwhelmingly positive, recommending the strategy to investors as a means of boosting expected returns.

The use of the IRR to evaluate investor returns may seem intuitive, since it takes into account the varied cashflows that are inherent in dynamic strategies such as VA. However, this paper demonstrates that the IRR recorded for any VA strategy is systematically biased up. This bias retrospectively increases the weight given in the IRR calculation to periods with strong returns and reduces the weight given to weaker returns.

This bias is not specific to VA. It affects the IRR of any dynamic strategy which links the scale of future investment to the returns achieved to date. This includes any strategy which is based on a target return or profit level, or which includes any systematic element of taking profits, or "doubling down" after taking losses. I demonstrate below shows that the modified internal rate of return (MIRR) is similarly biased.

The higher IRRs recorded for VA are likely to be entirely due to this retrospective bias. VA does not increase expected terminal wealth – indeed, it is likely to reduce it because it delays investment. I demonstrate below that VA is an inefficient strategy for any plausible investor risk preferences and quantify the resulting welfare losses. Certain types of weak form inefficiency in market returns could in principle justify the use of VA but it would be an inefficient means of profiting from such inefficiencies. VA may bring some behavioural finance benefits but, as discussed in section 7 below, simpler strategies are likely to be more attractive. Thus not only does VA not generate the higher expected profits that are claimed, it is also likely to significantly reduce investor welfare.

VA's proponents recommend the strategy on the grounds of its higher IRR. The contribution of the present paper is to demonstrate that (i) the IRR and MIRR are systematically biased indicators of expected profits for a wide range of dynamic strategies; (ii) the attractive IRRs achieved by VA are likely to be entirely due to this bias; (iii) VA is an inefficient strategy for any plausible investor risk preferences.

2. The Value Averaging Strategy

VA is similar in some respects to Dollar Cost Averaging, which is the strategy of building up exposure gradually by investing an equal dollar amount each period. DCA automatically buys an increased number of shares after prices have fallen and so buys at an average cost which is lower than the average price over these periods (Table 3.1 shows an example). Conversely, if prices rose DCA would purchase fewer shares in later periods, again achieving an average cost which is lower than the average price over this period (Table 3.2). As long as there is any variation in prices DCA will always achieve a lower average cost.

Table 3.1: Illustrative Comparison Of VA and DCA – Declining Prices

DCA and VA strategies are used to buy an asset whose price varies over time (the price could also be interpreted as a price index, such as an equity market index). DCA invests a fixed amount each period ($100). VA invests whatever amount is required to increase the portfolio value by $100 each period. Both strategies buy at an average cost which is below the average price.

		Dollar Cost Averaging (DCA)			Value Averaging (VA)		
Period	Price	Shares bought	Investment ($)	Portfolio ($)	Shares bought	Investment ($)	Portfolio ($)
1	1.00	100	100	100	100	100	100
2	0.90	111	100	190	122	110	200
3	0.80	125	100	269	153	122	300
Total		336	300		375	332	
Avg.price 0.90		Avg.cost:	0.893		Avg.cost:	0.886	

VA is a more complex strategy because the additional sum invested each period is not constant. The investor sets a target increase in portfolio value each period (assumed here to be a rise of $100 per period, although the target can equally well be defined as a percentage increase) and at the end of each period must make whatever additional investments are necessary in order to meet this target. Like DCA, VA purchases a larger number of shares after a fall in prices, but the response is more aggressive: Table 3.1 shows that in order to achieve its target portfolio value VA must make up for the $10 loss it suffered in period 1 by investing an additional $10 in period 2. Thus VA buys 122 shares in period 2, compared to 111 for DCA. The greater sensitivity of VA to shifts in the share price results in an even lower average purchase cost. Again, this is true whether prices rise, fall or merely fluctuate.

Table 3.2: Illustrative Comparison Of VA and DCA – Rising Prices

Strategies are as defined in Table 3.1. The price of the asset is here assumed to rise. Again, both strategies buy at an average cost which is below the average price.

		Dollar Cost Averaging (DCA)			Value Averaging (VA)		
Period	Price	Shares bought	Investment ($)	Portfolio ($)	Shares bought	Investment ($)	Portfolio ($)
1	1.00	100	100	100	100	100	100
2	1.10	91	100	210	82	90	200
3	1.20	83	100	329	68	82	300
Total		274	300		250	272	
Avg.price 1.10		Avg.cost:	1.094		Avg.cost:	1.087	

VA could in principle be applied over any time horizon, but its originator suggests quarterly or monthly investments (Edleson, 1991). A number of mutual funds now facilitate VA by offering schemes which automatically invest additional funds in amounts which are linked to the value of the investor's existing portfolio.

Despite its popularity, VA has so far been the subject of limited academic research. VA commits the investor to follow a fixed rule, allowing no discretion over subsequent levels of investment. As a result, it is subject to the criticism of Constantinides (1979), who shows that strategies which pre-commit investors in this way will be dominated by strategies which instead allow investors to react to incoming news. VA might seem to improve diversification by making many small purchases, but Rozeff (1994) shows that this is not the case for DCA. The same reasoning applies for VA: Both strategies start with a very low level of market exposure, so the terminal wealth will be much more sensitive to returns later in the horizon, by which time the investor is more fully invested. Better diversification is achieved by investing in one initial lump sum, and thus being fully exposed to the returns in each period. An investor who has funds available should invest immediately rather than wait.

Unlike DCA, VA's cashflows are volatile and unpredictable. Each period investors must add whatever amount of new capital is required to bring the portfolio up to its pre-defined target level, so these cashflows are determined by returns over the most recent period. Edleson envisages investors holding a 'side fund' containing liquid assets sufficient to meet these needs[9].

Although VA generates impressive IRRs, empirical studies show no corresponding outperformance on other performance measures. Thorley (1994) compares VA with a static buy-and-hold strategy for the S&P500 index over the period 1926-1991 and finds that it performs worse in terms of mean annual return, Sharpe ratio and Treynor ratio. Leggio and Lien (2003) find that the rankings of these three strategies depend on the asset class and the performance measure used, but the overall results do not support the benefits claimed for VA.

However, VA's proponents continue to stress its demonstrable advantage: achieving a higher expected IRR than alternative strategies (Edleson (1991), Marshall (2000, 2006)). This appears to be the key to VA's popularity. The following sections demonstrate that the IRR is raised by a systematic bias which allows VA to generate attractive IRRs even without increasing expected profits.

[9] Edleson (1991) and Marshall (2000, 2006) both calculate the IRR on the VA strategy without including returns on the side fund. We follow the same approach here in order to demonstrate that even in the form used by its proponents VA does not generate the higher returns that are claimed. Thorley (1994) rightly criticises the exclusion of the returns on cash in the side fund. However, including a side fund does not remove the bias: The modified IRR includes cash holdings, but I demonstrate in section 4 that this too is a biased measure of VA's profitability.

3. Simulation Evidence

VA is recommended by its proponents as a strategy which boosts expected returns in any market, even if the investor has no ability to forecast returns. Edleson (1991, 2006), Marshall (2000, 2006) and other proponents demonstrate that VA generates a higher IRR than alternative strategies even on simulated random walk data (corresponding dollar profits are not calculated). By contrast, Thorley (1994) shows VA generating lower average dollar profits than investing in one initial lump sum, but does not calculate the IRRs. In this section I use a consistent set of simulations to demonstrate that the IRR is a biased measure of the profitability of VA. The following section derives this result more formally and demonstrates how this bias arises.

We assume here that returns follow a random walk. This is consistent with the fact that investors who use VA are unlikely to believe that they are able to forecast short-term returns. Those who (rightly or wrongly) believe that they have such forecasting ability should prefer alternative strategies which – unlike VA – allow them some discretion over the timing of their investments. I consider in section 6 whether weak form inefficiencies in market returns could justify the use of VA.

The simulations also assume that this random walk has zero drift. This is the simplest assumption, and it is generous to VA. A more realistic assumption of upward drift would penalize VA since its relatively large initial holdings of cash would then earn a lower expected return than those invested in risky assets. For simplicity we also assume that the security that is purchased pays no dividend or other income. This assumption is similarly generous to VA.

Table 3.3 compares the average costs, IRRs and profits achieved by VA and DCA with those obtained by a simple strategy of investing in one initial lump sum. Both VA and DCA achieve significantly lower average purchase costs and higher IRRs, but VA appears to be the most attractive strategy when judged on either of these criteria. Yet, despite this, there is no significant difference between the dollar profits generated by these three strategies.

Table 3.3: Simulation Results: Performance Differentials

This table compares strategies which invest in an asset whose returns are assumed to follow a random walk with no drift. The first row compares DCA with a strategy which immediately invests the same total amount immediately in one lump sum. The second compares VA with this lump sum strategy. Following Marshall (2000, 2006), security prices are assumed to start at $10 and then evolve for five periods in each of 100,000 simulations. In each period returns are *niid* with mean zero and 10% standard deviation. DCA invests a fixed $400 each period; VA invests whatever amount is required to increase the portfolio value by $400 each period; the lump sum strategy invests $2000 in the first period. The expected terminal wealth of all three strategies will thus be identical if prices remain unchanged. Standard errors are shown in brackets. Asterisks *** indicate significance at 0.1%.

	Average Cost (cents)	IRR (%)	MIRR (%)	Profit ($)
DCA - Lump Sum	-7.80***	0.082***	0.222***	-0.387
	(0.35)	(0.007)	(0.007)	(0.704)
VA - Lump Sum	-19.75***	0.305***	0.461***	-0.31
	(0.34)	(0.007)	(0.007)	(0.72)

By buying more shares when they are relatively cheap, DCA always achieves an average purchase cost which is below the average price. As we saw above, VA responds more aggressively than DCA (by increasing the sum invested in the second period) and thus achieves an even larger reduction in its average purchase cost than DCA. All else equal, lower average costs would lead to higher profits, but all else is not equal here since the different strategies invest different total amounts. These dynamic strategies buy fewer shares after prices have risen and more after they have fallen. This reduces the average purchase price (compared to the counterfactual of buying equal numbers of shares in each period, and thus buying at an average cost equal to the unweighted average price). But profits are only increased by buying more shares before a rise, and fewer before a fall. DCA and VA achieve their lower average purchase costs by means of a retrospective response which has no effect on expected profits.

4. The Bias in the IRR

Edleson (1991) and Marshall (2000, 2006) focus exclusively on the IRRs achieved by VA. This might seem a reasonable approach, since the IRR takes account of the fluctuating cashflows that are an inherent part of the strategy. However, these IRRs are systematically misleading. In Chapter 4 I demonstrate that the aggregate IRR for the US equity market is biased down as investors “chase returns” by increasing their exposures following strong returns. This section uses the same approach to demonstrate that, by contrast, VA automatically biases the IRR up.

An investor’s portfolio value at the end of period *t* (K_t) is determined by the return in the previous period plus any additional top-up investment a_t made at the end of this period:

$$K_t = K_{t-1}(1 + r_t) + a_t \tag{1}$$

By definition, when discounted at the IRR, the aggregate present value of these investments equals the present value of the final value in period *T*:

$$K_0 + \sum_{t=1}^{T} \frac{a_t}{(1 + IRR)^t} = \frac{K_T}{(1 + IRR)^T} \tag{2}$$

Substituting from equation 1 allows us to eliminate a_t (following Dichev and Yu, 2009) and to demonstrate that the IRR is a weighted average of the returns in each period (r_t), where the weights reflect the present value of the portfolio at the beginning of each period:

$$IRR\sum_{t=1}^{T}\frac{K_{t-1}}{(1+IRR)^{(t-1)}}=\sum_{t=1}^{T}\left(\frac{K_{t-1}}{(1+IRR)^{(t-1)}}\times r_t\right) \tag{3}$$

Re-arranging further shows that the returns in any period may be above or below the IRR, but the weighted sum of these deviations is zero:

$$\sum_{t=1}^{T}\left(\frac{K_{t-1}}{(1+IRR)^{(t-1)}}(r_t-IRR)\right)=0 \tag{4}$$

Dividing the horizon in two shows the effect on the IRR of a single additional investment at the end of period *m* which has a value equal to *b*% of the portfolio at that time:

$$\sum_{t=1}^{m}\left(\frac{K_{t-1}}{(1+IRR)^{(t-1)}}(r_t-IRR)\right)+(1+b)\sum_{t=m+1}^{T}\left(\frac{K_{t-1}^{*}}{(1+IRR)^{(t-1)}}(r_t-IRR)\right)=0 \tag{5}$$

Additional investment after period *m* increases the weight given to later returns, compared to the weights based on the portfolio values K_t^* which would otherwise have been seen. If, for example, the periodic returns r_t up to period *m* were low, then these early *(r_t - IRR)* terms will tend to be negative, and subsequent terms will tend to be positive. A large new investment at this point would increase the weight given to subsequent *(r_t- IRR)* terms relative to the earlier terms so the IRR must increase in order to keep the weighed sum at zero. Similarly, investing less (or even withdrawing funds) after a period of strong returns will tend to reduce the relative weight given to later *(r_t- IRR)* terms, which would tend to be negative. This too would increase the IRR.

However, the impact on the IRR could reflect two very different effects. The IRR could be raised by relatively large additional investments taking place ahead of periods with relatively high returns. This would represent good investment timing and would clearly increase expected profits, but this effect cannot explain the high IRRs in the simulations since our assumption of a random walk means that future returns are unforecastable and investments will on average be badly timed as frequently as they are well timed.

However, a large new investment will not only increase the weight which the IRR calculation gives to future returns, it will also reduce the weight given to earlier returns (equation (3) shows that these weights sum to unity). This would be a retrospective adjustment which will boost the expected IRR even if (as in our simulations) there is no relationship between these intermediate cashflows and future returns. In this situation the IRR becomes a biased indicator of the profitability of this investment strategy, and we know that this bias is inherent in VA, since by construction disappointing returns are followed by larger net investments in order to raise the portfolio value to its target level.

Specifically, the net investment demanded by VA each period is determined by the degree to which organic growth in the value of the portfolio over the immediately preceding period (r_mK_{m-1}) fell short of the investor's target. The first summation in Equation 5 includes r_m so the level of new investment b will tend to be large (small) when the first summation is negative (positive). The second summation will be correspondingly positive (negative) and will be given more (less) weight as a result of this additional investment. All else equal, the weighted sum over all periods would become positive, but the IRR then rises to return the sum to zero. Thus VA biases the IRR up by automatically ensuring that the size of each additional investment is negatively correlated with the preceding return.

Phalippou (2008) shows that the IRRs recorded by private equity managers can be deliberately manipulated by returning cash to investors immediately for successful projects and extending poorly-performing projects. VA cannot change the end of the investment horizon in this way. Instead it achieves its bias by reducing the weight given to returns later in the horizon following good outturns, and increasing it following poor returns.

More generally, because the IRR is in effect a weighted average of individual period returns it can be biased by following any strategy which retrospectively reduces the weight given to bad outturns and increases the weight given to good outturns. This will be a property shared by any strategy which targets a particular level of portfolio growth, systematically takes profits after strong returns or “doubles down” after weak returns, since all these strategies invest more after poor returns and so give less weight in the IRR calculation to these prior returns (after strong returns they invest less than they otherwise would, thus increasing the relative weight given to these strong returns). It is by doing this automatically that VA raises its expected IRR.

Including Edleson’s “side fund” in the calculation is not sufficient to avoid this bias. We must also ensure that the size of this side fund is fixed in advance and not adjusted retrospectively. This can be seen from the bias in the modified internal rate of return (MIRR), and can be illustrated with a simple two period example. Suppose an investor initially allocates *a* to risky assets and *b* to the side fund, where it earns a risk-free return r_f. At the end of period 1 an amount *c* from the side fund is used to buy additional risky assets.

$$\textit{Terminal Wealth (TW)} = a(1+r_1)(1+r_2) + c(1+r_2) + (b(1+r_f) - c)(1+r_f) \qquad (5)$$

This measure is not affected by any retrospective adjusment, since the weight attached to r_1 is fixed in advance. Including the side fund in the calculation of the IRR means that intermediate cashflows just become a shift from one part of the portfolio to the other, leaving just the initial and terminal cashflows. Thus the IRR simply becomes the geometric mean return:

$$IRR = \sqrt{\frac{TW}{a+b}} - 1 \qquad (6)$$

This too is unbiased if *a* and *b* are both fixed in advance. The bias comes about because the side funds must be sufficiently large to meet the VA strategy's future cash needs, but this is a function of future returns and so is unknown. This tends to lead to the size of the side fund being set retrospectively to ensure that it is sufficient. This can be illustrated by considering the modified internal rate of return (MIRR), which assumes the existence of a side fund which is just big enough to fund subsequent cash injections (implying that $b(1+r_f)=c$ in the expression above for terminal wealth). Hence:

$$MIRR = \sqrt{\frac{a(1+r_1)(1+r_2)+c(1+r_2)}{a+c/(1+r_f)}} - 1 \qquad (7)$$

The MIRR is biased because the relative weight $a(1+r_2)/(a+c/(1+r_f))$ given to r_1 is adjusted retrospectively. VA automatically ensures that a low r_1 will be followed by a large cash injection *c*, so the expected relative weight given to r_1 is automatically reduced, increasing the MIRR. The weight on r_1 is changed after the event, so although this alters the MIRR it has no effect on expected terminal wealth. Thus VA also increases the expected MIRR because of a retrospective bias[10]. This bias is confirmed by the simulation results in Table 3.3, which show that VA generates a higher MIRR than investing in one initial lump sum, but without increasing expected terminal wealth.

[10] There is no bias if strong returns in period 1 lead to assets being sold and the proceeds added to the side fund. This is because only cash injections (new investments) are added in the denominator: the $c/(1+r_f)$ term is omitted if $c<0$. But in a multi-period setting the MIRR will only be unbiased if there are no additional cash injections in any period.

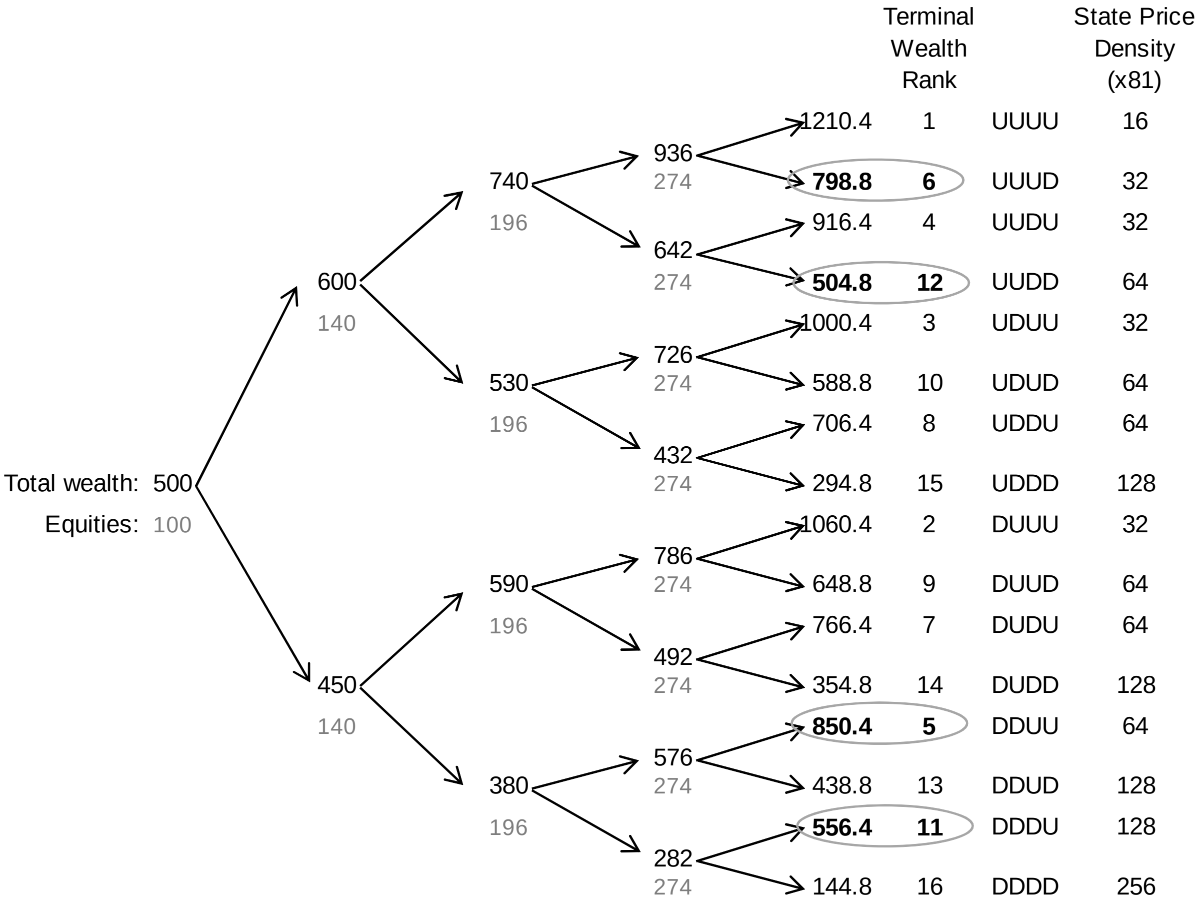
5. The Inefficiency of Value Averaging

The analysis above showed that VA does not generate the higher expected profits that its higher expected IRR would suggest. In this section I go one step further and demonstrate that VA is an inefficient strategy, with other strategies offering preferable risk-return characteristics. For now we maintain our assumption that asset returns follow a random walk. We will relax this assumption later, when we consider the use of VA in inefficient markets.

I here use the payoff distribution pricing model derived by Dybvig (1988b) to demonstrate that VA is inefficient. Figure 1 shows the simplest possible illustration of this technique, using a binomial model of the terminal wealth generated over four periods by a VA strategy. In a good outturn equity prices are assumed to double, whilst they halve in a bad outturn. The investor has chosen a portfolio growth target of 40% each period and initially invests 100 in equities. If the value of these equities rises in the first period to 200, then 60 is assumed to be transferred to the side fund, which for simplicity we assume offers zero return. Conversely, a loss in the first period sees the equity portfolio topped up from the side account to the target 140.

Figure 3.1: Simple Model of VA Strategy

This figure shows the total investor wealth (the upper figure at each point) for a VA strategy with a portfolio growth target of 40% each period. The lower figures show the amount of this total wealth which is held in equities. Equity values are assumed to double in a good outturn and halve in a bad outturn. Equity investment is adjusted back to the target value after each period using transfers into and out of the side account. For illustrative purposes funds in the side account are assumed to earn zero interest (Table 3.4 shows that inefficiencies persist with a higher risk free rate). All paths are assumed to be equally likely.



The inefficiency of this strategy can be demonstrated by comparing the ranking of the terminal wealth outturns and the state price densities (the state prices divided by the probability – for this tree they are $16(1/3)^{u}(2/3)^{d}$, where *u* is the number of up states and *d* the number of down states on the path concerned[11]). Higher terminal wealth outturns

[11] More generally, the state price densities of one period up and down states are $(1/(1+\mathrm{r}\Delta t))(1-((\mu - \mathrm{r})\Delta t/\sigma\sqrt{\Delta t}))$ and $(1/(1+\mathrm{r}\Delta t))(1+((\mu - \mathrm{r})\Delta t/\sigma\sqrt{\Delta t}))$ respectively, where r is the

generally come in the paths with lower state price densities, but not always. The best outturn is in the UUUU path, which has the lowest state price density. The second, third and fourth best outturns see three ups and one down. But the fifth best is DDUU, which beats UUUD into sixth place. Similarly, DDDU in eleventh place beats UUDD.

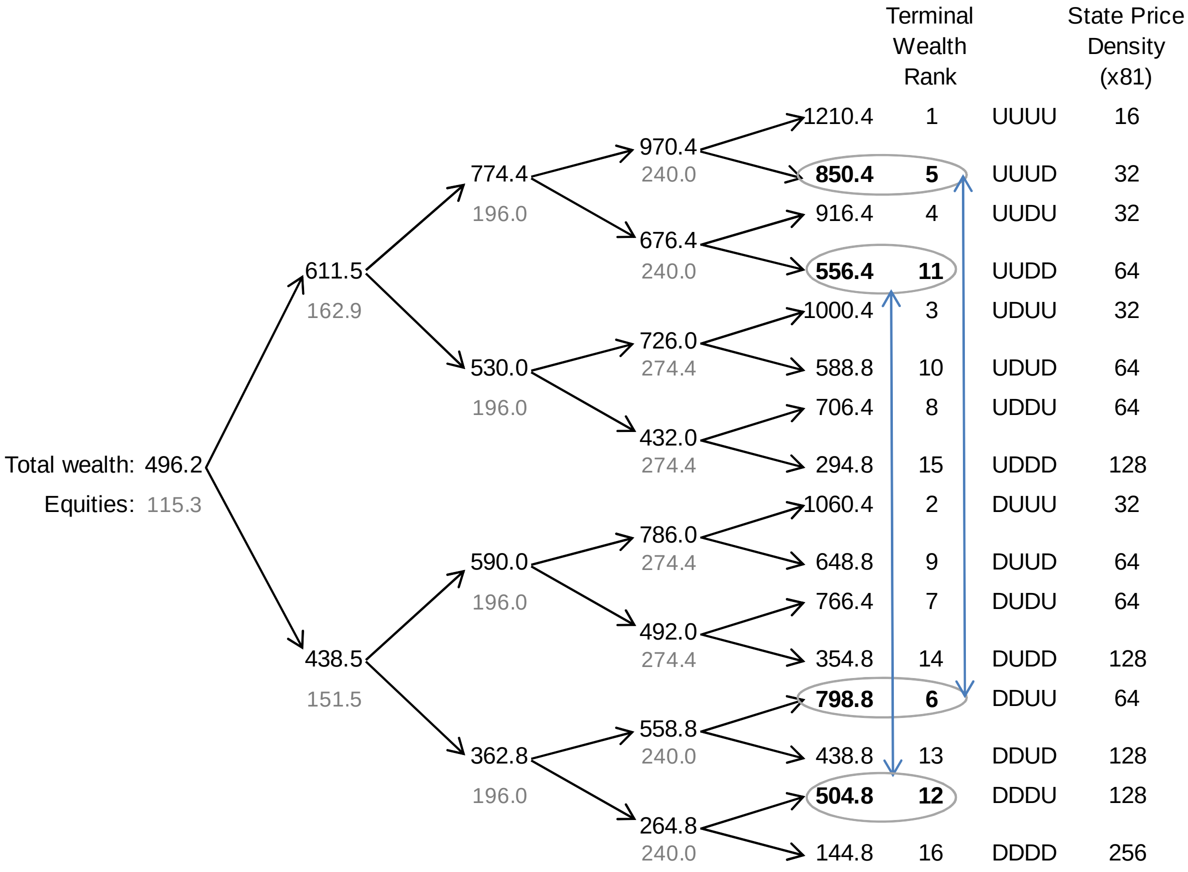
These results show that the VA strategy fails to make effective use of some relatively lucky paths (those with relatively low state price densities). This can be proved by generating a strategy which produces exactly the same 16 outturns with a smaller initial investment. We do this by altering our strategy so that the paths with the lowest state price densities (the largest number of up states) generate the greatest terminal wealth, so we swap the 5th highest outturn in Figure 1 with the 6th and the 11th highest with the 12th. We then work backwards through the tree using the state prices to calculate the equity and cash which must be held at each prior point. Ultimately this determines the initial capital which is needed. This new strategy is shown in Figure 3.2 and needs only 496.2 initial capital, rather than the 500 above, thus demonstrating the extent to which VA is inefficient. By generating the same set of possible outturns this alternative strategy must be taking the same level of risk as VA, no matter which measure of risk we use.

The reduction in the initial capital required is a measure of VA's inefficiency compared to our alternative strategy. This is a powerful technique because it demonstrates VA's inefficiency without needing to specify the investor's risk preferences. Producing the same set of outturns with less initial capital must be preferable regardless of the investors' risk preferences, profided only that terminal wealth is what investors care about, and that they prefer more terminal wealth to less.

continuously compounded annual risk-free interest rate and the risky asset has annual expected return μ and standard deviation σ. The corresponding one period risky asset returns are $(1+\mu\Delta t+\sigma\sqrt{\Delta t})$ and $(1+\mu\Delta t-\sigma\sqrt{\Delta t})$. See Dybvig (1988b).

Figure 3.2: Optimized Strategy Which Generates Identical Outturns To VA

The upper figure shows the total investor wealth at each point in a strategy in which the equity exposure (the lower figure at each node) has been set so as to replicate the total wealth outturns in Figure 1, but with these outturns optimized so that the largest terminal wealths are generated in the states with the lowest state price density. Compared with Figure 1, the outturns for UUUD and DDUU have been swapped, and for UUDD and DDDU. Equity returns are as assumed in Figure 1. The lower initial capital required for this optimized strategy to generate an identical set of outturns shows the degree to which the VA strategy is inefficient.



These results also confirm that VA is inefficient because it invests gradually, and thus has little risk exposure early in the investment horizon. VA generates lower terminal wealth in paths which include a comparatively large number of strong returns early in the horizon. Thus beats UUDD is beaten by DDDU, and UUUD is beaten by DDUU, showing that VA fails to take advantage of some early strong returns. This is the source of the inefficiency that we have demonstrated here.

The doubling or halving of equity values in each period is an extreme assumption – for typical levels of equity volatility this would imply several years between successive investments. This allows us to illustrate dynamic inefficiencies in a short tree, but it is unrealistic for most investors. For a more realistic strategy we consider an eighteen period tree. This has 2^{18} paths, and is the largest that was computationally practical.[12] This analysis continues to assume that returns follow a binomial distribution, but the inefficiency of VA extends to other distributions. Rieger (2011) generalizes Dybvig's results to show that path-dependent strategies which generate outturns which have a non-monotonic relationship with market returns will be sub-optimal no matter what distribution these market returns follow. VA is an example of such a path-dependent strategy.

Panel A in Table 3.4 shows the degree of inefficiency in VA strategies over a range of different time horizons and target growth rates. These were derived using a risk free rate of 5%, and risky asset returns with mean 10% and standard deviation 20% (all per annum). These efficiency losses remain very similar for a range of different volatilities (to save space these are not reproduced here).

These discrete time figures are likely to understate the true efficiency losses for two reasons. First, the limited number of paths which can be computed results in comparatively large differences between ranked terminal wealth outturns. Thus small potential inefficiencies will not be recorded if they do not reduce the terminal wealth on

[12] Dybvig (1988b) uses this technique to demonstrate the inefficiency of stop-loss and target return strategies which are invested either fully in the risky asset, or fully in the risk-free asset. The number of paths involved is thus limited since the tree is generally recombinant, and collapses to a single path on hitting the target portfolio value. By contrast, VA varies the exposure in successive periods so DU and UD paths will not result in the same portfolio value. Thus an *n* period tree has 2^n paths and computation rapidly becomes impractical as *n* rises.

one path sufficiently for it to fall below the terminal wealth achieved on at least one path with a higher state price density. This problem can be avoided by shifting to continuous time. This represents a simplification, since VA is intended to make any required additional investments at discrete (eg. monthly) intervals. But it has the advantage that all inefficiencies will be recorded since there will be an indefinite number of different paths with terminal wealths which differ only minutely. An expression for the efficiency losses resulting from VA is derived in the appendix. These continuous time estimates are indeed entirely unaffected by the level of price volatility assumed for the asset, and the efficiency losses (Panel B of Table 3.4) are substantially greater than the discrete time estimates.

Table 3.4: Measuring the Dynamic Efficiency Losses of Value Averaging

This table shows the additional initial capital required by a VA strategy compared with an optimized strategy which generates an identical set of final portfolio values. These figures are derived using the Dybvig PDPM model applied to a VA strategy over an 18 period tree with risk free rate 5%, expected market return 10% and volatility 20% (all per annum). The inefficiency is shown as a percentage of the average terminal portfolio value of the VA strategy. For the discrete time calculation an 18 period tree is used throughout, with the length of each period varied to achieve the total time horizon shown. The derivation of the corresponding continuous time losses is shown in the appendix.

Panel A: Discrete time estimates of efficiency losses over investment horizon

Target growth (per annum)	5 years	10 years	15 years	20 years
-10%	1.16%	4.93%	8.35%	10.57%
-5%	0.21%	1.88%	3.67%	5.06%
0%	0.00%	0.15%	0.62%	1.10%
5%	0.00%	0.00%	0.00%	0.00%
10%	0.00%	0.08%	0.44%	0.85%
15%	0.10%	1.37%	2.84%	4.02%
20%	0.67%	3.51%	6.23%	8.14%

Panel B: Continuous time estimates of efficiency losses over investment horizon

Target growth (per annum)	5 years	10 years	15 years	20 years
-10%	0.57%	4.33%	13.43%	28.73%
-5%	0.26%	2.01%	6.50%	14.59%
0%	0.06%	0.52%	1.72%	4.02%
5%	0.00%	0.00%	0.00%	0.00%
10%	0.06%	0.52%	1.72%	4.02%
15%	0.26%	2.01%	6.50%	14.59%
20%	0.57%	4.33%	13.43%	28.73%

Two results are clear in Table 3.4. First, VA becomes increasingly inefficient if the target growth rate is either very high or very low. Second, inefficiency increases dramatically as the time horizon is increased.

Intuitively, a VA strategy with a high target growth rate is likely to require substantial additional injections of funds over time to keep the portfolio value growing at its target rate. This will leave the investor's terminal wealth most sensitive to asset returns late in the horizon (when cash held is correspondingly low). Conversely, a low target return is likely to generate significant cash withdrawals, leaving the investor most exposed early in the horizon. Either extreme is inefficient compared to a strategy for which equity returns have equivalent impact on terminal wealth whenever they occur (as would be the case for a simple buy-and-hold strategy which immediately invests all available cash).

In practice target growth rates are likely to be in the higher part of the range shown in Table 3.4. There are three reasons for this. First, investors will naturally expect risky assets such as equities to generate an expected return equal to r_f plus a risk premium. Second, they are likely to overestimate this expected return in the mistaken belief that VA will boost returns above what could normally be expected on these assets. Third, VA is generally used as a means of investing new savings as well as generating organic portfolio growth, so the target growth rate is likely to be set above the expected rate of organic growth. Consistent with this, Edleson (1991) explicitly envisages that periodic cashflows will generally be additional purchases of risky assets rather than withdrawals of funds. Taking the risk premium to be 5% (as a very broad approximation), when we add investor overestimation of this risk premium and the desire to make further net investments, target growth rates are likely to be at least 5% higher than r_f, and quite plausibly 10% higher. Table 3.4 is calculated with r_f=5%, so the outturns shown for target growth rates in the range 10-15% are likely to be most representative.

Table 3.4 also shows that VA is much more inefficient over longer time horizons. VA is generally recommended as a long-term investment strategy (in particular for saving for retirement), so horizons of 10 to 20 years are likely to be more common than a

5 year horizon. Table 3.4 shows that over such time horizons, and with target growth rates in the range 10-15%, the dynamic inefficiency can be substantial.

Furthermore, all the figures in Table 3.4 (both discrete and continuous) should be regarded as conservative estimates of the welfare loss to investors. They show how much more cheaply an investor could achieve the same distribution of outturns as a VA strategy. This method allows us to derive these welfare losses without needing to make any assumption about the form of the investor's risk preferences, but there is no reason why an investor who abandons VA should actually choose an alternative strategy with exactly the same payoff distribution. Investors are likely instead to find other strategies even more attractive, implying that the actual welfare benefits of abandoning VA are higher than shown in Table 3.4

In particular, VA introduces a negative skew into the distribution of cumulative returns (compared to a lump sum investment), since larger additional investments are made following losses. For example, a series of negative returns could result in a VA strategy losing more than its initial invested capital as additional investments are made to keep the risk exposure at its target level. This would of course be impossible for a lump sum investment. Conversely, VA invests less following strong returns, restricting the upside tail. This negative skew will be welfare-reducing under many plausible utility functions, so abandoning VA is likely to bring such investors welfare benefits significantly larger than those shown in Table 3.4.

Thus whilst we can calculate plausible lower limits for the efficiency losses associated with VA, more realistic estimates are likely to be larger. Furthermore, the fact that there are efficiency losses for any distribution of returns and for any form of investor risk preferences comes in stark contrast to proponents' claims that VA outperforms alternative strategies.

6. Value Averaging In Inefficient Markets

In this section we consider whether VA could outperform in markets where asset returns contain a predictable time structure. However, it is worth stressing at the outset that this would be a much weaker argument in favor of VA than the outperformance in all markets (including random walks) which is claimed by VA's proponents. We also consider VA's performance against historical data.

This analysis is complicated by the fact that many popular performance measures will be inappropriate for assessing whether VA outperforms. The level of risk taken by VA depends on the growth target used, so differences in the expected return achieved by comparison strategies might simply reflect a different risk premium. This could normally be corrected for by comparing Sharpe ratios, but the negative skew in the cumulative returns generated by VA means that the Sharpe ratio will be misleading, since the comparatively small upside risk reduces the standard deviation of a VA strategy, even though investors are likely to prefer a larger upside tail. In addition, Ingersoll et al. (2007) show that performance measures such as the Sharpe ratio will be biased upwards when investment managers reduce exposure following good results and increase it following bad results. VA automatically adjusts exposures in this way, so there is also a dynamic bias which increases its Sharpe ratio.

Chen and Estes (2010) derive simulation results which explicitly include the cost of VA's side fund. These show that VA does indeed generate higher Sharpe ratios, but with greater downside risk. Given the negative skew, the Sortino ratio might be considered a more appropriate performance measure, but Chen and Estes show that VA generates a lower Sortino ratio than a lump sum investment. This is particularly discouraging since Ingersoll et al (2007) show that this ratio is also increased by the same dynamic bias as the Sharpe ratio.

Relaxing our previous assumption of weak-form efficiency, mean reversion in prices will tend to favor VA. Additional simulations (not reproduced here) suggest that single period autocorrelation has little impact on profits, but multi-period autocorrelation has a larger effect. For example, successive periods of low returns result in large cumulative additional investments which leave the portfolio well positioned for subsequent periods of high returns. Consistent with this, VA outperforms DCA in our earlier simulations when the terminal asset price ends up close to its starting value, and it underperforms DCA when prices follow sustained trends in either direction.

There is evidence of long-term reversals in some asset returns (eg. de Bondt and Thaler, 1985) but, conversely, there is also a large literature documenting positive autocorrelation in other markets (momentum or ‘excess trending’). The most relevant test for our purposes is whether VA outperforms when back-tested using historical returns. This will show whether any time structure in these market returns is sufficient to offset the innate inefficiency of VA.

Studies using historical data have not found that VA outperforms. Thorley (1994) calculates the returns to a VA strategy which invests repeatedly in the S&P500 index over a 12 month horizon for the period 1926-1991. He finds that the average Sharpe ratio of this strategy is below that of corresponding lump sum investments. Similarly, Leggio and Lien (2003) find that VA generates a Sharpe ratio which is lower than for lump sum investment in large capitalization US equities, corporate bonds or government bonds, with VA generating a larger Sharpe ratio only for small firm US equities. These results hold for both 1926-1999 and the more recent 1970-1999 period. The lower Sharpe ratios achieved by VA are particularly striking given the static and dynamic biases outlined above, which tend to bias the Sharpe ratio up.

This does not rule out the possibility that there are some markets which show time structures in their returns that VA could exploit but, as Thorley (1994) points out, even where suitable market inefficiencies can be detected, VA would be a very blunt instrument with which to try to profit from them. Other strategies are likely to be much more effective at extracting profits from such market inefficiencies, such as long/short strategies with buy/sell signals calibrated to the particular inefficiency found in historic returns in each market. Furthermore, any advantage gained by VA in such markets would have to outweigh the inherent inefficiency of the strategy, as demonstrated above. For all these reasons, market inefficiency is unlikely to be a convincing reason for using VA.

7. Behavioural Finance and Wider Welfare Effects

VA's proponents recommend the strategy on the basis of its higher IRR, making no claim that it has any wider benefits, but in this section we nevertheless consider whether wider welfare effects, such as behavioural finance effects, might explain why VA remains very popular.

Statman (1994) proposed several behavioural finance effects to explain DCA's popularity, and we now consider the extent to which they might apply for VA. First, prospect theory suggests that investors' utility functions over terminal wealth may be more complex than in traditional economic theory. However, this cannot explain VA's popularity, since we saw above that VA must be a sub-optimal strategy regardless of the form taken by investor risk preferences, since alternative strategies can duplicate VA's outturns at lower initial cost.

Statman also suggested that by committing investors to continue investing at a pre-determined rate DCA prevents investors from exercising any discretion over the timing of their investments, and so: (i) stops investors misguidedly attempting to time markets (investor timing has generally been shown to be poor); (ii) by giving investors no discretion over timing it avoids the feelings of regret that might follow poorly-timed investments. VA could plausibly bring similar benefits[13], but even in the light of such wider possible benefits, it is likely to remain a less attractive strategy than DCA. Both strategies commit the investor to adding cash according to a pre-specified rule, but VA's cashflows are unpredictable so this is likely to require more active investor involvement (compared to DCA's entirely stable and predictable cashflows), implying more potential for regret.

Furthermore, the need for a side fund of cash or other liquid assets to fund VA's uncertain cashflows is likely to lead investors to hold a higher proportion of their wealth in such assets than would otherwise be optimal, with correspondingly less invested in risky assets. Thus rather than overall portfolio allocations being chosen to maximize investor welfare, these strategic allocations may instead be determined by the liquidity needs of the VA strategy. This would imply a static inefficiency in addition to the dynamic inefficiency seen above.

The required size of the side fund will depend on the volatility of risky assets, but is likely to be substantial. With aggregate equity market volatility of around 15-20% per annum, a side fund of at least this fraction of the risky assets might be considered a bare minimum since we should anticipate occasional annual market returns substantially in

[13] The results derived in earlier sections assumed that investors always prefer greater terminal wealth to less, but this might not be true if regret is important, since investor utility would then depend on the path taken, rather than just the terminal wealth ultimately achieved.

excess of 20% below their mean. An alternative perspective is that another decade like 2000-2009 would see many markets stay flat or fall. For plausible levels of the target growth rate this would leave investors trying to find additional cash worth more than the original value of their investments.

Furthermore, VA requires investors to sell assets after any period in which organic growth in the portfolio exceeds the target growth rate. This may result in increased transaction costs compared to a buy-only strategy and, worse, could trigger unplanned capital gains tax liability. Edleson (2006) suggests that investors could reduce these additional costs by delaying or ignoring entirely any sell signals generated by the VA strategy, and that investors should in any case limit their additional investments to a level they are comfortable with. However, this re-introduces an element of investor discretion, implying possible bad timing and regret. By avoiding this DCA again appears to be the preferable strategy.

8. Conclusion

VA is recommended to investors as a method for raising investment returns in any market, even when prices follow a random walk. This paper shows that VA does indeed increase the expected IRR, but it does not increase expected profits. Instead the IRR is boosted by a retrospective bias which arises because VA invests more following poor returns and less following good returns. The same bias will be found for any strategy which varies its exposure in response to the return achieved to date. This includes all strategies based on a target return or profit level, and also those which systematically take profits following strong returns or "double down" following weak returns. The modified IRR is similarly biased.

In complete contrast to the outperformance that is claimed for it, VA is in fact an inefficient strategy. This paper identifies four sources of inefficiency: (i) VA is dynamically inefficient, except in the unlikely case that the target return is very close to the risk free rate (this is a powerful result since it applies regardless of the form taken by investor risk preferences); (ii) VA also introduces a downside skew to cumulative returns which is likely to be welfare-reducing for many investors; (iii) VA is likely to cause static inefficiency by requiring larger holdings of cash and liquid assets than would otherwise be optimal; (iv) VA may increase management costs, transaction costs and tax liabilities compared to a buy-and-hold strategy. Behavioural finance effects may be important enough to some investors that they outweigh all these inefficiencies, but for such investors VA is likely to be an inferior strategy to DCA, which has stable cashflows.

In short, VA has very little to recommend it. VA's popularity appears to be due to investors making a cognitive error in assuming that its higher IRR implies higher expected profits. More importantly, this is just one example of a dynamic strategy for which the expected IRR and MIRR are misleading indicators of expected profits. This is an important and very general point, since it is precisely for such dynamic strategies – with their variable periodic cashflows – that the IRR is likely to be used as a key performance metric.

References

Constantinides, G.M. (1979). “A Note On The Suboptimality Of Dollar-Cost Averaging As An Investment Policy.” *Journal of Financial and Quantitative Analysis* 14, 443-450.

Chen, H. and J. Estes, (2010). “A Monte Carlo Study Of The Strategies For 401(K) Plans: Dollar-Cost-Averaging, Value-Averaging, And Proportional Rebalancing.” *Financial Services Review* 19, 95-109.

De Bondt, W.F.M. and R. Thaler, (1985) “Does the Stock Market Overreact?” *Journal of Finance* 40, 793-805.

Dichev, I.D. and G. Yu, (2011) “Higher Risk, Lower Returns: What Hedge Fund Investors Really Earn.” *Journal of Financial Economics* 100, 248–263.

Dybvig, P.H. (1988a) “Inefficient Dynamic Portfolio Strategies or How to Throw Away a Million Dollars in the Stock Market.” *Review of Financial Studies* 1, 67-88.

Dybvig, P.H. (1988b) “Distributional Analysis of Portfolio Choice.” *Journal of Business* 61, 369-393.

Edleson, M.E. (1991) “Value Averaging: The Safe and Easy Strategy for Higher Investment Returns.” Wiley Investment Classics, revised edition (second edition 2006).

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

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

Leggio, K. and D. Lien (2003) “Comparing Alternative Investment Strategies Using Risk-Adjusted Performance Measures.” *Journal of Financial Planning* 16, 82-86.

Marshall, P.S. (2000) “A Statistical Comparison Of Value Averaging vs. Dollar Cost Averaging And Random Investment Techniques.” *Journal of Financial and Strategic Decisions* 13, 87-99.

Marshall, P.S. (2006) “A multi-market, historical comparison of the investment returns of value averaging, dollar cost averaging and random investment techniques.” *Academy of Accounting and Financial Studies Journal*, September.

Phalippou, L. (2008) “The Hazards of Using IRR to Measure Performance: The Case of Private Equity.” *Journal of Performance Measurement*, Fall.

Rieger, M. O. (2011) Co-monotonicity of optimal investments and the design of structured financial products, *Finance and Stochastics* 15, 27-55.

Rozeff, M.S. (1994) “Lump-sum Investing Versus Dollar-Averaging.” *Journal of Portfolio Management* 20, 45-50.

Statman, M. (1995) “A Behavioral Framework For Dollar-Cost Averaging.” *Journal of Portfolio Management* 22, 70-78.

Thorley, S.R. (1994) “The Fallacy of Dollar Cost Averaging.” *Financial Practice and Education* 4, 138-143.

Appendix: Continuous Time Analysis of VA's Inefficiency

This appendix uses the payoff distribution pricing model of Dybvig (1988a, 1988b) to derive the continuous time efficiency losses shown in Table 3.4. We assume an equity index (with zero dividends) which, relative to a constant interest rate bank account as numeraire, grows according to Geometric Brownian Motion as:

$$\frac{dS_t}{S_t} = \mu\,dt + \sigma\,dB_t \tag{A1}$$

This market offers a risk premium of μ and a Sharpe Ratio of μ/σ. We consider the degree of inefficiency for an investor who invests according to a fixed rule which determines the growth in the value $V_t = V_0\,g(t)$ invested in the equity market in each period from its initial V_0. Specifically in this case a value averaging strategy with target portfolio growth of α per period implies that V_t=$V_0e^{\alpha t}$. These amounts are also relative to the bank account as numeraire, so α = 0 corresponds to a value which grows at the interest rate. The investor's total wealth W_t grows according to:

$$dW_t = V_0 g(t)\left[\mu\,dt + \sigma\,dB_t\right]. \tag{A2}$$

We assume that the investor's initial wealth W_0 is sufficient to keep Vt on its target path, or that the investor can borrow enough for this purpose (indeed, we could set W_0=0 and assume that the strategy is entirely debt financed). These assumptions favour VA, since in practice no finite W_0 or credit line will ever be able to guarantee that adverse market outturns will not result in the VA strategy demanding more funds than the investor has available. This assumption implies that the distribution of terminal wealth at any later time T is normal with mean and variance given by:

$$E[W_t] = W_0 + V_0 \int_0^T \mu g(t)dt \tag{A3}$$

$$Var[W_t] = V_0^2 \int_0^T \sigma^2 g^2(t) dt \qquad \text{(A4)}$$

The normal distribution of these outturns is due to the fact that the equity market exposure follows a pre-determined target path, and does not depend on the returns made to date. This opens up the possibility of total losses exceeding the initial wealth W_0, as following earlier losses the strategy demands that the investor borrows to top the portfolio up to its required level (this is in contrast to the lognormal distribution of a buy-and-hold strategy). We now need to work out the cost of the cheapest way to buy a claim with this normal distribution. For fixed horizon *T* the market index evolves according to equation A5 (derived using the Ito integral):

$$S_T(u) = S_0 \exp\left\{(\mu - ½\sigma^2)T + \sigma\sqrt{T}u\right\} \qquad \text{(A5)}$$

where *u* is a standard normal variate. The pricing function for this economy is:

$$m(u) = \exp\left\{-½\left(\frac{\mu}{\sigma}\right)^2 T - \left(\frac{\mu}{\sigma}\right)\sqrt{T}u\right\} \qquad \text{(A6)}$$

This has expectation of one, and integrates with S_T to give E[$m(u)S_T(u)$]=S_0 or, scaling to a payoff equal to the normal variate *u*, E[$u\ m(u)$]= $\mu\sqrt{T}/\sigma$.

The exponential case

We now explicitly evaluate the minimum cost where $g(t)=e^{\alpha t}$. Substituting this into A3 shows that $E[W_T] = W_0 + M$, where $M = V_0 \int_0^T \mu e^{\alpha t} dt$

$$M = \begin{cases} V_0\mu[e^{\alpha T} - 1]/\alpha & \alpha \neq 0 \\ V_0\mu T & \alpha = 0 \end{cases} \qquad \text{(A7)}$$

$$Var[W_T] = S^2 = V_0^2 \int_0^T \sigma^2 e^{2\alpha t} dt$$

$$= \begin{cases} V_0^2\sigma^2[e^{2\alpha T} - 1]/2\alpha & \alpha \neq 0 \\ V_0^2\sigma^2 T & \alpha = 0 \end{cases} \tag{A8}$$

Dybvig (1988b) shows that the minimum cost of obtaining a specified set of terminal payoffs is given by the expected product of these payoffs with the corresponding state prices, where the payoffs and state prices are inversely ordered, so that the highest payoffs come in the lowest state price paths. Thus the minimum cost of obtaining the normally-distributed payoff W_0+M+Su is:

$$minimum\ cost = W_0 + M + S\mathrm{E}[um(u)] \tag{A9}$$

$$= W_0 + M - S\mu\sqrt{T}/\sigma \tag{A10}$$

This compares to the W_0 cost assumed for the VA strategy, so VA is inefficient by the magnitude $S\,\mu\sqrt{T}\,/\,\sigma - M$ which simplifies to:

$$V_0\mu\left\{\sqrt{\frac{T}{2\alpha}\left[e^{2\alpha T}-1\right]}-\left[e^{\alpha T}-1\right]/\,\alpha\right\}. \tag{A11}$$

Note that there is no inefficiency if μ or α are zero (implying that there is no opportunity cost to investing gradually), and the inefficiency is small if T is small. Furthermore, σ cancels out, so volatility plays no role in determining the size of the inefficiency. Intuitively, the inefficiency is also proportional to V_0 and the initial wealth W_0 plays no role at all.

Chapter 4

Measuring Investors' Historical Returns: Hindsight Bias In Dollar-Weighted Returns

Abstract

A growing number of studies use dollar-weighted returns as evidence that bad timing substantially reduces investor returns, and that consequently the equity risk premium must be considerably lower than previously thought. This paper demonstrates that this method is subject to hindsight bias (since prior returns influence levels of new investment) and derives a technique which removes this bias. The results show that for mainstream US equities dollar-weighted returns are low because of this hindsight bias - bad investor timing has very little effect. Thus low dollar-weighted returns should not lead us to adopt correspondingly low estimates of the risk premium.

JEL Classification: G11, G12, G15

I am grateful to Ilia Dichev, Stewart Hodges, Aneel Keswani, Gerhard Kling, Gilad Livne, Ian Marsh and Giorgio Questa for useful comments and to participants at the Midwest Finance Association 2012 annual meeting and a 2011 seminar at the University of Southampton. The usual disclaimer applies.

Measuring Investors' Historical Returns: Hindsight Bias In Dollar-Weighted Returns

1. Introduction

Few figures are of such central importance in finance as the equity risk premium, yet estimates vary widely. In particular, a growing number of studies argue that investors time their investments so badly that on average they earn returns which are significantly below the buy-and-hold return on the corresponding market index. They conclude from this that the risk premium must be substantially lower than had previously been thought.

This emerging literature stems from an influential paper by Dichev (2007), which argues that the impact of bad timing on aggregate investor returns can be deduced using a simple and elegant method. The geometric mean (GM) of monthly market returns gives the return that would be earned if investors followed a strict buy-and-hold strategy, immediately re-investing any dividends. This is contrasted with the dollar-weighted (DW) return (referred to in other contexts as the internal rate of return) which takes account of the net cashflows paid or received by the average investor ahead of the terminal period, such as share issues, dividend payments or share buybacks. The difference between these two rates is then used as a measure of the effect the timing of these cashflows has on investor returns.

Using this method Dichev concludes that poor timing has led to a substantial reduction in investor returns: a 1.3% per annum reduction for equities traded on NYSE and AMEX exchanges (1926 to 2002) and a 5.3% reduction for NASDAQ stocks (1973 to 2002), as shown in Table 4.1. This would imply that the equity risk premium earned by investors (and firms' cost of capital) must be considerably lower than previously estimated.

Table 4.1: Investor Timing Effects Identified by Previous Studies

This table shows the annualized Geometric Mean (GM) and Dollar-Weighted (DW) returns derived by previous studies for the markets and periods shown. A positive difference (final column) is interpreted as the reduction in the return received by investors as a result of bad timing. Distributions are defined as net cash distributions by firms to investors – a negative distribution represents an additional net investment (eg. as investors buy new share issues). The correlation coefficients in columns 4 and 5 are calculated on mean returns over the previous/subsequent three years (Dichev), and one year (Clare/Motson and Dichev/Yu).

Market	Period	Authors	Correlation of distributions with:		GM	DW	GM - DW
			Past returns	Future returns			
NYSE/AMEX	1926-2002	Dichev	-0.26	0.51	9.9%	8.6%	1.3%
NYSE/AMEX	1926-1964	Dichev			9.6%	8.0%	1.6%
NYSE/AMEX	1965-2002	Dichev			10.1%	9.4%	0.7%
NYSE/AMEX	1926-1951	Keswani/Stolin			7.5%	5.8%	1.8%
NYSE/AMEX	1951-1977	Keswani/Stolin			9.5%	9.7%	-0.2%
NYSE/AMEX	1977-2002	Keswani/Stolin			12.6%	12.9%	-0.3%
NASDAQ	1973-2002	Dichev	-0.57	0.28	9.6%	4.3%	5.3%
NASDAQ	1973-2006	Keswani/Stolin			10.4%	7.5%	2.9%
19 Internatn'l exchanges	1973-2004	Dichev	-0.24	0.16			1.5%
19 Internatn'l exchanges	1973-2004	Keswani/Stolin					0.7%
UK funds (all flows)	1992-2009	Clare/Motson	-0.18	-0.02	6.5%	5.7%	0.8%
UK funds (retail flows)	1992-2009	Clare/Motson	-0.37	0.09	6.5%	5.4%	1.2%
UK funds (institn'l flows)	1992-2009	Clare/Motson	-0.03	-0.08	6.5%	6.2%	0.3%
US mutual funds (all)	1991-2004	Friesen/Sapp					1.6%
Hedge funds (7190 funds)	1980-2006	Dichev/Yu	-0.22	0.04	10.0%	6.4%	3.6%

Keswani and Stolin (2008) show that the divergence between GM and DW returns for NYSE/AMEX stocks is sensitive to the exact start and end dates chosen. Dichev finds that the difference falls to 0.7% per annum in the second half of the period, but Keswani and Stolin find that it disappears entirely if the time series is split at different points. They also find that for NASDAQ stocks the difference shrinks substantially when four years' subsequent data are included, and that the difference recorded for 19 international stock exchanges was influenced by a dramatic increase in the proportion of stocks included in these indexes. However, studies using the same method have found GM-DW return differences in other markets which are similar to those reported by Dichev. Friesen and Sapp (2007) find a difference of 1.6% per annum for US mutual funds, Clare and Motson (2010) a difference of 0.8% for UK mutual funds, and Dichev and Yu (2011) a difference of 3.6% for hedge funds. A consensus has thus emerged that the aggregate effects of bad investor timing have been substantial.

These studies all use the difference between GM and DW returns to measure the impact of bad investment timing. Dichev (2007) notes that this difference can result from either (i) the correlation of investor cashflows with future asset returns, or (ii) the correlation of these cashflows with past asset returns. He interprets both of these as representing good/bad investor timing. This is clearly true of the correlation with future asset returns, but the following section demonstrates that the correlation of cashflows with past returns is very different: It alters the DW calculation retrospectively, and thus represents a hindsight effect which has no corresponding impact on investors' expected wealth. This effect should not be confused with genuine timing effects.

Section 3 derives a method for quantifying and removing this hindsight effect. The results show that for mainstream US equities (those traded on NYSE and AMEX) the great majority of the difference between DW and GM returns has been due to the hindsight effect, and very little has been due to bad investor timing. DW returns are low because aggregate investment flows reflect past returns rather than future returns.

Section 6 shows that the effect of bad investment timing in NASDAQ stocks is also much smaller than is suggested by the DW return. Furthermore, Table 4.1 shows that distributions in other markets are generally much more strongly correlated with previous returns than with future returns. This suggests that the relatively low level of the DW returns recorded for these markets is also likely to be largely due to the hindsight effect.

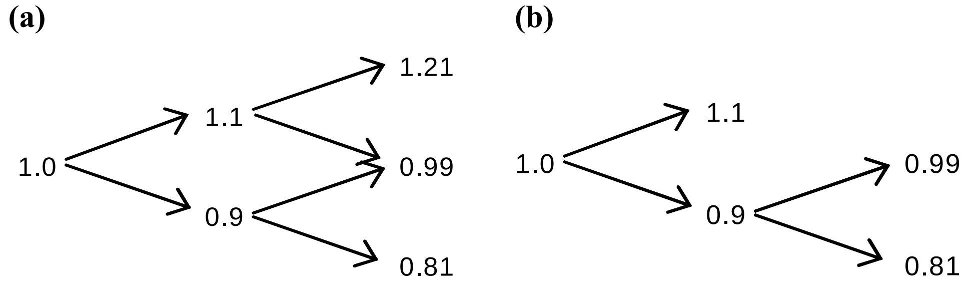
This contribution of the present paper is: (i) it helps to resolve the current debate about the equity risk premium by showing that low DW returns do not imply correspondingly low risk premia; (ii) it derives a new method which can separate the hindsight effect from genuine bad timing in any context in which DW returns are used. This method is likely to find applications in many other fields, since the DW return is still commonly used in project finance and investment management.

The structure of this paper is as follows: Section 2 demonstrates the hindsight effect in DW returns which can be mistaken for the effect of bad investor timing. Section 3 sets out a method for decomposing the difference between GM and DW returns into the hindsight effect and the genuine impact of investor timing. Subsequent sections apply this decomposition to data for NYSE/AMEX stocks (sections 4 and 5) and NASDAQ stocks (section 6). Conclusions are drawn in the final section.

2. Identifying the Hindsight Effect

The source of the hindsight effect can be illustrated with reference to the simple game illustrated below, in which the player faces a gain/loss of 10% in each of two rounds (Figure 4.1a). If we assume that the outturns in each round have a probability of exactly 50%, then the expected terminal wealth equals the initial stake.

Figure 4.1: Illustrative Game Showing the Quit-Whilst-Ahead Effect
Payoffs in a game which involves a gain or loss of 10% (with equal probability) in each of two rounds. In version (b) the player quits the game after a win in the first round. This improves some performance measures even though the expected terminal wealth remains equal to the initial stake in both (a) and (b).



The player may instead be able to quit the game following a win in the first round (Figure 4.1b). The expected terminal wealth for this truncated tree is still 1.0, but by quitting following the initial win the player can affect other performance measures.

For example, quitting early allows the player to claim a higher expected win rate. The full tree structure shows a 50% success rate. For example if we are trying to show “heads” in a fair coin toss then we have: HH, HT, TH, TT, giving success rates of 100%, 50%, 50%, 0%. But if the player quits if the first outcome is a head, then the tree shrinks to H, H, TH, TT, and the success rates shift to 100%, 100%, 50%, 0%, giving an impressive overall average of 62.5%. Quitting whilst ahead, and thus preserving a 100% winning record, affects this performance measure in a way which appears to suggest that the player has the ability to forecast the coin. The opposite incentive – to keep gambling when behind – can also be found. One simple example from outside the realm of finance is the child who agrees to toss a coin to settle an issue but, having lost, demands “best of three”.

Quitting whilst ahead also affects the expected Internal Rate of Return (IRR, the term which is generally used for the dollar-weighted return in investment management). Table 4.2 shows that the full tree shown in Figure 4.1a gives an average IRR of close to zero (fractionally negative due to the arithmetic/geometric mean inequality). This rises to 2.4% if the player quits after a win in the first round, since quitting locks in the early gains and gives the same IRR as if another win was guaranteed in the second round. This quit-whilst-ahead bias is similar to the familiar problem of the re-investment assumption used in calculating the yield to maturity on bonds.

Table 4.2: IRRs of Illustrative Two-Round Game

This table shows the cashflows and associated internal rates of return of the two games shown in Figure 4.1. An initial investment of 1 unit is assumed. The average IRR is the simple unweighted average of the IRRs calculated for the four scenarios.

	Lose-lose	Lose-win	Win-lose	Win-win	Avg.IRR
(a) Game Played over two periods					
0	-1	-1	-1	-1	
1	0	0	0	0	
2	0.81	0.99	0.99	1.21	
IRR	**-10.0%**	**-0.5%**	**-0.5%**	**10.0%**	**-0.25%**
(b) Player quits if ahead after round one					
0	-1	-1	-1	-1	
1	0	0	1.1	1.1	
2	0.81	0.99			
IRR	**-10.0%**	**-0.5%**	**10.0%**	**10.0%**	**2.4%**

Phalippou (2008) notes that private equity managers can boost their recorded IRRs by altering the time horizon of their investments – rapidly returning all cash to investors for successful projects and extending the life of projects which have performed poorly. I show below that IRRs can also be raised when the time horizon is fixed.

Ingersoll et al. (2007) show that investment managers can manipulate conventional performance measures by reducing risk exposure following a good performance and increasing exposure after a poor performance. The underlying strategy is to quit whilst ahead, but gamble more following poor outturns. They show that measures such as the Sharpe ratio and Jensen's alpha can be manipulated by this means, although they do not cover the IRR in their analysis. Individual investors have no corresponding incentive to alter the IRR recorded for their own savings, but the typical pattern of investor cashflows tends to introduce this effect accidentally. To demonstrate this we need to examine the reasons for this effect more formally.

Dichev (2007) derives net distributions from data for market returns and market capitalization using the clean surplus identity identified by Peasnell (1982). If in any period the market capitalization K_t is less than would have been suggested by applying the monthly rate of return r_t to the previous capitalization, then the difference must represent a net distribution d_t of funds to investors. A negative net distribution represents an additional investment (eg. as investors subscribe for a new share issue).

$$d_t = K_{t-1}(1 + r_t) - K_t \tag{1}$$

If we regard the market capitalization K_t as the aggregate portfolio value across all investors, then when we discount at the internal rate of return (r_{dw}), the present value of future cashflows and the final liquidation value by definition sum to the value of the initial investment:

$$K_0 = \sum_{t=1}^{T} \frac{d_t}{(1 + r_{dw})^t} + \frac{K_T}{(1 + r_{dw})^T} \tag{2}$$

As set out in Dichev and Yu (2011), substituting equation (1) into equation (2) eliminates the distributions and shows that the IRR can be considered to be a dollar-weighted average of the individual monthly returns. Specifically, the relative weight that this DW return puts on the market return in any month (r_t) is determined by the NPV of the assets that the investor holds in this market at the start of this period (discounted at the DW return):

$$r_{dw}\sum_{t=1}^{T}\frac{K_{t-1}}{(1+r_{dw})^{(t-1)}}=\sum_{t=1}^{T}\left(\frac{K_{t-1}}{(1+r_{dw})^{(t-1)}}\times r_t\right) \tag{3}$$

This formula shows how distributions and injections of additional funds re-weight the monthly returns r_t. The resulting shifts in the DW return may represent either a genuine effect on expected investor wealth or a hindsight effect in the calculation. To illustrate these different effects Table 4.3 shows the weights involved in calculating the DW return over a 10 period investment horizon. The first scenario shown in the table has no further cashflows after the initial investment. Returns are assumed to be independent and identically distributed (IID), so the ex ante expectation would then be that the ten returns will be given equal weight. We would expect the portfolio value to increase over time, but at a rate equal to the DW return, implying that the expected NPV of this portfolio would be equal in each period.

After the event there is likely to be some variation in these NPVs, due to volatility in $r_{t,}$ but to keep the illustration simple this effect is assumed to be small. If we invest a further amount, equal to the current portfolio value, after period 9, then the weight given to the period 10 return will be increased from 1/10 to 2/11. All earlier periods now have 1/11 weight, keeping the weights summing to unity. This is the second scenario illustrated in Table 4.3.

Table 4.3: Illustrative Effects of Net Distributions on Return Weights

Table 3 shows the expected weights given in the DW return calculation to the returns in each period of a ten-period investment horizon, given the cash injections/distributions shown in the first column. Four different cashflow profiles are considered. For simplicity returns are assumed IID, with low volatility.

	1	2	3	4	5	6	7	8	9	10
(1) No injections/distributions	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10
(2) Injection (=K_t) after period 9	1/11	1/11	1/11	1/11	1/11	1/11	1/11	1/11	1/11	2/11
(3) Injection (=K_t) after period 1	1/19	2/19	2/19	2/19	2/19	2/19	2/19	2/19	2/19	2/19
(4) Distrb'n (=K_t/2) after period 1	2/11	1/11	1/11	1/11	1/11	1/11	1/11	1/11	1/11	1/11

If instead a corresponding injection had been made after period 1, then the weight on subsequent periods would be raised only to 2/19 (scenario 3). An injection or withdrawal cannot have a substantial effect on the weights given to a large number of subsequent periods since this would raise the overall sum of the NPVs across all periods, with limited impact on the relative weights. But this injection has a substantial impact on the first period's weight, which falls from 1/10 to 1/19. Indeed, if we had instead distributed half the portfolio after period 1, halving the value of the remaining portfolio, then we would expect the first period return to be given twice the weight of each subsequent return (scenario 4).

Thus injections/distributions can affect the weights given to previous returns just as much as the weights given to future returns. For example, comparing scenarios 2 and 4 shows that we can just as easily boost the expected weight given to r_1 (by distributing half the portfolio after period 1) as the weight given to r_{10} (by doubling the size of the portfolio after period 9). In addition, comparing scenarios 2 and 3 shows that the effect of a given distribution/injection depends on its timing within the overall investment horizon – a result confirmed in the appendix using simulated data.

We can re-arrange equation (3) further to show the deviation of periodic returns from the DW return. Periodic returns r_t will be either above or below r_{dw}, but the weighted sum of these differences must be zero:

$$\sum_{t=1}^{T}\left(\frac{K_{t-1}}{(1+r_{dw})^{(t-1)}}\left(r_t - r_{dw}\right)\right) = 0 \tag{4}$$

This gives us a convenient form in which to consider the effect on the DW return of a distribution *d* (expressed as a percentage of portfolio value) at the end of period *m*:

$$\sum_{t=1}^{m}\left(\frac{K_{t-1}}{(1+r_{dw})^{(t-1)}}\left(r_t - r_{dw}\right)\right) + (1-d)\sum_{t=m+1}^{T}\left(\frac{K_{t-1}^{*}}{(1+r_{dw})^{(t-1)}}\left(r_t - r_{dw}\right)\right) = 0 \tag{5}$$

The distribution reduces the weight given to future returns in calculating the DW return, by reducing the future portfolio values to a fraction (*1-d*) of what they otherwise would have been (K_t*). A negative distribution (a further investment, for example as the result of a share issue) correspondingly increases future portfolio values. In the extreme, an investor could liquidate the entire portfolio (*d*=1). The DW return would then be calculated just on the returns up to period *m*, giving no weight to subsequent market returns.

Equation (5) shows that the two types of correlation that affect the GM-DW difference act in very different ways. A negative correlation between distributions and future returns would tend to boost the DW return, with negative distributions (injections) raising the start-of-period portfolio value ahead of periods of above-average returns, and positive distributions lowering it ahead of weaker returns. This would represent good investor timing. Unfortunately this correlation is generally positive (see Table 4.1), with investors tending to reduce their exposures ahead of periods of above-average returns and increase them ahead of poor returns.

The correlation of distributions with previous returns can also affect the DW return by retrospectively altering the relative weight given to earlier returns. This correlation tends to be negative (eg. with above-average[14] returns tending to be followed by negative distributions). This reduces the expected DW return by increasing the relative weight given to subsequent returns and correspondingly decreasing the weight given to these earlier strong returns.

The arithmetic appears similar for the backward-looking and forward-looking correlations, but these effects are very different. The forward-looking correlation works by altering investors' portfolio size ahead of unusually strong/weak returns. Thus the change in the weight given to these returns in the DW return calculation corresponds to a change in investors' exposure to these returns. By contrast, the correlation of distributions with past returns does not affect the portfolio value until after the relevant returns have already taken place – the relative weight given to these returns in the DW return calculation is adjusted retrospectively. A forward-looking correlation between distributions and future returns represents good/bad timing, and clearly affects investor welfare. The backward-looking correlation does not.

There is also an important distinction to be made in the information content of these different effects. Ingersoll et al. (2007) state an important principle: That a manipulation-proof performance measure must not reward information-free trading. The correlation of distributions with future returns clearly depends on trades which have a high information content, since they forecast future returns. The correlation of distributions with prior returns instead affects the DW return by means of trades which have a very low information content – all that is required is that the investor is sometimes

[14] The correlation coefficient is calculated using the arithmetic mean rather than the DW, but these two measures will of course be highly correlated.

able to judge that returns to date have been unusually high or low compared with likely returns in future. This is a much easier task.[15]

An investment manager with no forecasting ability can boost his recorded DW return by making large distributions following a lucky period of strong returns, thus giving less weight to subsequent returns and correspondingly more weight to the returns already recorded. This is a form of the quit-whilst-ahead effect discussed above. Conversely a negative distribution could be used to increase the relative weight given to future returns after disappointing returns to date.

For illustration we can consider a situation in which periodic returns are drawn from a distribution with a fixed mean μ. An investor with a negative forward-looking correlation will tend to invest more (negative distribution) ahead of periods where $r_t>\mu$. This investor will achieve higher returns over time because her forecasting ability means that her ex ante conditional expectation (conditioned on these forecasts) is greater than μ.

An investor with no forecasting ability is still able to boost his expected DW return by retrospectively re-weighting the returns in previous periods. This can boost the expected DW return, but not in any way which is likely to help meet his underlying investment objectives since his ex ante expected return in each period is still μ. The evidence in Table 4.1 suggests that distributions tend to show a significant negative correlation with past returns, reducing the DW return.

[15] It may be difficult to judge whether the return to date in any specific period differs significantly from the long-term mean, but investors can be opportunistic: If there is ever a period when the return to date has reached levels which are clearly different from any plausible estimate of the long-term mean then investors can use net distributions at this point to alter the DW return (for example, the cumulative return drops to well below zero in the early years of our NYSE/AMEX dataset – see Table 4.4). By contrast, forecasting future returns is always likely to be difficult.

Monte Carlo simulations confirm that this effect can spuriously affect the DW return. Friesen and Sapp (2007) show the results of simulations where returns are NIID. The ex ante expected return each period is identical, so any weighting of these ex ante returns must give the same average, regardless of the relative weights used. Volatility in ex post returns will drag the geometric mean below this arithmetic mean, but the simulations show that when distributions are negatively correlated with previous returns the DW return is significantly lower than the geometric mean. By construction, there is no correlation between distributions and future returns, so this reduction must be due to ex post re-weighting of past returns. The simulations presented in the appendix to the present paper confirm this result.

Chapter 3 shows that the same hindsight effect is also responsible for the superior returns claimed for value averaging (a formula investment strategy which requires investors to make regular periodic investments to keep their portfolio growing at a pre-specified target rate). This strategy builds in a strong correlation of periodic investments with prior returns, since a smaller (larger) additional investment is required following strong (weak) market returns, thus giving relatively less (more) weight to future returns, which are likely to be lower (higher). This gives rise to an IRR which is greater than the GM return even in simulated random walk data where the ex ante expected return in each period is constant by construction.

The following section sets out a method for decomposing the observed difference between the GM and DW returns into these two very different components.

3. A Method for Decomposing the Effects of Distributions

Equation (5) allows us to calculate the effect of each distribution on the DW return, but it cannot in itself distinguish between the forward-looking and backward-looking effects identified above. This is because in any historic dataset (i) the weights given to the periodic returns r_t sum to unity, and (ii) if the average return up to period *m* is, for example, below the whole-sample average, then it must subsequently be above this average. Thus a positive net distribution in period *m* has two simultaneous effects, which in this case both decrease the DW return: (a) a reduction in the weight given to later above-average returns (bad timing), and (b) a retrospective increase in the relative weight given to the lower returns seen earlier (hindsight effect).

At first sight these two effects may seem inextricably linked. However, they can be separated if we evaluate the effect of each distribution on the counterfactual assumption that all returns subsequent to this are constant at a neutral level which represents the average return. On this assumption a distribution cannot be well or badly timed, so any effect on the DW return must be due to the hindsight effect. We can then obtain the total hindsight effect by stepping through the historic data and summing the effect on the DW return of each successive distribution.

Section 5 shows that the results obtained using this method are robust to a wide range of different assumed levels for future returns, but setting future returns equal to the whole sample GM is the assumption which best isolates timing effects. On this assumption the DW return is initially equal to this GM return, but the DW return will shift when we substitute in the historical data to the extent that there is any systematic relationship between the timing of distributions and returns. By contrast, if we initially assume returns are equal to a figure other than the GM, then substituting in the actual historical data will alter the DW return even if there is no relationship between distributions and returns (or indeed, even if there are no distributions).

I start by assuming that the expected return in each period is equal to the whole-sample GM (9.9% for our NYSE/AMEX data) and that each distribution is zero. The DW return over the investment horizon as a whole will thus initially be equal to this GM. I then substitute in the historical value for the return in the first period (r_1), recalculate the DW return for the entire series, and record the amount by which this is different from our initial DW return estimate. Next I substitute in the historical distribution for that period, recalculate the DW return again and note how much this has changed from our previous estimate (which was based on r_1 and assumed values for all other data). This sequence reflects the assumption in equation (1) that distributions are made at the end of each month, after the return for the month is known. I then repeat this process for each subsequent period in turn, until all the initial assumptions made for r_t and d_t have been over-written with historical data.

Substituting in the actual distribution data can be interpreted as analogous to the process by which a cynical investment manager could alter the DW return. Each month, once the monthly return is known, the manager decides on the net distribution. Having no short-term forecasting ability, he assumes that all future returns are average, but if returns to date are significantly different from this assumed average, then a net distribution/injection can immediately increase the expected DW return by increasing the weight given to previous good returns or reducing the weight given to previous bad returns (a form of the quit-whilst-ahead strategy outlined above). The sum of the effects of each distribution on the DW return represents the total effect due to these distributions being related to past returns – in other words, the total hindsight effect.

If distributions turn out to anticipate future returns, then this will become apparent when these subsequent returns are substituted into the calculation. The effect of successive return data on the estimated DW return is likely to be noisy, but if there is no relationship between these returns and earlier distributions then our initial assumption (that these are equal to the whole horizon GM return) means that these effects will tend to cancel out over time. The cumulative effect will be positive only to the extent that previous net distributions resulted in relatively large start-of-period portfolio values (with correspondingly large NPVs) for periods when the returns were high, and relatively low NPVs ahead of periods when returns were low. This captures the effect of the good/bad timing of previous net distributions on the DW return. As discussed above, this is a genuine effect on investors' terminal wealth (rather than just a measurement issue) and will only come about if previous distributions contain information about future returns.

The assumed DW return will initially be equal to the GM return, but by recalculating it after each new piece of data is substituted in, it will gradually converge to the historic DW return (in *2T* steps as I substitute in *T* returns and *T* distribution figures). The aggregate effect of including the distribution data captures the total effect resulting from any relationship between these distributions and previous returns (hindsight effect) and the aggregate effect of the monthly return data reflects the degree to which these returns are related to previous distributions (the good/bad timing of these distributions). These two components sum to give the total difference between the GM and DW returns.

For illustration, Table 4.4 shows the impact of returns and distributions in the early years of our sample. Substitution of historic data is shown interrupted in 1931, so historical data is shown up to this date, whilst future returns are still set at the 9.9% average and future distributions at zero. In 1929 investors in aggregate invested additional cash equivalent to 9.8% of their existing portfolios. This subsequently turned out to be very bad timing given subsequent negative returns. However, even before any further return data were included, this cash injection had an immediate impact on the expected DW return by increasing the weight given to assumed future returns and reducing the weight given to the annualized return up to 1929, which was then well above 9.9%. This re-weighting resulted in the immediate -0.06% hindsight effect shown for 1929. Conversely, the large cash distribution in 1931 reduced the weight given to future assumed returns and boosted the weight given to the return to date, which by then was far below 9.9%. This gave rise to a hindsight effect, which immediately reduced the expected DW return. This illustrates how we can separate the forward-looking (good timing) and backward-looking (hindsight) effects of distributions on DW returns by stepping through the data with all future returns assumed to be average. The following sections apply this method to data for US equity returns.

Table 4.4: Impact of Distributions on the DW Return (NYSE/AMEX)

This table shows for early years of the NYSE/AMEX dataset the effect each year's return (timing effect) and distribution (hindsight effect) have on the expected DW return for the whole investment horizon (1926-2002). For illustration, the table interrupts the process with actual data substituted up to 1931, whilst future returns are still assumed constant at 9.9% per annum with no net distributions. For clarity the table shows annual data, but the underlying calculations use monthly data, so the precise effects depend on the timing of these distributions within each year.

	Annual return	Annualised return to date	Net distribution	Timing effect	Hindsight Effect
1926	9.7%	9.7%	-3.0%	0.00%	0.00%
1927	33.4%	21.0%	-3.1%	0.23%	-0.01%
1928	39.0%	26.7%	0.2%	0.28%	-0.01%
1929	-15.0%	14.7%	-9.8%	-0.30%	-0.06%
1930	-28.8%	4.2%	-5.3%	-0.51%	0.00%
1931	-44.4%	-6.1%	6.6%	-0.80%	-0.03%
1932	9.9%	.	0%	.	.
1933	9.9%	.	0%	.	.
.	.	.	.	.	.
2002	9.9%	.	0%	.	.

4. Decomposing the Effects on NYSE/AMEX Stocks

I initially use the same dataset as Dichev (NYSE and AMEX stocks January 1926 to December 2002), and the same method to infer net distributions from the capitalization and return figures. I then step through the entire dataset adding first the monthly return, then the monthly distribution, calculating the DW return after each piece of data is added. The GM is 9.86% and the DW 8.60% (1926 to 2002). However, the overall -1.26% per annum difference decomposes into an annualized -0.21% from adding the return data and -0.95% from adding the distribution data (Table 4.5). This shows that the large majority of the GM-DW difference is due to the hindsight effect, with only a limited effect from bad investor timing.[16]

Table 4.5: Decomposition of Timing and Hindsight Effects (NYSE/AMEX)

This table shows the cumulative effect on the DW return of substituting in (i) return data (timing effect), (ii) distributions (hindsight effect). The DW return is recalculated after each successive substitution. By construction, for monthly data DW return = GM return + timing effect + hindsight effect, although this summation is only approximate for the annualized returns shown here.

	GM return	Timing effect	Hindsight effect	DW return
Jan. 1926 - Dec. 2002	9.86%	-0.21%	-0.95%	8.60%
Jan. 1926 - Dec. 2011	9.58%	-0.08%	-0.88%	8.54%

This method allows us to identify the impact of each new data point as I step through the data. Figure 4.2 presents the annualized return to date and the annual net distribution as a proportion of the implied portfolio value at the time. Figure 4.3 shows the corresponding incremental effects on the DW return resulting from adding successive data for returns (the investor timing effect) and distributions (the hindsight effect). The

[16] Some papers have found bad investor timing using data on equity issues (e.g. Ritter, 1991, Loughran and Ritter, 1995), although others question these results (e.g. Brav and Gompers, 1997 and Schultz, 2003). The present paper does not revisit this well-established debate: Instead it shows that whatever their statistical significance, the economic significance of these bad timing effects is far smaller than is suggested by the studies detailed in Table 4.1.

timing effect is noisy, but small in aggregate, whereas the hindsight effect is consistently negative in the early part of the period. It is also reassuring that Figure 4.3 shows no massive monthly jumps. This suggests that the DW return calculation has found the same underlying root to the IRR polynomial in every calculation, with only modest changes due to the new data. If it ever jumped between multiple solutions then we might expect to see a far larger monthly move.

Figure 4.2: Returns to Date and Net Distributions (NYSE/AMEX stocks)

The line shows the annualized cumulative return to date from the start of the dataset in January 1926. The bars show net distributions as a percentage of the implied total market capitalization before the distribution. A positive distribution is a return of cash to investors, a negative distribution is a net investment.

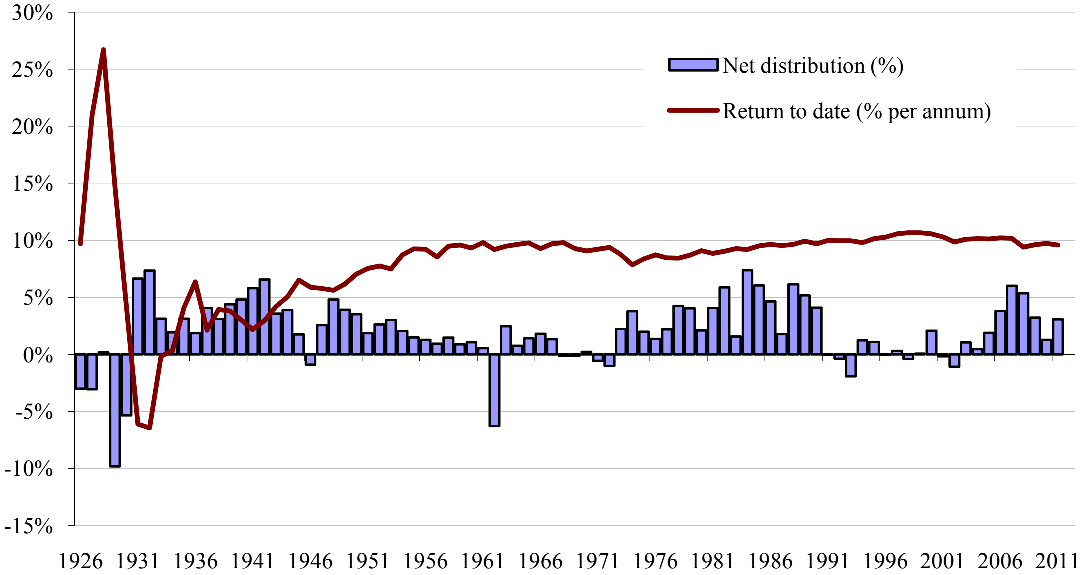
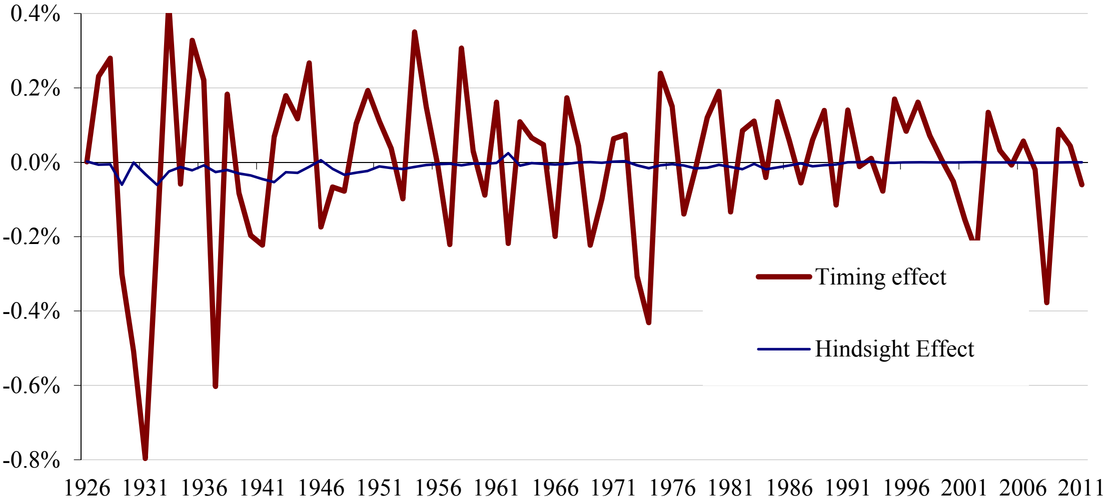


Figure 4.3: Timing & Hindsight Effects in DW Returns (NYSE/AMEX Stocks)

This figure shows the change in the expected DW return over the whole horizon (Jan. 1926 – Dec. 2011) resulting from substituting in (i) the monthly market return ('timing effect', the volatile bold line), and then (ii) the monthly aggregate net distribution ('hindsight effect', the thinner, more stable line). The DW return is calculated on the initial assumption that future returns are equal to the geometric mean (9.9% per annum) and that future distributions are zero. The underlying calculations use monthly data, but annual effects are shown here for clarity.



The large cash distributions in 1931 and 1932 reduced the weight given to future returns and boosted the weight given to returns to date, which by then were far below average. This gave rise to an immediate hindsight effect which reduced the expected DW return. The incremental hindsight effect remained negative in the late 1930s and early 1940s as consistently large distributions were made whilst the return to date was below 5%. Distributions in later years generally had little impact since by this stage the return to date had inevitably converged towards the overall average.

Extending the dataset diminishes the estimated timing effect from -0.21% per annum (1926 to 2002) to -0.08% (1926 to 2011). The sensitivity analysis in the following section shows that the difference is due to the very unusual long term uptrend in returns over the 1926-2002 period. More recent poor returns offset this effect, so the much smaller timing effect shown for the extended data series should be regarded as the better estimate.

5. Sensitivity Analysis

In this section I examine the robustness of these results as we relax the initial assumptions made above: (i) that future returns are equal to the GM of 9.9% per annum; (ii) that future distributions are zero. Our assumption that distributions are made at the end of each month makes minimal difference: When we assume instead that they are made at the start of each month, the decomposed effects differ by less than 0.01%.

Table 4.6 sets out the timing and hindsight effects derived using a wide range of assumptions for future returns. The timing effect calculated on such counterfactual assumptions is relatively uninformative: Assuming returns which are well below the historical mean naturally leads to a more positive timing effect as returns subsequently tend to be higher than this (high assumed returns lead to correspondingly negative return surprises). Our interest is instead in the hindsight effect. The table covers a huge range of assumed average returns, but this paper's key finding is robust, since for any plausible figure in the middle part of this range the hindsight effect is clearly substantial and negative, and accounts for a large part of the -1.3% historical GM-DW difference.

Table 4.6: Decomposition on Alternative Return Assumptions

This table shows the cumulative effect of monthly returns (timing effect) and distributions (hindsight effect) on the expected DW return for NYSE/AMEX stocks 1926-2002. Future returns are initially assumed constant at the levels shown in the first column. By construction, for monthly returns the DW return = assumed GM + timing effect + hindsight effect, although this summation is only approximate for the annualized returns shown here.

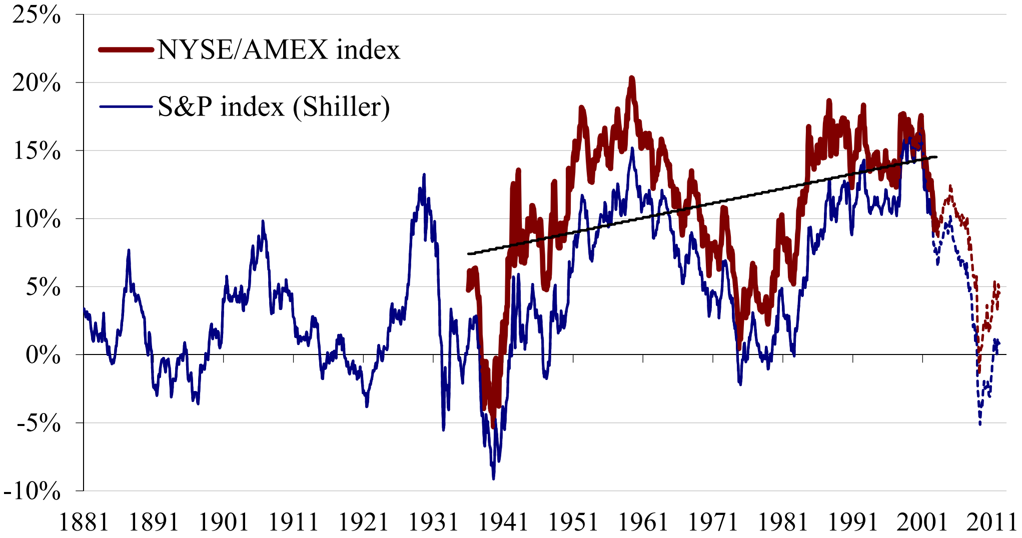
Assumed GM	Timing effect	Hindsight Effect	DW return
5%	3.49%	-0.05%	8.60%
6%	2.76%	-0.29%	8.60%
7%	2.00%	-0.50%	8.60%
8%	1.24%	-0.68%	8.60%
9%	0.46%	-0.83%	8.60%
10%	-0.32%	-0.97%	8.60%
11%	-1.10%	-1.09%	8.60%
12%	-1.89%	-1.20%	8.60%
13%	-2.68%	-1.29%	8.60%
14%	-3.47%	-1.37%	8.60%
15%	-4.25%	-1.44%	8.60%

Moreover, we do not need to interpret these assumptions as reflecting investor expectations. Considering how a cynical investment manager could attempt to raise his DW return helped lead us to the decomposition set out above, but this should be seen as just an analogy. As discussed in section 3, the key requirement is that assumed future returns are constant, with no relationship with past returns or distributions. This ensures that any relationship between distributions and future returns is captured in the "timing effect" column as the subsequent return data is substituted in, whilst any backward-looking relationship between distributions and previous returns is captured in the "hindsight effect" column. This holds regardless of whether the return assumption actually reflects investor expectations. The key advantage of setting the assumed future returns equal to the historical geometric mean is that this removes the effect of consistent return surprises in either direction, leaving only the pure effect of the timing of investment flows compared to periods of above/below average return.

I also investigate the effect of changing the initial assumption for future distributions (set to zero for all periods in the decompositions above). Setting each instead to 0.081% of market capitalization (giving an average distribution equal to that in the historic sample) substantially alters the decomposition, with the aggregate hindsight effect shifting to -1.16% and the timing effect almost vanishing (-0.005%). The reason can be seen in Figure 4.4, which shows that returns on NYSE/AMEX stocks (cumulated over 10 year periods to reduce short-term noise) trended upwards over our sample period of January 1926 to December 2002. Given this trend, any early distribution would appear to be bad timing compared to our initial assumption that future distributions were zero.

Figure 4.4: Long-Term Equity Returns

This figure shows the annualized 10 year geometric mean market return up to the date shown. A linear time trend has been added, fitted to NYSE/AMEX returns Jan. 1926-Dec. 2002. The longer time series shows the S&P index return (source: Shiller, 2005). Data after Dec. 2002 is shown as dotted lines.



However we should not accept the effects of this trend at face value. First, bad investment timing is generally interpreted as a short-term cyclical effect as investors chase returns during booms. Apparent bad timing caused by this very long term trend is a very different effect. As most investors have horizons which are substantially shorter than this 77 year time series, they had no realistic option to time their investments better. Moreover it is entirely implausible to suppose that this upward trend will continue in future – this would imply that expected equity returns are currently over 15% per annum and will continue to rise by almost 1% per decade.

Figure 4.4 also shows the similar S&P index returns going back to 1871 and forward to 2011. This shows that the 1926 to 2002 period was almost unique in showing such a sustained uptrend. Almost any other period of equal length would have given us very different results. Thus if we are to obtain results that can be plausibly applied to the future (e.g. in estimating the expected risk premium) we need to strip out the effects of this trend before decomposing the residual into the timing and hindsight effects. This can be achieved either by de-trending the return series or de-meaning the distribution. For robustness I do both, on a variety of different assumptions, both individually and combined.

Table 4.7 presents the results using four alternative treatments of the distribution data. We have already seen the first two, which use unadjusted historical distribution data. The third and fourth variants de-mean the distribution data by subtracting a percentage of portfolio value such that (a) the average dollar distribution is zero, (b) the average distribution as a percentage of market capitalization is zero. These distributions are all set to zero ex ante and then replaced sequentially by the adjusted historical data. The last four variants repeat the first four, but with return data from which a log-linear trend has been extracted (thus keeping the whole-horizon GM return unchanged).

Table 4.7: Decomposition with Alternative Corrections for Trend in Returns (% pa.)

This table shows the cumulative effect of monthly return and distribution data on the DW return for NYSE/AMEX stocks 1926-2002 using a range of measures to correct for the long-term uptrend in returns and the positive mean distribution. The first two use raw distribution data with an initial assumption that future distributions are (1) zero, (2) the fixed percentage of implied market capitalization which gives the average dollar distribution seen in the historical data. The following decompositions use distribution data which has been de-meaned by subtracting a percentage of market capitalization such that (3) the average dollar distribution is zero, (4) the average percentage of market capitalization which is distributed is zero. The last four variants repeat the first four, but with return data from which a log-linear trend has been removed.

Return Data	Distribution data and starting assumption	Timing effect	Hindsight Effect	DW return
1. Unadjusted r_t	Unadjusted d_t (d_t initially set at zero)	-0.21%	-0.95%	8.60%
2. Unadjusted r_t	Unadjusted d_t (d_t initially fixed % of capitalization)	0.00%	-1.16%	8.60%
3. Unadjusted r_t	De-meaned d_t (average d_t* zero)	-0.11%	-0.53%	9.17%
4. Unadjusted r_t	De-meaned d_t (average d_t*=0% of capitalization)	-0.07%	-0.33%	9.43%
5. Detrended r_t	Unadjusted d_t (d_t initially set at zero)	-0.03%	-0.49%	9.30%
6. Detrended r_t	Unadjusted d_t (d_t initially fixed % of capitalization)	0.06%	-0.58%	9.30%
7. Detrended r_t	De-meaned d_t (average d_t* zero)	-0.05%	-0.36%	9.42%
8. Detrended r_t	De-meaned d_t (average d_t*=0% of capitalization)	-0.06%	-0.28%	9.49%

The first two variants use unadjusted historical data, thus ending up with the historical DW return of 8.6%. By contrast the other variants (3 to 8) adjust the historic data to remove the effect of the long-term trend in returns. These new variants all give substantially higher final DW returns but, reassuringly, these lie within a limited range 9.17% to 9.49% Thus all these methods suggest that this long-term uptrend accounted for a substantial part of the raw difference between the GM and DW returns. Even before decomposing the residual into hindsight and timing effects it is clear that these two together have much less effect once we strip out the long-term uptrend in returns. Moreover, decomposing the remaining difference shows that the timing effect is very small in all cases, ranging from -0.11% to +0.06%.

The timing effect also collapses to near zero when I extend the dataset to 2011 (Table 4.5). This is because the poor returns in the years 2003-2011 largely removed the long-term trend. However, it is important that this sensitivity analysis has established that it is the shorter period 1926-2002 which is unrepresentative. The much smaller timing effect recorded for the extended dataset (-0.08% per annum) corresponds closely to the figures obtained by stripping out the effects of the long-term uptrend in the earlier returns data, and these should be regarded as the most representative estimates of the effect of bad timing on investor returns.

In conclusion, after adjusting for (i) the unsustainable uptrend in returns, and (ii) the hindsight effect in the DW return, we find that bad timing actually had only a very small impact on the return received by investors. Thus, in contrast to the claims made elsewhere, bad investor timing does not justify reducing our estimates of the equity risk premium.

6. Decomposing the Return Difference for NASDAQ Stocks

NASDAQ stocks show a much larger difference than NYSE/AMEX stocks, with a GM of 9.6%, but a DW return of only 4.2% (January 1973 to December 2002). When I decompose this difference using the method set out above, we see that the large majority (-4.0%) is due to bad investor timing, with only -1.0% due to the hindsight effect (see Table 4.8).

Table 4.8: Decomposition of Timing and Hindsight Effects (NASDAQ)

This table shows the cumulative effect of monthly returns (timing effect) and distributions (hindsight effect) on the expected DW return for NASDAQ stocks. By construction, for monthly returns DW return = GM return + timing effect + hindsight effect, although this summation is only approximate for the annualized returns shown here.

	GM return	Timing effect	Hindsight effect	DW return
Jan. 1973 - Dec. 2002	9.6%	-4.0%	-1.0%	4.2%
Jan. 1973 - Dec. 2011	9.4%	-1.8%	-1.0%	6.3%

The main effect comes from investors' terrible timing during the dot-com boom. Additional funds equivalent to 8.5% of market capitalization were invested in 1999 and 13.7% in 2000, just ahead of the crash (see Figure 4.5). Naturally, the recorded effect of bad timing rises as we increase our assumption for future returns, but it is reassuring that the hindsight effect remains fairly stable (see Table 4.9). Moreover, the decomposition shows very little sensitivity to the assumed level of future distributions (consistent with the absence of any long-term trend in returns). Thus our estimates of the hindsight effect are robust to shifts in both these assumptions.

Figure 4.5: Distributions and Annual Returns (NASDAQ Stocks)

The line shows the annual return on NASDAQ stocks (Jan. 1973 – Dec. 2011). The bars show annual net distributions as a percentage of the market capitalization just before the distribution ($K_{t-1}(1+r_t)$). A positive distribution is a return of cash to investors, a negative distribution a net investment.

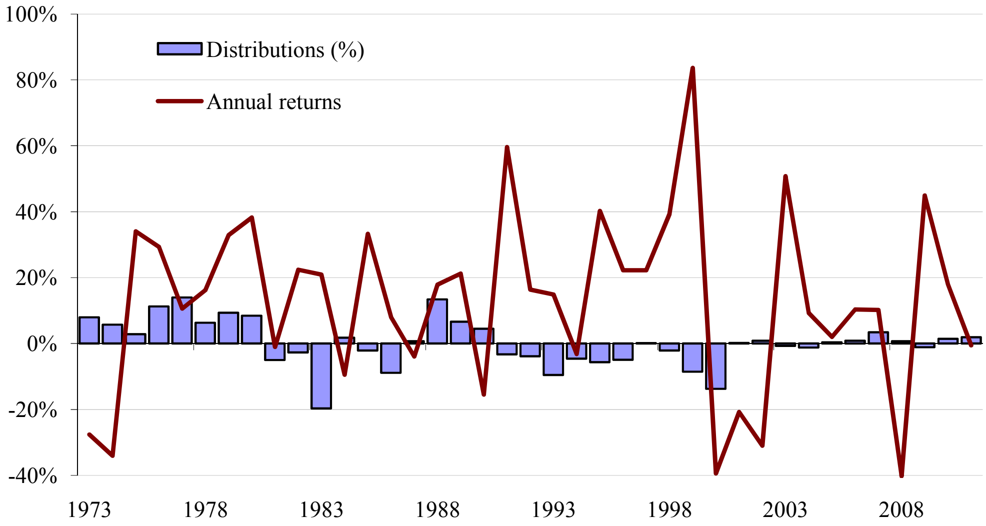


Table 4.9: NASDAQ Return Decomposition: Alternative Assumptions
This table shows how cumulative timing and hindsight effects vary as we alter our assumption for future returns. Returns are initially assumed constant at the levels shown in the first column before historical returns are substituted in. For monthly returns DW return = assumed GM return + timing effect + hindsight effect, but this summation is only approximate for the annualized returns shown. Coverage: NASDAQ stocks 1973-2002.

Assumed GM	Timing effect	Hindsight effect	DW return
5%	0.63%	-1.34%	4.25%
6%	-0.41%	-1.26%	4.25%
7%	-1.42%	-1.19%	4.25%
8%	-2.40%	-1.12%	4.25%
9%	-3.37%	-1.06%	4.25%
10%	-4.31%	-1.00%	4.25%
11%	-5.24%	-0.94%	4.25%
12%	-6.15%	-0.89%	4.25%
13%	-7.04%	-0.84%	4.25%
14%	-7.91%	-0.80%	4.25%
15%	-8.76%	-0.75%	4.25%

As noted by Keswani and Stolin (2008), the difference between the GM and DW returns shrinks markedly if the dataset is extended beyond 2002. Table 4.8 shows that the cumulative timing effect changes from -4.0% (1926 - 2002) to only -1.8% (1926 - 2011). One reason for this is the positive returns seen after 2002, but the timing effect would have diminished even if the additional data were unexceptional. As shown in Section 2 (and is confirmed by the simulations in the appendix), for any given relationship between distributions and subsequent returns, timing effects are far more powerful for distributions close to the end of the investment horizon, since they then have a large effect on the weights given to subsequent returns in the DW return calculation. The same pattern of distributions and subsequent returns would tend to have much less impact further from the end of the horizon, since the distributions would then affect start-of-month portfolio values over a larger number of subsequent periods, implying less impact on their relative weights in the DW return calculation. Using artificial data for these extra years (returns set equal to the GM up to 2002 (9.6%) and distributions set to zero) reduces the aggregate timing effect to only -1.7% (from -4.0% for 1973-2002). This confirms that the reduction is due to the extension of the dataset, rather than being specific to the subsequent historic outturns.

The historical data 1973 to 2011 still show a -1.8% effect from bad investor timing, but we face two problems in assuming that bad timing will continue to have such an effect in future. First, this effect stems from what should be seen as a single massive event – the dot-com crash – so we must question whether this is statistically significant. Second, we must expect the measured bad timing effect to shrink further as more data is added, pushing the 2000 to 2002 crash further away from the end of the investment horizon.

7. Conclusion

A growing number of papers use the difference between the dollar-weighted (DW) return and the geometric mean (GM) return as a measure of the effect of bad investment timing. They generally find that poor timing has reduced the return actually received by investors to well below the buy-and-hold return on the assets concerned. As a result they conclude that estimates of the equity risk premium need to be revised down substantially. Given the central role that this figure plays in finance, this would have profound implications. However, the present paper finds that the DW return will be affected if net investor cashflows are related to either future asset returns (which would clearly affect investors' expected wealth) or past asset returns (which would represent a hindsight effect in the DW return).

This paper derives a method which separates these two effects. This shows that bad investment timing accounts for very little of the overall difference between the GM and DW returns for mainstream US equities (those traded on the NYSE and AMEX exchanges). The great majority is just due to the hindsight effect. Thus low DW returns should not lead us to adopt correspondingly low figures for the equity risk premium.

The method derived here clearly has applications in other fields where DW returns are used, notably project finance and investment management. Good timing by investment managers should clearly be separated from the impact of the hindsight effect (whether deliberate or accidental). For this purpose a hindsight-corrected DW return can be derived by adding the timing effect to the GM return or, equivalently, subtracting the hindsight effect from the DW return:

$$R_H \textit{ (hindsight-corrected DW return)} = \textit{GM return} + \textit{timing effect} \quad (6)$$

$$= \textit{DW return} - \textit{hindsight effect} \quad (7)$$

More specifically, future research could investigate the degree to which funds have benefited from the hindsight effect. This could have come about if funds tend to choose between alternative cashflow options by comparing the projected IRRs, or if funds which benefit from this effect by luck tend to have higher survival rates. As we saw above, the effects can be substantial even for broad equity market indices. They could be much larger for individual funds since these may have considerably greater volatility in both their returns and their cashflows.

References

Brav, A., and P. A. Gompers (1997) "Myth or Reality? The Long-Run Underperformance of Initial Public Offerings: Evidence from Venture and Non-Venture Capital-Backed Companies." *Journal of Finance,* 52, 1791-1821.

Clare, A., and N. Motson (2010) "Do UK Investors Buy at the Top and Sell at the Bottom?" *Cass Business School Working Paper.*

Dichev, I. D. (2007) "What Are Stock Investors' Actual Historical Returns? Evidence from Dollar-Weighted Returns." *American Economic Review*, 97, 386-401.

Dichev, I. D., and G. Yu. (2011) "Higher Risk, Lower Returns: What Hedge Fund Investors Really Earn." *Journal of Financial Economics,* 100, 248-263.

Friesen, G. C., and T. R. A. Sapp. (2007) "Mutual Fund Flows and Investor Returns: An Empirical Examination of Fund Investor Timing Ability." *Journal of Banking and Finance,* 31, 2796-2816.

Gompers, P. A., and J. Lerner. (2003) "The Really Long-Run Performance of Initial Public Offerings: The pre-NASDAQ Evidence." *Journal of Finance*, 58, 1355–1392.

Ingersoll, J.; M. Spiegel; W. Goetzmann and I. Welch. (2007) "Portfolio Performance Manipulation and Manipulation-Proof Performance Measures." *Review of Financial Studies,* 20,1503-1546.

Jensen, M. C. (1968) "The Performance of Mutual Funds in the Period 1945-1964." *Journal of Finance* 23, 389-416.

Keswani, A., and D. Stolin. (2008) "Dollar-Weighted Returns to Stock Investors: A New Look at the Evidence." *Finance Research Letters*, 5, 228-235.

Loughran, T., and J. R. Ritter. (1995) "The New Issues Puzzle." *Journal of Finance*, 50, 23-51.

Peasnell, K. V. (1982) "Some Formal Connections between Economic Values and Yields and Accounting Numbers." *Journal of Business Finance & Accounting,* 9, 361–381.

Phalippou, L. (2008) “The Hazards of Using IRR to Measure Performance: The Case of Private Equity.” *Journal of Performance Measurement*, 12, 55-66.

Ritter, J. (1991) “The Long-Run Performance of Initial Public Offerings.” *Journal of Finance*, 46, 3-27.

Schultz, P. (2003) “Pseudo Market Timing and the Long-Run Underperformance of IPOs.” *Journal of Finance*, 63, 483-517.

Sharpe, W. F. (1966) "Mutual Fund Performance." *Journal of Business*, January, 119-138.

Shiller, R. J. (2005) *Irrational Exuberance*. Princeton University Press, 2nd edition, and www.irrationalexuberance.com.

Zweig, J. (1997) “Funds That Really Make Money for Their Investors.” *Money Magazine*, 1st April 1997.

Appendix: Simulation Evidence

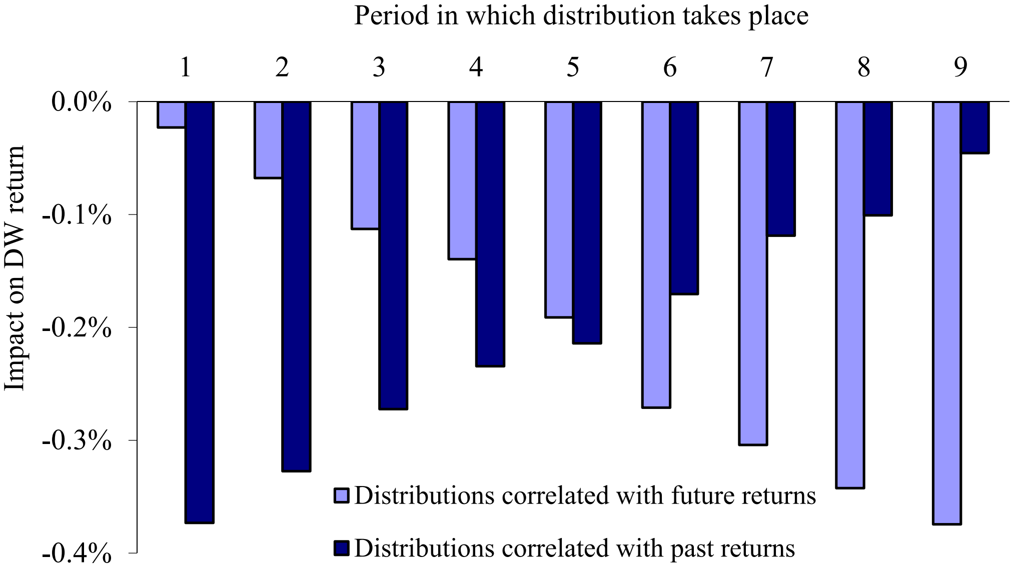
This appendix uses Monte-Carlo simulations to show that the correlation of distributions with prior returns can shift the average DW return away from the geometric mean even when the ex ante expected return in each period is identical by construction. I also confirm that the size of this effect depends on the timing of the cashflows within the investment horizon.

Returns are generated over an investment horizon comprising ten periods. These returns are normally, identically, and independently distributed (NIID), and for convenience the mean is set so that the GM return over the ten periods averages zero. We then consider the impact that a single net distribution after each of periods 1-9 has on the DW return (all assets are assumed to be liquidated in the tenth period). Net distributions are either (a) negatively correlated with the previous return (as investors chase returns by investing more following strong returns), or (b) positively correlated with the return in the following period (bad investor timing). Note that these are the signs of the correlations generally found in the empirical studies shown in Table 4.1.

Figure 4.6 shows that each of these correlations pulls the average DW return below the GM return (the opposite correlations – not shown – have a positive effect). The forward-looking correlation reduces the conditional expected return in the period following the distribution (conditioned on the amount which remains invested). By contrast, where the distributions are only correlated with past returns the ex ante expected return in each period is identical by construction, so we should regard the shift in the DW return as a hindsight effect produced by the retrospective shifts in the weights given to past returns.

Figure 4.6: Impact on DW Returns of Correlation between Distributions and Returns

The chart shows the effect on the DW return of correlation between distributions and returns in different periods. Each simulated path is of ten periods, with returns in each period NIID with standard deviation 20% and geometric mean zero. A single distribution is included for each path in the period shown. This distribution is set as a percentage of portfolio value which is (i) the previous period's percentage return multiplied by -1, or (ii) identical to the following period's percentage return. Net distributions in all other periods are zero for each path. The impact of these cashflows in pulling the DW return below the GM return is calculated for each of 5000 simulated return paths for each of the distribution patterns shown. The GM is, of course, unaffected by these cashflows, so the difference represents the negative impact of these cashflows on the DW return.



The size of these effects depends on when each distribution comes within the investment horizon, but the average size of the effect across all periods is roughly the same for the forward-looking and backward-looking correlations. This confirms the underlying symmetry apparent in equation (5): That a distribution can in principle affect the DW return just as effectively by re-weighting either past returns or future returns.

A distribution which is correlated with returns in the coming period has most effect in period 9, since it can then have a substantial effect on the NPV of the portfolio value at the start of period 10, and hence on the relative weight given to this return in the DW return calculation. By contrast, a similar distribution after period 1 alters the portfolio value in all future periods, and so has little effect on their relative weights. But such early distributions strongly affect the relative weight given to previous returns. Thus a forward-looking correlation has more impact near the end of the investment horizon and the backward-looking correlation has more impact near the beginning.

Decomposing the effects on the DW return for NYSE/AMEX stocks (Section 4) shows that the major impact comes from the correlation of distributions with prior returns early in the investment period. The same pattern of distributions and returns would have had less impact on the DW return if our sample had started earlier. This explains why the overall differential between the GM and DW returns is very sensitive to the exact start and end dates chosen for the analysis (as was observed by Keswani and Stolin, 2008).

For NASDAQ stocks we found instead that the major impact is the bad timing of the large net investments made at the height of the dot-com boom – ahead of the subsequent bust. But, again, the fact that these flows took place very near the end of the initial 1973-2002 investment horizon gave them the maximum effect on the DW return. We found that the effect is reduced as more recent data is added, pushing these large distributions away from the end of the horizon, and we should expect further reductions as subsequent data is added.

Chapter 5

Diversification Returns, Rebalancing Returns and Volatility Pumping

Abstract

It is widely claimed by academics and practitioners that periodic rebalancing of portfolios to keep asset weights constant will directly boost geometric returns by buying on downticks and selling on upticks. This paper refutes this claim by showing that comparable improvements arise even without rebalancing. Volatility pumping is a strategy which appears to benefit from high asset volatility and large rebalancing trades. We show that the real source of return on such strategies is an implied risk premium on these assets, and that volatility unambiguously reduces the expected geometric return.

JEL Classification: G10, G11

First I would like to acknowledge my co-authors for this chapter: Keith Cuthbertson and Nick Motson. I am also grateful to Max Bruche, Ales Cerny, Michael Dempster, Aneel Keswani, Ian Marsh, Frank McGroarty, Anthony Neuberger, Richard Payne, Scott Willenbrock, anonymous referees and the participants at the seminar at Cass Business School on 10 July 2013 for useful comments. The usual disclaimer applies.

Diversification Returns, Rebalancing Returns and Volatility Pumping

1. Introduction

Rebalancing is an important part of many investment strategies. Investors with target asset allocations need to rebalance when their actual exposures diverge significantly from desired levels, and strategies such as volatility pumping deliberately aim to profit from such rebalancing. It is widely claimed that "rebalancing returns" add significantly to overall expected returns because the rebalancing trades consistently buy on downticks and sell on upticks.

This paper rejects this claim. Instead we find that in normal circumstances the apparent impact of rebalancing on portfolio geometric returns is entirely because a rebalanced portfolio tends to remain better diversified, with lower volatility and hence lower "volatility drag". Only if asset returns exhibit negative autocorrelation does rebalancing have any direct effect on portfolio returns, since this autocorrelation makes it profitable to sell previous winners and buy previous losers. This has important implications for investors.

Confusion on this issue appears to have arisen because of the difficulty in making meaningful comparisons between rebalanced and unrebalanced portfolios, since even when the portfolios are initially identical the composition of an unrebalanced portfolio tends to shift over time. We derive like-for-like comparisons between rebalanced and unrebalanced portfolios which show that in the absence of autocorrelation of returns the difference is entirely explained by volatility drag.

We examine in detail the popular rebalanced strategies: (i) "volatility pumping" with a portfolio comprising one risky asset and risk-free deposits, (ii) the more general multi-asset rebalanced strategies put forward by Fernholz and Shay (1982). We show that the returns on these strategies too are entirely explained by volatility drag, with no evidence of the buy-low-and-sell-high effects that proponents claim. This misleading claim encourages investors to hold volatile assets so as to increase the scale of the rebalancing trades which are claimed to generate profits. Volatility pumping is likely to be systematically inefficient, with sub-optimal asset allocations and unnecessarily high trading costs. More efficient portfolios can be constructed simply by diversifying effectively and thus minimizing volatility drag.

2. Rebalancing Return Versus Diversification Return

A key objective of this paper is to compare "rebalancing returns" with two other effects: “diversification returns" and “volatility drag". These effects all relate to the geometric mean (GM) returns achieved by a portfolio. Use of the GM is intuitively attractive, since it is more closely related to investors’ terminal wealth than the more frequently used arithmetic mean (AM).[17]

Volatility drag will be familiar to many investors, and can be understood directly from the standard relationship between the AM and GM[18]:

[17] Using the GM as a target raises a number of questions. For example, maximising the GM will maximise the welfare of an investor whose utility is a logarithmic function of terminal wealth. It is less clear that it is an appropriate target for investors with other utility functions. This topic has previously been the subject of a long and rancorous debate, which we do not wish to revisit here. For our purposes it is sufficient to note that investors are encouraged to choose strategies such as volatility pumping or rebalanced strategies on the basis of their expected GM returns. This paper seeks to clarify how these apparently attractive GM returns come about.

[18] Booth and Fama (1992) derive this relationship using a Taylor expansion for the expected continuously compounded return E[log(1+r)] around E[r]. This derivation makes no assumption

$$E[GM] \approx E[AM] - \frac{\sigma^2}{2} \tag{1}$$

This tells us that, all else equal, an asset or portfolio will generate a lower expected GM if it has a higher volatility. This is because the GM is a concave function of terminal wealth (GM=$^1/_T$log(Terminal wealth/Initial wealth)[19]). If we compare two strategies with identical expected terminal wealth (TW), the one with lower volatility will generate a higher E[GM] because this concave relationship effectively penalizes both exceptionally high and exceptionally low terminal wealth outturns. Volatility drag is thus inherent in the compounding relationship between periodic returns and terminal wealth.

Equation (1) holds for any asset or portfolio. Here we apply it to two portfolios: a diversified portfolio *p* and a portfolio containing a single asset *i*. Subtracting one of the resulting equations from the other gives us:

$$E[GM_p] - E[GM_i] \approx (E[AM_p] - \tfrac{1}{2}\sigma_p^2) - (E[AM_i] - \tfrac{1}{2}\sigma_i^2) \tag{2}$$

For simplicity we assume that all these assets are independently and identically distributed. Without this assumption, assets with larger expected returns are likely over time to comprise a larger proportion of an unrebalanced portfolio, raising the expected return of the portfolio as a whole (Fernholz and Shay (1982) similarly assume that all

about the nature of the asset or portfolio which generates these returns (except that the distribution of returns is differentiable with finite derivatives), yet it is an accurate approximation when compounding over small time periods over which E[r] is also small. The precise expression derived by Fama and Booth is E[GM] = E[AM] - $\sigma^2/2(1+E[r])^2$, but we use it here in the form in which it is most normally cited, which is of course still a good approximation for small E[r].

[19] In discrete time we have the corresponding expression GM=TW$^{1/n}$ – 1, which is similarly concave.

assets have identical expected growth rates). By removing this effect, our assumption of IID asset returns simplifies the analysis − it is also generous to rebalanced portfolios[20]. As we show below, even with this assumption, unrebalanced portfolios give expected geometric returns equal to those of rebalanced portfolios with equal levels of volatility. Without it, unrebalanced portfolios would outperform.

By definition, the expected portfolio AM is the weighted average of the expected AM of it component assets: $E[AM_p]=\Sigma w_i E[AM_i]$. With E[AM] assumed equal for every asset, $E[AM_p]=E[AM_i]$ regardless of the composition of the portfolio (we assume zero leverage, so portfolio weights always sum to unity). This gives us:

$$E[GM_p] - E[GM_i] \approx \tfrac{1}{2}(\sigma_i^2 - \sigma_p^2) \tag{3}$$

Booth and Fama (1992) define the "diversification return" as the degree to which the expected GM of a portfolio is greater than the weighted average of the expected GMs of its component assets. Our assumption of IID asset returns means that every asset has an identical E[GM] and so the weighted average of these component returns is the same for any unleveraged portfolio of these assets. Equation 3 thus represents the diversification return of shifting from a single asset to a portfolio of similar assets. It also tells us that this diversification return is entirely due to the associated reduction in portfolio volatility. This derivation makes it clear that the diversification return and

[20] If the assets had identical σ_i^2 but different expected AMs then an unrebalanced portfolio would be likely to become more concentrated over time in the assets that have the highest AMs, thus boosting the expected AM of the portfolio as a whole (this would also imply some increase in variance as the portfolio becomes less well diversified and hence suffers greater volatility drag, but for modest shifts in portfolio composition this effect is small). Thus assuming that all assets have identical expected AM returns is not only the simplest assumption – it is also likely to be the most favourable assumption for rebalanced portfolios.

volatility drag are the result of the same underlying process. Indeed the diversification return can be seen as simply the reduction in volatility drag caused by improved diversification. By contrast the "rebalancing return" is portrayed as resulting from a very different process, as rebalancing trades consistently buy on downticks and sell on upticks.

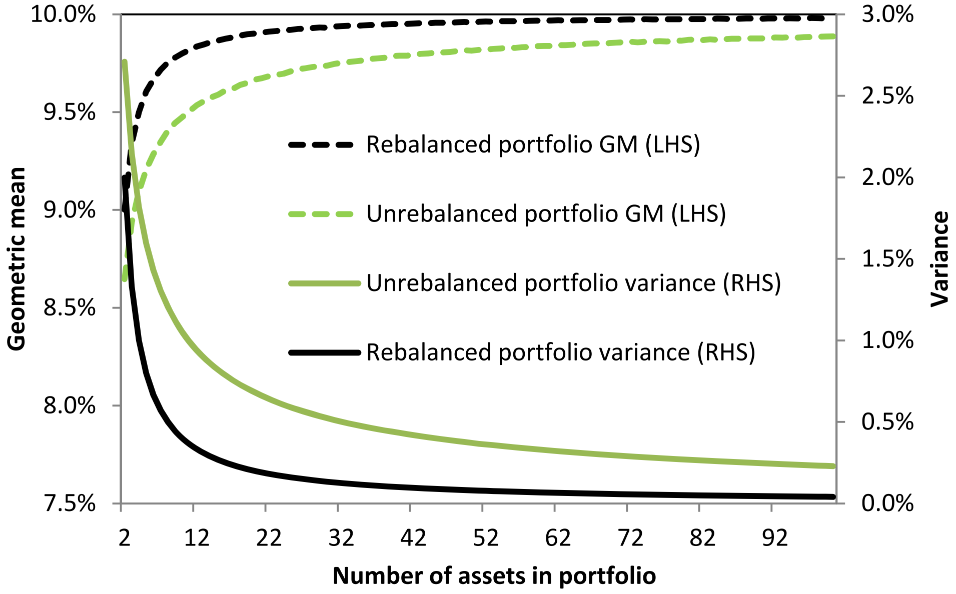
Without rebalancing, portfolio asset weights tend to vary over time. This makes it difficult to derive a simple like-for-like comparison between rebalanced and unrebalanced strategies, since there are two effects which might be thought to cause the E[GM] on rebalanced and unrebalanced portfolios to differ even when asset returns are IID: (i) unrebalanced strategies will tend to become less well diversified over time, leading to higher volatility and hence greater volatility drag; (ii) in addition, proponents claim that the corresponding rebalanced portfolio will benefit from "rebalancing returns". The fact that such otherwise identical strategies must be expected to have different levels of volatility makes it hard to distinguish these two effects.

Our solution is to simulate rebalanced and unrebalanced strategies with a wide range of different variances. We use two separate methods to achieve this: (a) comparing portfolios with different numbers of risky assets; (b) comparing portfolios of two risky assets with different initial weightings given to these assets. There are many other variants that we could use, but these two suffice to show that the different expected GMs of these portfolios are entirely explained by their differing variances (and hence different levels of volatility drag), with rebalancing only relevant to the extent that it affects these variances. We find no "rebalancing returns".

We first consider portfolios containing *N* risky assets which we vary from *N*=2 to 100. For each value of *N* we simulate (i) an equally-weighted portfolio which rebalances at the end of each month to return the weight on each asset to *1/N*; and (ii) an unrebalanced portfolio where each weight is initially *1/N*, but is subsequently allowed to evolve in line with the relative returns on assets in the portfolio. For each portfolio we conduct 10,000 simulations, each over a horizon of 100 years, with monthly asset returns assumed to be normally distributed and serially independent, with annualized (arithmetic) mean 10% and standard deviation 20% per annum for each asset.

Figure 5.1
Rebalanced and unrebalanced portfolios – varying number of assets

This chart shows the annualized GM and variance of rebalanced and unrebalanced portfolios comprising different numbers (from 2 to 100) of component assets. This gives a total of 198 variants. Asset returns are assumed normal and IID, with annualized arithmetic mean 10% and variance 4%. Each path is calculated over 100 years, and the figure reports the average annualized GM and variance over 10,000 simulated paths.

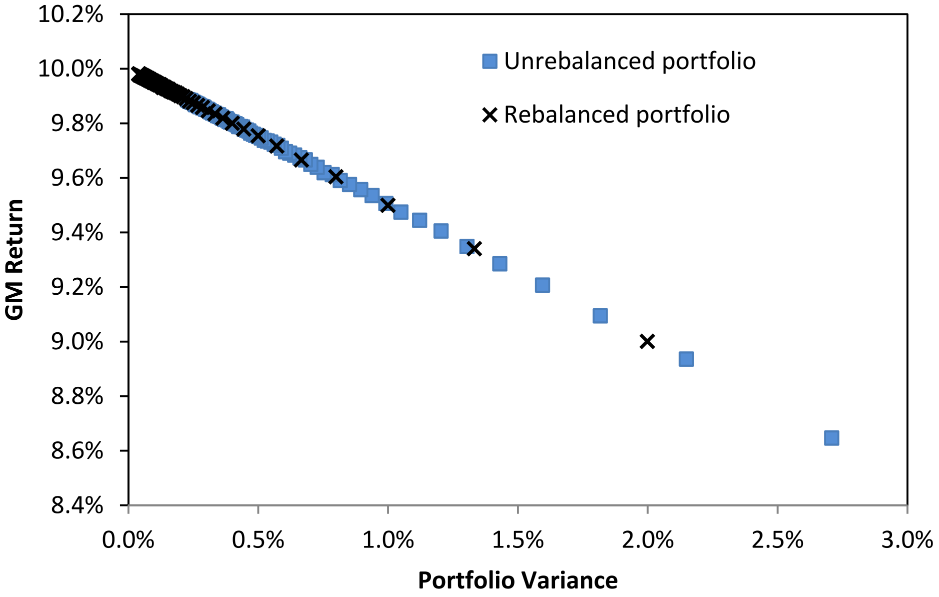


The results are shown in Figure 5.1. For both rebalanced and unrebalanced portfolios the portfolio variance falls as *N* rises. However for each level of *N* the unrebalanced portfolios have higher expected variances. Asset returns are all assumed identically distributed, so an equally weighted portfolio gives the minimum variance. Without rebalancing, the weights on each asset tend to diverge over time, leaving the portfolio less effectively diversified. The expected AM return remains constant for every portfolio (since all assets have identical expected AMs), but the geometric mean returns of these portfolios increase as their variances decrease, consistent with E[GM] ≈ E[AM] - $\sigma^2/2$.

Figure 5.2 presents the same results, but with the average variance for each set of simulations plotted against the corresponding expected GM. The results for the rebalanced and unrebalanced portfolios now coincide. This shows that the choice of whether or not to rebalance affects the expected GM only to the extent that it affects the portfolio variance, and hence generates different levels of volatility drag. By contrast, if rebalancing generated returns by "buying low and selling high" as proponents suggest, we should expect different GMs for these portfolios even after correcting for their different variances.

Figure 5.2
GM vs. Variance for Rebalanced and Unrebalanced Portfolios
(Varying Number of Assets)

This chart shows the same simulation results as in Figure 5.1, but plots the average annualized GM and variance for portfolios of each size N=2 to 100. The figure reports the average annualized GM and variance over 10,000 simulated paths for rebalanced and unrebalanced portfolios. The results show that the expected GM return depends on the average level of portfolio variance, but that for a given level of variance it makes no difference whether the portfolio is rebalanced or not.



Next we examine the relationship between the expected GM and portfolio variance for portfolios of two risky assets with a range of different initial asset weights (Figure 5.3). A fixed 50:50 weighting is the minimum variance portfolio, with unrebalanced portfolios seeing higher variances as the portfolio weights subsequently drift over time. However if the initial portfolio weights are highly unequal (with one asset accounting for 86% or more of the portfolio) then the drift of these portfolio weights in the unrebalanced portfolios on average reduces variance because it can lead to weights becoming substantially more equal over time.

Figure 5.3

Rebalanced & Unrebalanced Portfolios – Varying Initial Portfolio Weights

This chart shows the annualized GM and variance of portfolios comprising two assets. The unrebalanced portfolio initially starts with the weight shown for asset A, but the weights are then allowed to evolve in line with relative asset returns. For the rebalanced portfolio the weight of asset A is returned at the end of each period to its initial value. Initial portfolio weights are given 101 different values (from 0% to 100% asset A) for both rebalanced and unrebalanced portfolios − a total of 202 variants. Asset returns are assumed normal and IID, with annualized arithmetic mean 10% and variance 4%. Each path is calculated over 100 years, and the figure reports the average GM and variance over 10,000 simulated paths.

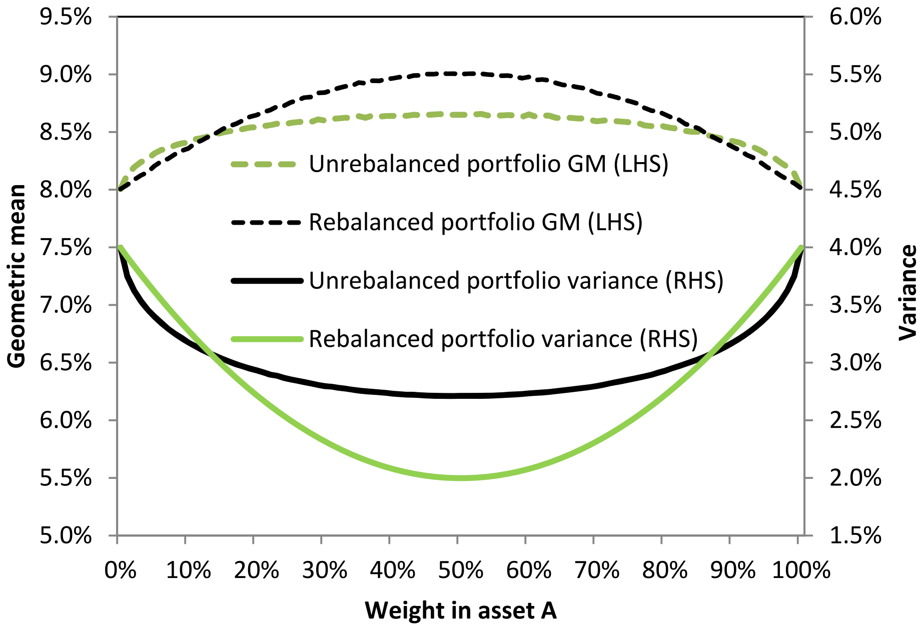
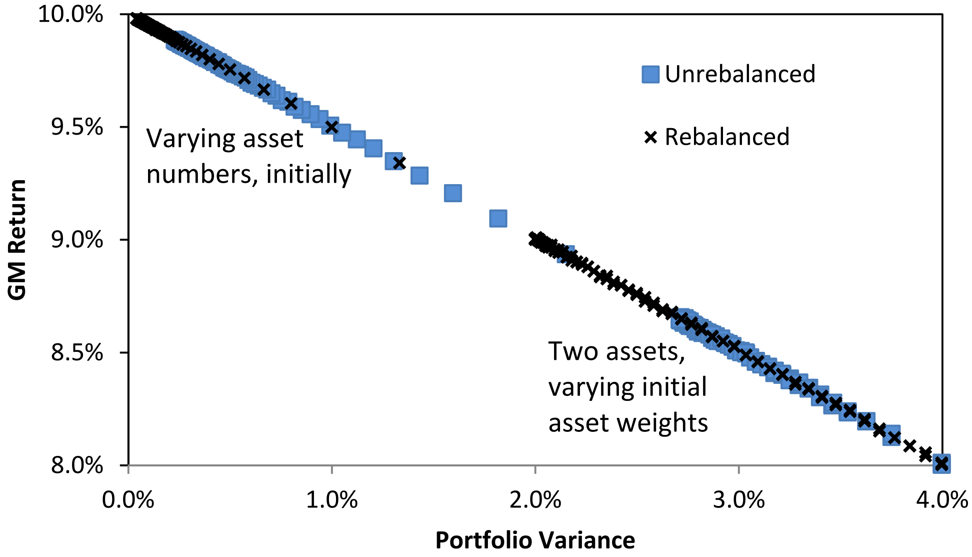


Figure 5.4 plots the same results in terms of mean realised variance versus mean GM return for each set of simulations. The results coincide for rebalanced and unrebalanced portfolios just as they did for our earlier simulations. Furthermore, on this figure we have combined the results of both sets of simulations (i) varying the number of assets (with equal initial weights) and (ii) varying the initial weights in a two-asset portfolio. This shows that all these simulations describe the same linear relationship (E[GM] ≈ E[AM] - $\sigma^2/2$), confirming that rebalancing only affects the average GM to the extent that it affects the average portfolio variance. The average GMs of our simulated

portfolios differ by a maximum of only 0.8 basis points from those implied by this equation[21]. Thus the different E[GM]s of the rebalanced and unrebalanced portfolios can be entirely explained by the different degrees to which they suffer from volatility drag, and we have no evidence of "rebalancing returns" caused by the rebalancing trades themselves being profitable.

Figure 5.4
GM vs. Variance for Rebalanced and Unrebalanced Portfolios

This chart shows the same simulation results as in Figures 5.1 and 5.3, but plots the average annualized GM against the variance. The upper part of the figure shows the relationship between GM and variance when we vary the number of assets in the portfolio (which are initially equally weighted). The lower part of the figure shows the relationship between GM and variance for a two asset portfolio for different initial weights in asset A. The results show that rebalancing affects the expected GM return via its "volatility drag" effect on portfolio variance (following the linear relationship E[GM] ≈ E[AM] – $\sigma^2/2$), but it has no direct impact on the GM.



[21] Annex 1 shows that even though the portfolio variance shifts over time for an unrebalanced portfolio, when compounding over multiple short periods equation 1 is still a good approximation for the whole-horizon expected GM as a function of the average expected AM and average variance over this horizon. Using monthly (rather than continuous) compounding is inherently an approximation, but these results show that in these simulations it is a very good approximation.

Fernholz and Shay (1982) state that a fixed weights portfolio “buys on a downtick and sells on an uptick”, and Luenberger (1997) that it will automatically “buy low and sell high”. Dempster et al. (2009) rightly note that such statements presume negative autocorrelation of returns. The rebalancing process will by construction sell some of an asset after a period in which it outperformed the rest of the portfolio, but this sale is only profitable if it takes place before a period (of whatever duration) of relative underperformance. Indeed, if rebalancing really did buy low and sell high, then it would increase the expected AM as well as the expected GM, but none of the proponents of rebalancing that we cite above claim that it does, and our simulations clearly show that it does not.

Similarly, Willenbrock (2011) argues that “the underlying source of the diversification return is the rebalancing”, and Qian (2012) states that a “diversified portfolio, if left alone and not rebalanced, does not provide diversification return”. These statements are misleading. Rebalancing can be used to keep the portfolio at its minimum-variance weights and hence maximize the diversification return, but this does not imply that the diversification return will be zero in unrebalanced portfolios. Our simulations clearly show that unrebalanced portfolios can achieve substantially higher expected GMs than their component assets.

Whether asset returns are autocorrelated in practice is an empirical question. Rebalancing will be profitable in markets which tend to mean-revert, and loss-making in markets which tend to trend (as assets which underperform in one period also tend to underperform in the next, and vice versa). The misleading conclusion that rebalancing boosts expected GM returns even without any mean-reversion encourages investors to pursue strategies which may be inappropriate for the markets concerned.

3. Rebalancing with One Risky Asset

We now consider a simpler situation: a portfolio consisting of risk-free deposits and a single risky asset with variance σ_a^2. This example is used by Fernholz and Shay (1982) and Qian (2012), and forms the basis of the volatility pumping strategy. This is an important example because it is used to encourage investors to hold volatile assets and poorly diversified portfolios so as to maximise the scale of rebalancing trades. These authors assume that the risk free and the risky asset have identical expected geometric means. For simplicity we follow Dempster et al. (2007) and Qian (2012) in also normalising these returns to zero[22].

Under these assumptions these authors find that the expected geometric mean return of a portfolio which is 50% risky asset and 50% risk-free is $\sigma_a^2/8$. The fact that this return is achieved by combining two assets which each have zero expected GM makes this seem almost like achieving something out of nothing. Furthermore, maintaining the 50% asset weights requires the investor to rebalance by selling some of the risky asset after it has generated a positive return and buying some following a negative return. The positive GM return generated by this strategy is interpreted as resulting from these rebalancing trades. As Fernholz and Shay (1982) put it:

> "...a balanced cash-stock portfolio will buy on a downtick and sell on an uptick. The act of rebalancing the portfolio is like an infinitesimal version of buying at

[22] Without this normalisation, the AM and GM figures in table 1 would all be increased by the risk-free rate. However the key result would remain unchanged: that the risky asset must by implication have E[AM] which is greater than the risk-free rate, and that this should be seen as the underlying source of the positive expected GM on the 50/50 portfolio, which has half the E[AM] of the risky asset but only one quarter of the volatility drag.

> the lows and selling at the highs. The continuous sequence of fluctuations in the price of the stock produces a constant accrual of revenues to the portfolio."

The language used here suggests that these price movements are temporary and rapidly reversed. By construction, rebalancing sells a proportion of the assets that outperformed in the most recent period and buys those that underperformed. The profitability of these trades depends on these price moves subsequently reversing, but in a geometric Brownian motion assets that outperform in one period are as likely to outperform in future periods as they were in the first. Specifically, Fernholz and Shay (1982) assumes a Brownian motion in which the risky asset has an expected growth rate of zero, so after the asset price has initially diverged from its original value, the expected geometric return on any shares bought or sold at this new price is zero). This applies in every period. Any rebalancing trade shifts some wealth from one asset into another which is as likely to outperform as underperform the asset it replaces. This shift will raise the expected portfolio growth rate if it improves diversification and so reduces volatility drag. But the language used by proponents of rebalancing strategies suggests a very different effect is at work.

We can also demonstrate that the rebalancing trades are not the source of the increased GM return by deriving the expected size of this increase without using any dynamic expressions, but instead merely using the standard arithmetic/geometric mean relationship (E[GM] ≈ E[AM] - $\sigma^2/2$). This equation is of general applicability, and makes no presumption about rebalancing − it applies to the returns of both rebalanced and unrebalanced portfolios, and indeed to all positive numbers. In this case it tells us that the risk-free asset must have zero expected AM (since it is assumed to have zero GM and zero variance), but that the risky asset must have a positive expected AM of $\sigma_a^2/2$

(see Table 5.1). This positive arithmetic mean is generally not made explicit in discussions of this strategy, thus helping to maintain the impression that the expected geometric mean return on the 50:50 portfolio is caused by the rebalancing trades buying low and selling high.

Table 5.1: Expected AMs and GMs derived using E[GM]≈E[AM] - $\sigma^2/2$

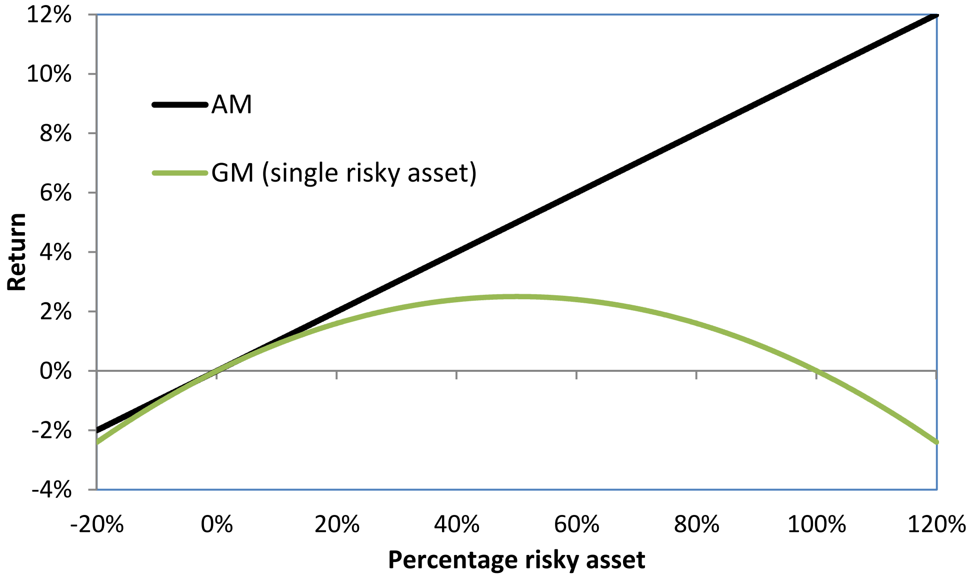
	E[AM]	Variance	E[GM]
Risky asset	$\sigma_a^2/2$	σ_a^2	0
Risk-free asset	0	0	0
50:50 fixed weight portfolio	$\sigma_a^2/4$	$\sigma_a^2/4$	$\sigma_a^2/8$

The 50:50 portfolio has an expected AM equal to half that of the risky asset, but only one quarter of the variance. Thus it must have a GM of $\sigma_a^2/8$. Fernholz and Shay (1982) derive this result using stochastic calculus to model the dynamics of the portfolio, but it follows directly from the standard AM:GM relationship which applies for all assets and portfolios, regardless of whether they are rebalanced. This return is better interpreted as arising because the risky asset itself has a positive arithmetic mean return. For the 100% risky asset portfolio this positive expected AM is perfectly offset by volatility drag. By contrast, the 50/50 portfolio has an expected AM which is half as large, but it suffers only one quarter of the volatility drag, leaving it a positive expected GM. The positive E[GM] of the rebalanced portfolio is thus explained entirely by the reduced volatility drag, so there is no evidence of the buy-low/sell-high effects which are claimed.

Portfolio rebalancing played no part in deriving the size of the expected GM for this portfolio, but in practice if there is no rebalancing then the proportion *a* of the portfolio which is held in this single risky asset is likely to vary from one period to the next. The expected AM in any period increases in direct proportion to a, whilst the variance of the portfolio increases with a^2. Thus the expected GM in any period has a quadratic relationship[23] with a, with the maximum expected GM at $a = 0.5$ as shown in Figure 5.5.

Figure 5.5: Volatility Pumping – Single Risky Asset

This chart shows the arithmetic and geometric means of portfolios comprising a risk-free asset (zero expected AM return and zero variance) and a risky asset (expected AM return 2%, and variance 4%). Equation (1) implies that the risky and risk-free assets each have zero expected GM, but for portfolios with a positive weight on each asset the expected GM must be positive.



Rebalancing is required to maximise E[GM] by keeping the portfolio composition at 50/50. If a falls to zero then the portfolio is composed entirely of risk-free asset, with zero GM. If a rises to 1 then the portfolio is entirely risky asset and the volatility drag will completely offset the positive E[AM], leaving E[GM] zero. However, as long as risky asset follows a distribution such as the lognormal which does not allow prices to fall to precisely zero, the proportion of the portfolio which is accounted for by the risky asset must be strictly greater than zero and less than 100%, so the expected GM of the

[23] $E[GM] = E[AM] – \sigma_p^2/2 = a\ \sigma_a^2\ /2 - a^2\sigma_a^2/2 = a(1- a)\ \sigma_a^2/2$. This result is derived by Qian (2012) and, in continuous time form, by Fernholz and Shay (1982).

unrebalanced portfolio will be strictly positive in every period, and hence also positive for the horizon as a whole[24]. Thus it is extremely misleading to label the E[GM] of the rebalanced strategy the "rebalancing return", since unrebalanced portfolios must also have a positive E[GM]. Indeed for short time horizons the proportion of the risky asset in the portfolio should not be expected to diverge substantially from its initial 50%, so the expected GM will be only slightly below the expected GM of the rebalanced portfolio.

The rebalancing trades are mistakenly regarded as the source of the return on the rebalanced portfolio, so the larger rebalancing trades associated with more volatile assets are regarded as more desirable: "The pumping effect is obviously most dramatic when the original variance is high. After being convinced of this, you will likely begin to enjoy volatility, seeking it out for your investment rather than shunning it" (Luenberger, 1997).

It is worth considering the welfare effects of this volatility pumping strategy. First, far from being a source of profits, these rebalancing trades are likely to be costly due to transaction and market impact costs. Investors would be better advised to rebalance to the minimum extent that is consistent with keeping the portfolio composition acceptably close to their target weights. Second, the desire to maximise these transactions may push investors into sub-optimal asset allocations. If there is only a single risky asset, with an expected GM equal to the risk-free rate, then the 50:50 fixed weight portfolio will indeed be the most attractive option for an investor who wishes to maximise his expected portfolio GM. But in practice there are likely to be many alternative risky assets which are less than perfectly correlated. This allows superior strategies to be constructed. If, for example, two or more assets have the same expected

[24] As shown in Annex 1, a simple average of the expected GM over all periods in the horizon gives 1/n $\Sigma E[\ln(1+r_t)]$, which can be rearranged to give the expected GM for the horizon ($E[\ln(\Pi(1+r_t)^{1/T}]$). This must be positive given that $E[\ln(1+r_t)]$ is positive in every period.

AM and variance then a portfolio of them (however weighted) will have the same expected AM, but a lower variance, and thus a higher expected GM. Combining this multi-asset portfolio with a fixed *a%* of cash will (for any *a*>0) generate an expected GM which is greater than that shown in Figure 5.5.

Fernholz/Shay (1982) is still very widely cited in support of rebalancing strategies. It supports its claim that such strategies benefit from “buying on downticks and selling on upticks” by noting:

> “Let *Z* be a balanced cash-stock portfolio with Π=½ [Π is the proportion invested in the risky asset], and let *W* be a passive [unrebalanced] cash-stock portfolio which starts with equal proportions of cash and stock. From equation (22) we see that every time that *W* returns to equal proportions ─ and this will occur infinitely often with probability one ─ *Z* will be greater than *W* ”

It is true that the passive portfolio will return to equal proportions infinitely often with a probability that tends to one as time tends to infinity. This appears to imply that every rebalancing trade must eventually end up generating a profit, but this is misleading. The unrebalanced portfolio always contains a fixed amount of cash ($W_0/2$, where W_0 is initial wealth). The value of the portfolio holding of the risky asset (S_t) is likely to diverge from this, triggering rebalancing trades. If S_t subsequently returns to its initial price (S_0) then rebalancing trades made since it last diverged from S_0 will end up having been profitable. If the risky asset initially underperforms (outperforms), then the rebalancing trades will have bought (sold) some of this asset, which will turn out to have been profitable if S_t rises (falls) back to S_0. The probability of S_t returning to S_0 clearly tends to one as time tends to infinity, since an ever-increasing proportion of future possible paths will at some stage hit this value. However the same is true of the proportion of paths for which reach any arbitrary value S*, so the relevance of this fact is clearly questionable.

Furthermore, even though the proportion of these paths which at some point return to S_t=S_0 tends to 1, each period the set of paths which have not yet returned to S_0 will on average diverge further from S_0 than in the previous period. When considering the expected future values of S_t, we need to weight together the increasing proportion of paths which at some stage will have returned to S_0 and the decreasing proportion of paths which are diverging from S_0 by (on average) ever-larger amounts and for which a rebalancing trade made when it first diverged from S_0 will be recording ever-larger losses. It is only by taking both these groups into account that we can make meaningful statements about the expected return on the rebalanced portfolio.[25]

Indeed, it is straightforward to demonstrate that the expected value of the unrebalanced portfolio is in fact greater than for the rebalanced portfolio[26]. We assume that the price of our risky asset follows a standard geometric Brownian motion:

$$\frac{dS}{S} = \mu dt + \sigma dB \tag{4}$$

This integrates (using the Ito integral) to give:

$$S_t = S_0 e^{\left(\mu - \frac{\sigma^2}{2}\right)t + \sigma B(t)} \tag{5}$$

[25] Similarly, we have the standard result that the expected length of time required for a random walk to converge to any arbitrarily distant level is infinite (when the GBM has zero drift, $\mu=\sigma^2/2$, as Fernholz/Shay assume). The time to convergence for many paths may be very short, and the proportion of paths which reach this level inevitably rises over time, but the remaining paths that have not yet converged will on average have moved in the opposite direction.

[26] We are grateful to an anonymous referee for pointing out this derivation.

Without loss of generality we normalise S_0=1. Fernholz and Shay (1982) and other papers also assume that the risky asset has zero expected growth rate, which implies that $\mu=\sigma^2/2$. This can be interpreted as the risky asset having a positive expected AM return which is exactly balanced out by volatility drag, leaving zero expected GM:

$$E[\log(S_t)] = E[\sigma B(t)] = 0 \tag{6}$$

We can derive an expression for the corresponding expected value of the risky asset by using a Taylor expansion of the exponent, simplifying using the standard properties of the Wiener process (E[B(t)]=0 and E[B^2(t)]=t). The accuracy of this approximation depends on standard assumptions that higher moments are well behaved.

$$E[S_t] = E\left[e^{\sigma B(t)}\right] = e^{\frac{\sigma^2 t}{2}} \tag{7}$$

A portfolio P_r which is constantly rebalanced to keep 50% in the risky asset and 50% in a risk free asset (with return normalised to zero, for simplicity) will always have $\mu=\sigma^2/4$ and standard deviation σ/2 (values that are half those of the asset itself). Substituting these into equation 5 gives us:

$$P_r = e^{\frac{\sigma^2}{8}t+\frac{\sigma}{2}B(t)} \tag{8}$$

This shows that even though the risky asset has an expected growth rate of zero, the rebalanced portfolio has an expected growth rate of σ^2t/8. This is exactly the situation that we considered in discrete time in Figure 5.5 and Table 5.1. Only the risky asset generates growth, and the rebalanced portfolio holds half as much as a portfolio which is entirely composed of the risky asset, but has one quarter of the volatility drag. Again, using a Taylor expansion to simplify gives us:

$$E[P_r] = e^{\frac{\sigma^2 t}{4}} \tag{9}$$

The corresponding unrebalanced portfolio is simply a fixed 0.5 invested in the risky asset, and an initial 0.5 invested in the risky asset:

$$E[P_u] = 0.5 + 0.5e^{\frac{\sigma^2 t}{2}} \tag{10}$$

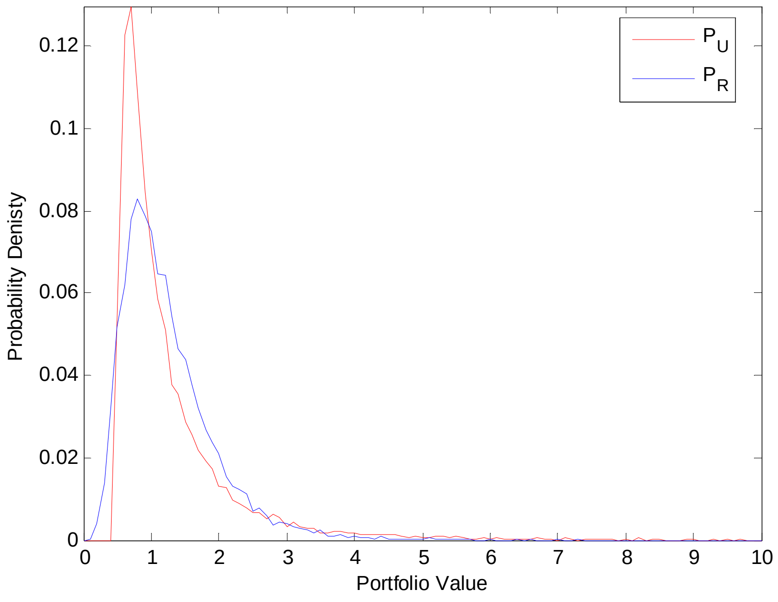
By inspection, the ratio $E[P_r]/E[P_u]$ clearly tends to zero as the time horizon increases, and our simulations confirm that the average terminal value is higher for unrebalanced portfolios. This may seem odd given that in the ever-increasing proportion of paths where the risky asset price returns to S_0 a rebalancing trade (which would have sold some of the risky asset after it initially outperformed and bought after it initially underperformed) would have turned out to be profitable. Simulations help resolve this apparent contradiction.

Figure 5.6 shows that P_u outperforms at both tails of the distribution. Rebalancing trades are profitable on average on paths where S_t tends to mean revert, leaving only small positive or negative cumulative returns. Conversely, when S_t makes large cumulative moves in either direction (i.e. tends not to mean-revert) then P_r underperforms P_u.

As time passes the left tail of the P_r distribution becomes vanishingly small, since the expected growth rate of this portfolio is positive, and fewer and fewer outturns see P_r below 0.5. But on the right tail, P_u shows more extreme outturns than P_r. Over time this tail represents a smaller and smaller probability space, but the average size of P_u-P_r in these cases keeps increasing. This ever-more-extended, but ever-less-likely tail explains why $E[P_u]>E[P_r]$ even though the probability that $P_r>P_u$ tends to 1 as time passes.

Figure 5.6
Probability Density of the Terminal Wealth of Rebalanced and Unrebalanced Portfolios after 100 Years

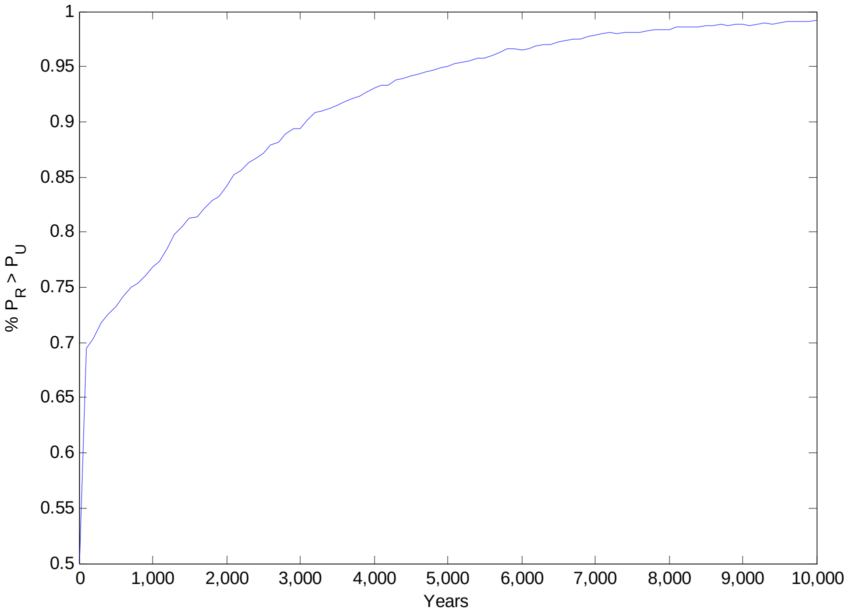
We follow Fernholz and Shay (1982) in assuming asset returns follow a geometric Brownian motion with zero expected geometric return. We consider two portfolios with starting value \$1. Both invest \$0.50 in a risk-free asset which has an interest rate of zero and \$0.50 in the risky asset. The rebalanced portfolio rebalances back to 50/50 asset mix every month. We assume σ is 10% per annum for the risky asset (other simulations, not reported here, show our results are robust to alternative assumptions). The chart shows the distribution of terminal wealth over 10,000 simulated paths of a 100 year time period.



Fernholz and Shay (1982) focuses not on the expected value of the rebalanced and unrebalanced portfolios, but on the fact that the risky asset returns to its initial value “infinitely often with probability 1”. In such cases the rebalanced portfolio will outperform the unrebalanced, but it is very misleading to assume in practice that this probability is close to 1. Figure 5.7 shows that for typical parameter values it takes several millennia for the probability that (p_r>p_u) to get anywhere close to unity. Over horizons of up to 100 years P_r outperforms P_u in less than 70% of our 10,000 simulations. Thus in practice it is extremely misleading to assume that the probability that P_r>P_u is anywhere near 1. Fernholz and Shay (1982) base their arguments on the properties of the payoff distributions as time tends to infinity, but the distributions faced by investors in practice will be very different.

Figure 5.7
Proportion of Outcomes where $P_r > P_u$

This chart shows the proportion of the simulated paths for which the terminal wealth of the rebalanced portfolio P_r is greater than the unrebalanced portfolio P_u. This proportion is shown for simulations with a wide range of different time horizons. The parameters of the simulated values are the same as for Figure 5.6.



4. Conclusion

It is widely claimed that rebalancing strategies generate "rebalancing returns" by buying on downticks and selling on upticks. This paper demonstrates instead that the difference between the expected GMs of rebalanced and unrebalanced portfolios can be entirely explained by their different degrees of volatility drag, with no evidence of "rebalancing returns". This paper also shows that the arguments used by key proponents of rebalancing strategies are based on properties of returns at infinite horizons which are not applicable over practical investor lifetimes unless we assume that risky asset prices tend to mean revert.

These misleading arguments have important implications, since they encourage investors to hold portfolios which are concentrated in volatile assets so as to increase the scale of the resulting rebalancing trades. Investors would be better advised to seek to minimize volatility drag by diversifying effectively and to rebalance no more than is necessary to keep their portfolio compositions adequately close to their target allocations.

References

Booth, David and Fama, Eugene (1992), "Diversification Returns and Asset Contributions," *Financial Analysts Journal* 48, 26-32.

Bouchey, P, Nemtchinov, V, Paulsen, A and Stein, D.M. (2012) “Volatility Harvesting: Why Does Diversifying and Rebalancing Create Portfolio Growth?” *Journal of Wealth Management*, Vol. 15, No. 2, 26-35

Cover, Thomas M. (1991) “Universal Portfolios”, Mathematical Finance Vol.1, Issue 1, 1-29.

de La Grandville, Olivier (1998) “The Long-Term Expected Rate of Return: Setting It Right”, *Financial Analysts Journal*, vol. 54, no. 6, 75-80.

Dempster, Michael AH, Igor V. Evstigneev, and Klaus R. Schenk-Hoppé (2007) "Volatility-induced financial growth." *Quantitative Finance* 7.2, 151-160.

Dempster, Michael AH, Igor V. Evstigneev, and Klaus R. Schenk-Hoppé. (2009) “Growing wealth in fixed-mix strategies.” in Maclean, L, Thorp, E and Ziemba, W. “The Kelly Capital Growth Investment Criterion: Theory and Practice”, World Scientific Publishing.

Fernholz, E. Robert and Maguire, Cary (2007) “The Statistics of Statistical Arbitrage”, *Financial Analysts Journal*, Vol. 63, No. 5, 46-52.

Fernholz, Robert, and Brian Shay (1982). "Stochastic Portfolio Theory and Stock Market Equilibrium", *Journal of Finance*, vol. 37: 615-624.

Luenberger, David G. (1997, 2nd ed 2013) “Investment Science”, Oxford University Press.

MacBeth, James D. (1995) “What's the Long-Term Expected Return to Your Portfolio?” Financial Analysts Journal, Vol. 51, No. 5, 6-8.

Qian, Edward (2012) “Diversification Return and Leveraged Portfolios”, *Journal of Portfolio Management*, Vol. 38, No. 4, 14-25.

Willenbrock, Scott (2011) “Diversification Return, Portfolio Rebalancing, and the Commodity Return Puzzle”. *Financial Analysts Journal*, Vol. 67, No. 4: 42–49

Annex 1

Fama and Booth (1992) show that the continuously compounded holding period return is well approximated by the expected return expressed in continuously compounded terms minus a fraction of the variance of the simple returns.

$$\mathrm{E}\left[\log(1+\mathrm{r})\right] \approx \log(1+\mathrm{E}[\mathrm{r}]) - \frac{\sigma^2}{2(1+E[r])^2} \tag{A1}$$

This equation holds in each period, so we can sum each side over periods 1 to *T*:

$$\sum_{\mathrm{t}=1}^{\mathrm{T}} \mathrm{E}\left[\log(1+\mathrm{r}_\mathrm{t})\right] \approx \sum_{\mathrm{t}=1}^{\mathrm{T}} \log(1+\mathrm{E}[\mathrm{r}_\mathrm{t}]) \; - \; \sum_{\mathrm{t}=1}^{\mathrm{T}} \frac{\sigma_t^{\;2}}{2(1+E[r_t])^2} \tag{A2}$$

Rearranging and dividing through by *T*:

$$\frac{1}{\mathrm{T}}\mathrm{E}[\log\left(\prod_{\mathrm{t}=1}^{\mathrm{T}}(1+\mathrm{r}_\mathrm{t})\right)] \approx \frac{1}{\mathrm{T}}\sum_{\mathrm{t}=1}^{\mathrm{T}} \log(1+\mathrm{E}[\mathrm{r}_\mathrm{t}]) \; - \; \frac{1}{\mathrm{T}}\sum_{\mathrm{t}=1}^{\mathrm{T}} \frac{\sigma_t^{\;2}}{2(1+E[r_t])^2} \tag{A3}$$

$$\approx \mathrm{E}[\mathrm{r}_\mathrm{t}] - \frac{1}{2\mathrm{T}}\sum_{\mathrm{t}=1}^{\mathrm{T}} \sigma_t^{\;2} \tag{A4}$$

If each period is short then E[r$_t$] will be small and the continuously-compounded GM and AM above will be close to their more commonly-used discretely compounded equivalents. Thus we end up with a form of the standard relationship E[GM] ≈ E[AM] - $\sigma^2/2$ which applies even if the distribution of r$_t$ varies over time: the expected GM over the whole multi-period horizon is approximately equal to the average expected return over these periods minus half of the average variance. The linear relationships shown in figures 5.2 and 5.4 confirm that this relationship holds for the unrebalanced portfolio, whose E[AM] and variance shift over time.

Chapter 6

Conclusion: cognitive error in measuring investment returns

This thesis consists of four substantive papers and a literature review. The underlying theme is that investors are encouraged to follow some strategies for reasons that seem convincing, but are in fact based on misleading analysis. The contribution of this thesis takes three forms. First, each of these papers contradicts some aspects of the current academic literature:

> In most situations, buying at a lower average cost implies higher expected returns. Chapter 2 explains why investors are mistaken in believing that the same is true for Dollar Cost Averaging (DCA). It also demonstrates that recent academic research is wrong to argue that using DCA is beneficial for certain types of non-variance investor risk preferences. I show instead that DCA is inefficient regardless of the form taken by such preferences.
>
> The (smaller) literature on Value Averaging (VA) claims that it generates superior returns. Chapter 3 demonstrates that VA is an inefficient strategy. It also demonstrates that the IRR and MIRR will be systematically biased measures of investor returns for a wide range of dynamic strategies involving target returns, stop-losses or profit taking.
>
> An increasing number of published papers argue that investor timing has been much worse than was previously thought, and has significantly reduced aggregate returns to investors. Chapter 4 demonstrates that the technique used by these papers (using the difference between the IRR and GM returns to measure the impact of investor timing) is flawed, and that for US equities almost the whole of this differential is due

to a hindsight effect. This has important implications for our estimates of the equity risk premium.

Chapter 5 shows that the 'rebalancing returns' claimed for volatility pumping strategies are likely to be due to reduced volatility drag and that superior strategies can be constructed which offer better diversification and lower transaction costs.

In writing this thesis I have not sought to be contrary for its own sake. Instead this thesis took shape as I came across published research that seemed to me counter-intuitive. I then decided to investigate. If the analysis in this thesis is robust in its criticisms, then the points above represent a meaningful contribution to the academic literature.

The second contribution is that Chapter 4 also derives a technique for adjusting the IRR so as to remove the hindsight effect, leaving only the genuine timing effect of investment flows. This new technique is likely to have applications in a wide range of fields where the IRR is used.

The final contribution lies in the possibility of improving investor decision-making. Academics are generally reticent about prescribing particular courses of action to investors and rightly so, since behavioural finance has amply demonstrated that investors' underlying motives are much more complex than conventional finance theory had assumed. Prescriptive advice based on these assumptions would have been misplaced. However, this thesis has identified widespread advice in the popular media – and sometimes in academic journals – that recommends strategies for reasons that are demonstrably false. Investors will continue to make decisions based on their own idiosyncratic preferences and beliefs, but identifying falsehoods in the advice that they are given represents a contribution to good investment practice.