Working Version: Review of the Your Future Your Super Performance Test

Detailed Paper

20 November 2020

David Bell, Emily Barlow, Andrew Boal, Kim Bowater, Nick Callil, David Carruthers, Matthew Griffith, Clayton Sills, Tim Unger

Table of Contents

1. Introduction 3
2. Analysis 4
   2.1. What drives consumer outcomes? 4
   2.2. The effectiveness of the YFYS performance metric at identifying 'poor' funds 15
3. Undesirable outcomes 28
   3.1. How will funds invest in the presence of the YFYS performance test? 29
   3.2. How will the YFYS performance test directly impact consumers? 37
   3.3. How will the YFYS performance test impact industry structure? 39
4. Solutions 40
   4.1. Principles of an investment performance test 40
   4.2. Addressing the principles to develop an effective performance test 42
   4.3. A family of solutions 43
5. Conclusion 46

Appendix 1 – Brief literature review on performance persistence 48
Appendix 2 – Approximation of implementation volatility 51
Appendix 3 – Benchmarking 'noise' created by YFYS performance test 52
Appendix 4 – Estimating 'implementation noise' which comes from YFYS benchmarking approach 56
Appendix 5 - Proposing a single metric (working version) 57

1. Introduction

The Government recently announced the Your Future Your Super (YFYS) reforms as part of the 2020 Budget. There is merit in protecting consumers with a performance test. However, it needs to be an effective performance test with limited undesirable outcomes. Unfortunately, our analysis suggests the YFYS performance test does not meet these goals: it will be ineffective at identifying poor performing funds while introducing a range of undesirable outcomes. We are concerned that the detriments of the YFYS performance test may outweigh the benefits.

To summarise our analysis:

- An investment performance metric has shortcomings: it doesn't account for the total consumer outcome (by considering product design), fails to account for changes made through time by super funds, while the evidence that past performance is a predictor of future performance (performance persistence) is modest (total fees and governance appear informers of future outcomes).

- Placing these issues aside we assessed the YFYS performance metric. The statistical effectiveness of the YFYS performance metric at identifying poor performing funds is found to be weak. The YFYS performance metric faces three major challenges: (1) timeframe (8 years may not be sufficient), (2) it focuses on one (likely minor) component of performance (implementation[1]) rather than investment performance in total, and (3) the benchmarking process generates inaccuracies which create benchmark 'noise'. Our statistical analysis reveals that, in its current form, the performance metric will be ineffective at identifying poor performers and have a high likelihood of falsely identifying good performers as poor.

- We believe the YFYS performance test will result in undesirable outcomes relating to how funds invest, may have some adverse impacts on consumers, and create a distorted industry structure (the potential for 'zombie' funds). We expect there to be a detrimental effect on both industry performance and individual consumer outcomes.

We detail a range of alternative solutions and self-assess through the lens of consumer outcomes. To summarise:

[1] Here, implementation performance includes asset level performance assessed against ascribed sector benchmarks and the performance of tactical asset allocation calls (performance from having an asset allocation which is different from the fund's strategic target).

- A solution which involves regulator assessment is, as it stands, the only way to acknowledge the qualitative issues and the through-time changes made by funds. It can also incorporate metric(s).

- Better metrics exist than the YFYS performance metric and a well-designed collection of metrics has advantages over a single metric.

- The YFYS performance metric can be improved. At best it can be more stable and assess all fund-types consistently. Nonetheless, it will remain statistically ineffective at assessing implementation performance, which itself is a small component of total investment performance.

We are positive about the opportunity to improve consumer outcomes. There is a great opportunity to implement an effective performance test which protects consumers from being exposed to funds which are likely to underperform in the future, whilst limiting undesirable outcomes. We are happy to share models and discuss this work and solutions in further detail.

2. Analysis

2.1. What drives consumer outcomes?

2.1.1. Superannuation is just a component of the retirement system

Superannuation is a complex system. In Diagram 1 we provide an overview through the lens of the outcomes it delivers. Investment performance is an aggregate of three components: (1) the level of risk, (2) asset allocation, and (3) implementation. Investment performance is delivered within a product design which may vary the risk target through time to account for expected future contributions (known as a lifecycle strategy). The superannuation product applies to a range of consumers with unique characteristics relating to their own income experiences. A range of possible outcomes is the result with an observed outcome effectively being one realisation of a process.

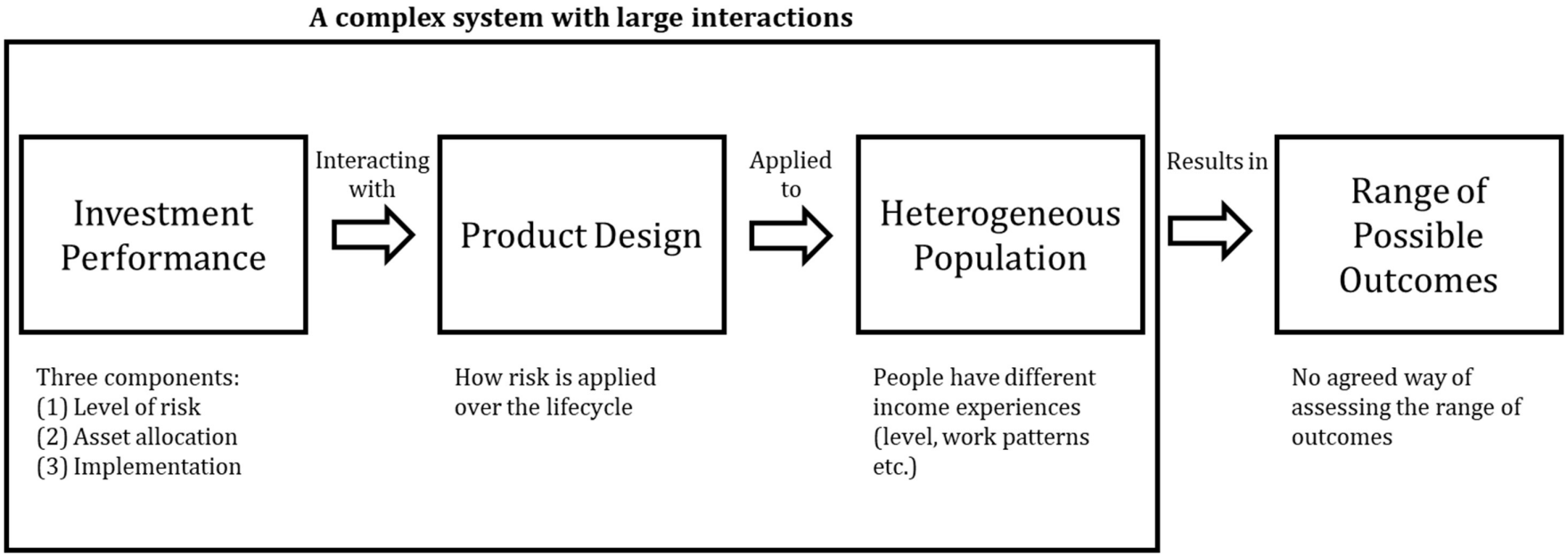

Diagram 1: Overview of superannuation system from the perspective of the outcomes it produces.

Diagram 1 largely describes an accumulation framework whereas superannuation forms part of the retirement system. The way that superannuation interacts with other parts of Australia's three-pillar framework is important. Technically, this makes the problem more complex, especially so when we consider retirement income as a through-time income stream.

Whether the focus is on accumulation or retirement income, the overall assessment should account for the range of possible outcomes. At present this is an area of weakness for industry, regulators and policymakers: there is no generally accepted approach to assessing an outcome of this nature.

Academia leads on this issue: the traditional approach used by leading academics for over 50 years[2] considers the range of outcomes and effectively assesses each possible outcome using what is known as a utility (or preference) function[3].

In Australia there is an emerging body of research using such techniques. MDUF (2017) was a working group effort to develop an industry standard utility function, while Warren (2019) explores how utility functions can be tailored to better reflect different investor-types. Both these examples apply to retirement income.

A higher order assessment of retirement system outcomes and superannuation outcomes in general requires the use of more advanced techniques to assess the range of possible outcomes. Unfortunately, no review has taken on this challenge to date.

2.1.2. Focusing on investment performance

Accordingly, the focus of superannuation assessment is on performance. As mentioned previously superannuation performance is impacted three primary attributes (also reflected in Diagram 2):

1. Risk: the overall level of investment risk taken on
2. Asset allocation: how that risk is allocated across asset classes over time
3. Implementation: performance achieved within each asset class (sometimes called 'implementation alpha')

[2] Samuelson (1969) and Merton (1969) are considered seminal in this respect.
[3] A utility or preference function could be considered a mechanism to score the experience of each possible outcome in the hands of a representative individual.

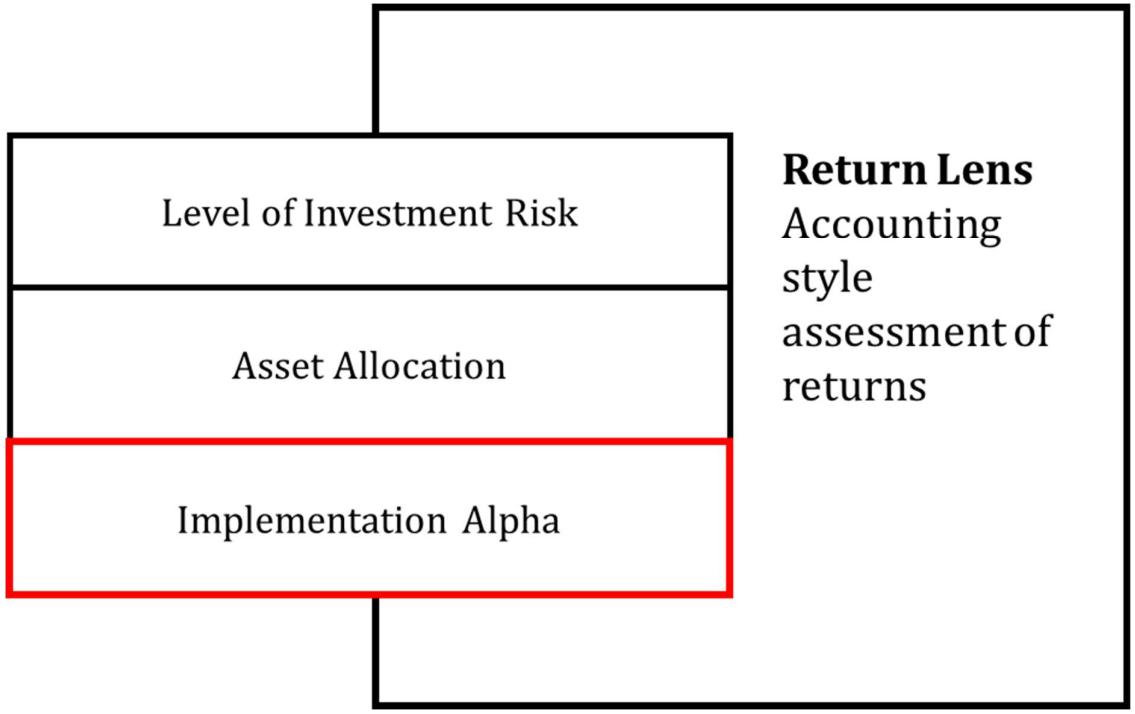

Diagram 2: Simplified summary of components of super fund investment performance. The red box reflects the focus of the Your Future Your Super performance metric.

We note that risk and asset allocation can be separated into longer-term (strategic) and short-term (tactical) decisions. Overall, this represents a multi-dimensional problem which makes performance assessment a complex exercise. However, it has been heavily explored in the area of investment

management where the Brinson-Fachler model (1985) broadly accounts for (1), (2) and (3). This implicitly assumes that (1) and (2) have stable, long-term targets whereas for super funds the long-term positioning is not fixed and is a responsibility of the trustee[4].

Which of the three areas has the most impact on super fund returns? This topic has been heavily explored in both industry and academia. Brinson et al (1986) calculate that investment policy (in our case the long-term targets for (1) and (2)) explain on average nearly 94% of variation in total US pension plan returns. This result was controversial, largely because it seems to have been misinterpreted. Subsequent research by Ibbotson and Kaplan (2000) found the following:

1. Asset allocation explains about 90% of the variability of a pension fund's returns over time
2. Asset allocation explains about 40% of the variation of returns among funds
3. On average across funds, asset allocation policy explains slightly more than 100 percent of the level of returns

How should these results be interpreted?

- The results do not invalidate the importance of implementation, which could represent a constant (hence low volatility) leakage of returns
- No component of performance should be ignored as each can have a significant impact on consumer outcomes and each is an active trustee decision

Three simple cases demonstrate how each component may impact consumer outcomes:

1. Risk: allocating 10% more to growth assets over the last 8 years would have added 57bp pa over that time period. Further corroboration can be found by estimating the slope per unit of growth exposure in the APRA Heatmap results.

2. Asset allocation: deciding to allocate 10% more to global shares rather than Australian shares would have added 43bp pa for the 8yrs.

3. Implementation: the Productivity Commission found examples where the implementation performance of a number of funds (including administration fees) was > 20bp pa.

The above three cases do not determine which component has the greatest impact on total investment outcomes. However, we are sure that the impact of risk and asset allocation can be significant. The presence of multiple sources of return generates interaction challenges. While it is convenient to

[4] Note that Brinson-Fachler detail the existence of interaction effects (i.e. a decision to take an active position in a sector involves taking an active position in the implementation of that sector as well) which further adds to complexity.

attribute performance to individual sources, this may be incongruent with how trustees align their decision-making when focused on total member outcomes. Consider Example 1:

Example 1

A trustee believes that high yield bonds have good risk-adjusted return potential. However, they are cognisant that, due to the liquidity profile and market structure of this segment, implementation performance is likely to be negative, even when investing passively (further background). The trustee accounts for this and still believes the total benefit to members merits an allocation.

Example 1 appears a well-considered decision through the lens of consumer outcomes. The expected outcome is positive performance from asset allocation which more than offsets the negative outcome from implementation. We consider it important that all components of investment performance are accounted for otherwise investment decision-making is distorted away from best member outcomes.

A point of definition relating to the YFYS performance metric: the implementation measure used in the YFYS metric includes the traditional definition of implementation described above as well the impact of any tactical asset allocation (TAA) decisions made. From hereon our definition of implementation is consistent with that in YFYS. Expanding our definition to include TAA does not notably affect the issues raised above. To minimise confusion, we focus little on TAA in this paper.

2.1.3. Return is not the only realised outcome

As detailed in Diagram 3, returns are not the only output from the investment process: realised experiences of variability are also an output. Here we use volatility as a descriptive measure of realised risk.

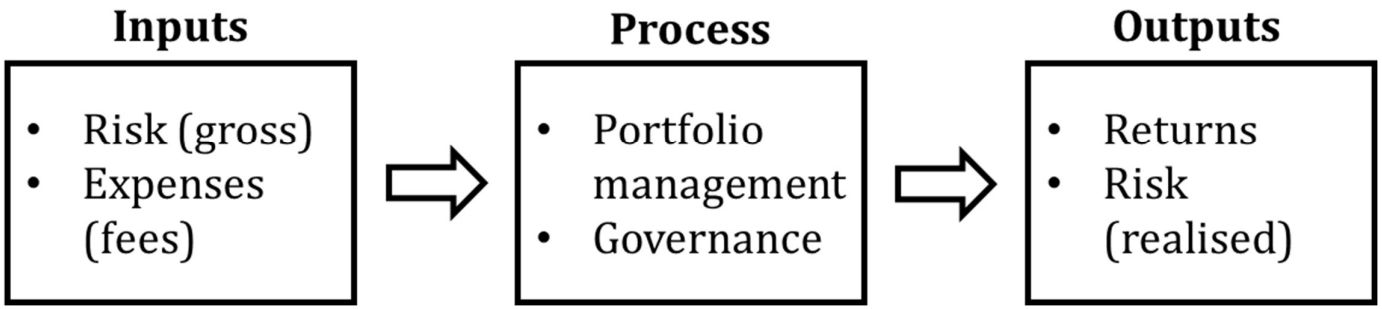

Diagram 3: Process representation of investment management. Note that chosen inputs (risk and expenses) account for the assessed opportunity set (expected returns, forecast volatility, correlations etc.), fund membership characteristics (demographic profile), and fund-level constraints such as liquidity.

In Diagram 3 gross risk accounts for the aggregate volatility sourced from each portfolio exposure, a net measure of risk accounts for the expected relationship between risky relations (i.e. diversification benefits, typically estimated with correlations), while realised risk accounts for the actual portfolio experience. Volatility reduction is beneficial, all else equal, as lower volatility reduces the range of retirement outcomes and reduces the risk of adverse outcomes considered unacceptable to consumers.

When assessing performance outcomes, it is important to consider whether risk can be controlled for as an input or whether it needs to be considered as an output. This depends on whether realised risk moves proportionately with gross risk. This relationship is not consistent across super funds. In designing and implementing an investment strategy it is reasonable to expect that a trustee would account for the impact of risk reduction strategies and diversification strategies. A simple case study in Table 1 illustrates how different investment strategies can yield different reductions in realised volatility.

Case 1: Simple Growth Portfolio		Case 2: Diversified Growth Portfolio	
	Allocation		Allocation
Australian shares	50%	Australian shares	50%
Global shares, unhedged	25%	Global shares, unhedged	25%
Global shares, hedged	25%	Global shares, hedged	15%
		Emerging market shares, unhedged	10%
		Australian listed property	5%
		Global listed infrastructure, hedged	5%
	Measure		Measure
Portfolio gross volatility:	12.6%	Portfolio gross volatility:	12.7%
Portfolio net volatility:	11.3%	Portfolio net volatility:	10.8%
Volatility reduction:	-11.0%	Volatility reduction:	-15.3%

Table 1: Comparison of two alternative growth portfolios to compare volatility reduction (data: 30/4/2009 – 30/9/2020).

Table 1 reveals different degrees of volatility reduction across two growth portfolios. This is traditionally a part of the portfolio where volatility reduction is more difficult to achieve (due to a common equity beta factor). The potential for volatility reduction is understated here: it is likely to be substantially larger across other parts of the portfolio. Some approaches such as portfolio protection strategies are specifically designed to reduce realised volatility, while there is evidence of funds who target significant diversification benefits as a foundation of their portfolio construction.

The expectation that funds should focus on net risk outcomes when determining portfolio construction, combined with the inconsistency of volatility reduction across funds instructs us to advocate for an approach that recognises net volatility as a valuable output.

In summary, risk is important and gross risk (the gross risk of the exposures) may be an unreliable proxy for the realised net risk.

2.1.4. Persistence of investment performance

Performance persistence (whether under or outperforming funds continue to under or outperform in the future) is a subset of the active performance debate, a highly contested part of the academic literature. Much of the analysis is based on US mutual fund performance. A small number of analogies can be applied to superannuation:

- Fees are a drag on performance

- There is a case for active management which runs along the following lines:
  - Most academic research is based on US mutual funds which in aggregate experience negative returns post-fees
  - But institutional investors can typically access investment managers at fees far lower than mutual fund fees
  - Arguably US stocks represent the most efficient market and there is evidence of higher active returns in other markets
- Persistence:
  - High fees are an indicator of future poor performance
  - There is some evidence of persistence amongst the poorest performing funds, partly due to high fee loads
  - There is mixed evidence of sustained elevated levels of outperformance by individual investment managers over the long-term. While some research identifies that persistent outperformance can be explained by factor exposures rather than a unique 'skill' factor, arguments based on less efficient markets (relative to US stocks) and lower institutional fees support a case for long-term outperformance

A different research stream focusing on asset owners identifies fund governance as being strongly aligned with investment performance (a good overview by Willis Towers Watson). Clark and Urwin (2008) find that *"Good governance by institutional asset owners makes a significant incremental difference to value creation as measured by their long-term risk-adjusted rate of return"*, while Ambachtsheer (2007) suggested the return difference between a poorly governed and well-governed fund was likely in the range of 100 – 200 bp pa. While some instruments to quantitatively assess governance have been developed it remains an area that requires qualitative assessment.

Anecdotally, investment performance could be considered a low signal-to-noise activity making future prediction weak. This motivates an important reflection: will a performance test be effective? We explore this later in this paper.

A short literature review is included in Appendix 1.

2.1.5. Accounting for product design

As identified in Diagram 1 performance interacts with product design and a person's individual situation to produce a range of possible outcomes. To assess the impact on consumer outcomes would necessitate modelling this interaction and assessing the range of outcomes. The purpose of Diagram 4 is to illustrate the relative importance of investment performance at different stages of life. In this example, of an individual experiencing continual employment, performance in the final year of employment is expected to have greater than 100 times the impact on retirement balance as performance in the first year. The product-design response to this is to develop strategies which ensure a more consistent "dollar-at-risk" profile through time which represents the case for life-cycle products: higher investment risk at young ages which declines through time.

The benefits of product design were largely overlooked by the Productivity Commission who chose to focus on the traditional measure of annualised returns. Such an approach is only effective at assessing the performance of a one-off contribution. We acknowledge that any alternative through-life approach would have been complex and potentially involved multiple cohorts etc. Nonetheless the impact is significant: a submission made by Bell (2019) to APRA demonstrated how a poor performing but well-designed superannuation fund could produce equivalently valued outcomes as a good performing but basically designed super fund.

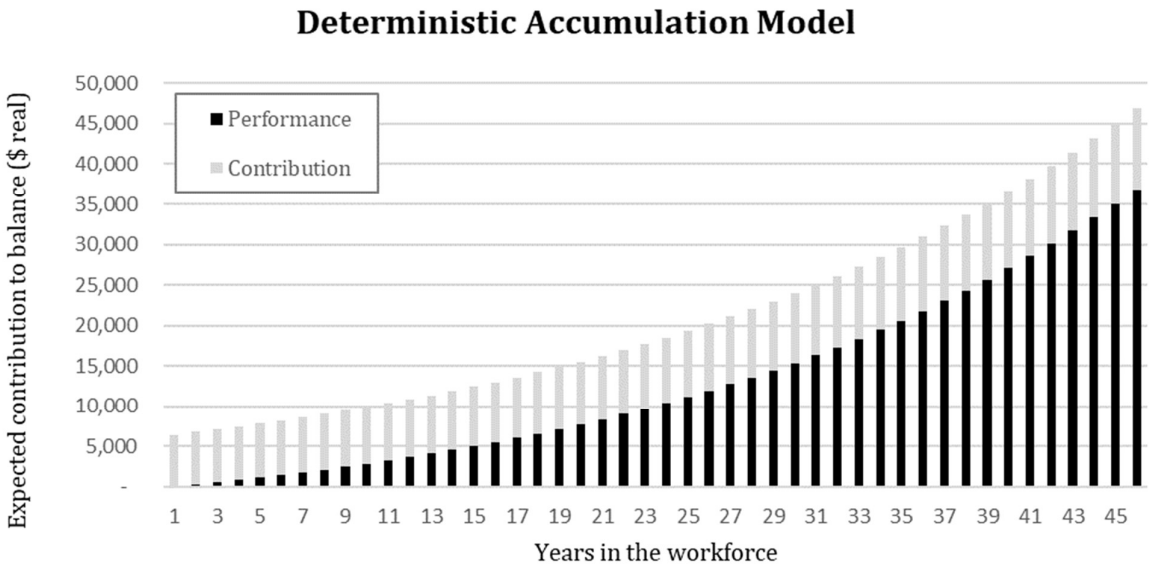

Diagram 4: Deterministic accumulation model to identify annual expected contributions to accumulation balance from investment returns and contribution payments. Assumptions: continued employment through life, initial wage of $80,000, annual wage growth of 1% pa, investment return of 4% pa real.

2.1.6. Summary – what drives consumer outcomes?

As identified in Diagram 1 consumer outcomes, even just through the accumulation stage, are a complex interaction of returns and product design applied to unique consumers. Undeniably product design is important, but unfortunately policy, reviews and regulators appear have not assessed the interaction.

This leaves us with investment returns, or more specifically future investment returns. We have demonstrated that there are three components to investment outcomes in superannuation (risk, asset allocation and implementation). Each can have a significant impact on total investment outcomes and so should be accounted for. When we question whether past returns inform future returns the evidence is limited, except in the area of total fees.

Governance is another recognised predictor of future returns amongst asset owners. We also identified that returns are not the only output of the investment process: net risk is an outcome and gross risk (an input into the return generation process, as per Diagram 3) is not a reliable estimate of net risk, especially given the strategies specifically adopted by some funds to manage their net risk profile (direct risk management trades and diversification strategies).

Ultimately of the reason for introducing the YFYS performance test is to create an environment of enhanced accountability. In this regard it is important that the test focuses on the performance outcomes that impact consumers (total investment performance), that it acknowledges net risk rather than gross risk (to include portfolio construction in the assessment) and to include all relevant fees. Incorporating qualitative analysis would allow the assessment of governance. Regarding fees we clarify that there are two categories: investment fees and administration fees. We view that a pure assessment of investment outcomes would consider only investment fees while a broader measure of consumer outcomes would include investment and administration fees. We note that if administration fees are to be included it would be appropriate to account for the range of services provided.

This is reflected in Diagram 5, which expands Diagram 3. We believe that it is important for consumers that all the outputs identified in Diagram 5 are considered in any assessment of performance.

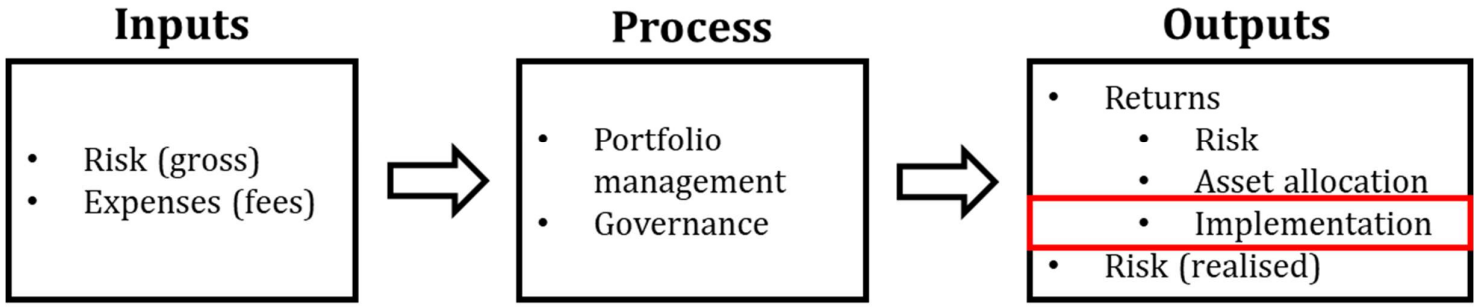

Diagram 5: Expanded process representation of investment management. The red box reflects the focus of the YFYS performance test. Note that chosen inputs (risk and expenses) account for the assessed opportunity set (expected returns, forecast volatility, correlations etc.), fund membership characteristics (demographic profile), and fund-level constraints such as liquidity.

2.2. The effectiveness of the YFYS performance metric at identifying 'poor' funds

In this section we investigate the effectiveness of the YFYS performance metric for identifying poor performing funds. The YFYS performance metric faces three major challenges detailed in Diagram 6; for each of these challenges we explore the impact on statistical effectiveness. Before this we detail our major concern with a statistical performance test designed to protect consumers from experiencing poor future performance: funds are not stationary (they make changes through time).

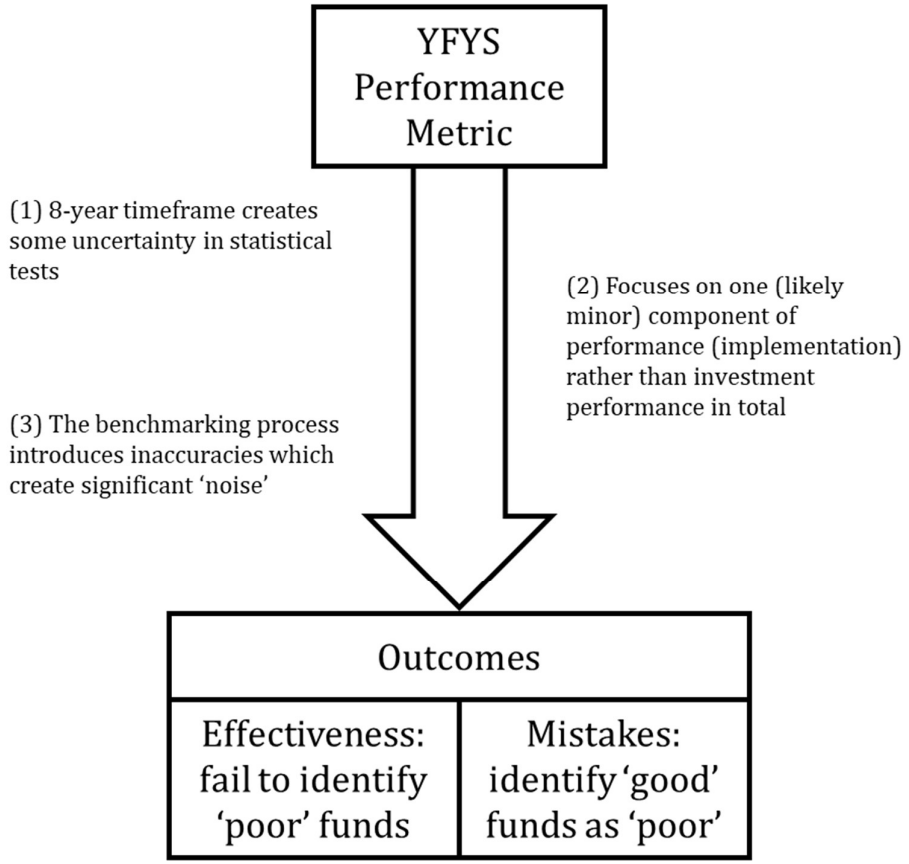

Diagram 6: Detailing concerns around the effectiveness of the YFYS performance metric.

2.2.1. Funds are not stationary

Performance tests are backwards looking and it is common practice, enforced by regulators globally, to disclose that past investment performance has limitations as a predictor of future performance. There are various reasons for this. One is the existence of investment style biases (for example value investing in stocks) which may experience prolonged periods of underperformance and outperformance. Another reason is that any trustee performing their duties would take steps to ensure that underperformance is investigated, reviewed and addressed. Examples of important changes include a review of investment beliefs and approach, changes to organisational structure, resourcing, and board composition. This means that if super funds are viewed as a process which turns risk and fees into investment outcomes (as per Diagram 3), then the process is not a stationary one. Statistically this violates the ability to draw insights from observations of past outcomes.

Diagram 7 illustrates the risk of a backward-looking performance test as a guide for future performance for consumers. Consider Fund A and Fund B as detailed in Diagram 7: if Fund A failed the YFYS performance test, we can't see how consumers would be better off outside of Fund A in the future, if we let Fund B represent the alternative.

Subsequent statistical analysis is based on the (incorrect) assumption of a stationary process (i.e. the fund is assumed to make no changes through time).

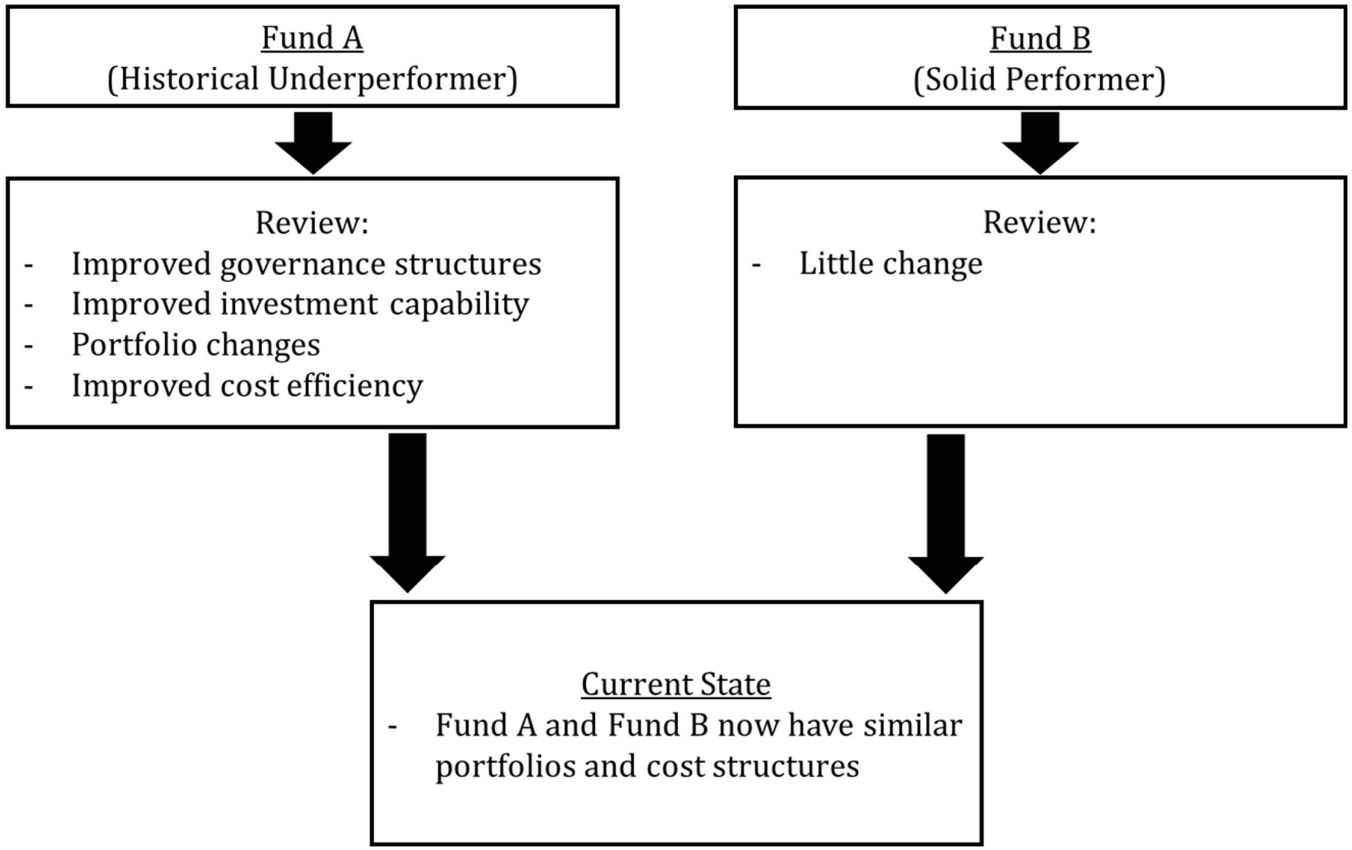

Diagram 7: Hypothetical example illustrating how two funds may reach a present position of having similar portfolio characteristics, fees and investment capabilities.

Some changes can be quantitatively observed such as reductions in administration fees. It is possible that changes of this nature can be acknowledged in a performance metric. However, changes in areas such as governance, which can also inform future performance, are best assessed qualitatively.

2.2.2. Is 8 years long enough?

All investment processes inherently experience a degree of variability. Time smooths out the variability of a process and enables us to take greater certainty from statistical findings. We assume:

- The YFYS performance test is effective at identifying investment performance (we subsequently acknowledge that the YFYS performance test is not accounting for the other two components of investment outcomes)

- The YFYS performance test has no additional noise other than the variability of the investment process i.e. the active risk of the fund (we subsequently acknowledge that the benchmarking process introduces significant external noise into the process)

Based on these assumptions we can undertake some statistical tests to assess the effectiveness of the test. These are detailed in Table 2.

Mistakes: falsely identifying a 'good' fund as a poor performer	**Effectiveness: failing to identify a 'poor' fund as a poor performer**
'Good fund' properties: - Expected excess return: 0% pa - Volatility of implementation perf.: 0.7% ann.	'Poor fund' properties: - Expected excess return: -0.75% pa - Volatility of implementation perf.: 0.7% ann.
Probability of error: 2%	Probability of error: 16%
Interpretation: For every 50 'good' funds, this test will likely mis-identify 1 as a poor performer because over 8-year intervals a good fund may experience annualised performance below the threshold level.	Interpretation: For every 6 'poor' funds, this test will likely mis-identify 1 as a good performer because over 8-year intervals a poor fund may experience annualised performance above the threshold level.

Table 2: Assessment of the effectiveness of the YFYS performance test in assessing implementation performance in a controlled environment with an eight-year window. We set the threshold level of underperformance at -0.5%. Details of how the 0.7% volatility of implementation performance are detailed in Appendix 2.

We can see from Table 2 that, due to the eight-year time horizon, even if the performance test is perfectly appropriate, it will have a reasonable probability (15.6%) of failing to identify 'poor' performing funds. The risk of mistakes (falsely identifying a 'good' fund as a poor performer) is low. A solution is to consider a longer assessment period. However, for the effectiveness to be less than 5%, the assessment period would need to be 20 years.

2.2.3. Focusing on all components of investment performance

As identified in Diagram 5 the YFYS performance test focuses only on implementation and ignores other components of investment performance. To investigate the effectiveness of the YFYS performance test in identifying 'poor' total investment performers we consider a variety of 'poor' funds where the poor performance comes from different performance components. We assume that:

- The YFYS performance test has no additional noise other than the variability of the investment process (we subsequently acknowledge that the benchmarking process introduces significant external noise into the process)

We first focus on the effectiveness of identifying 'poor' performers. Table 3 details our results.

	Case 1 A 'good' performing fund	Case 2 Underperformance expectation 100% from implementation	Case 3 Underperformance expectation split between implementation and risk / asset allocation	Case 4 Underperformance expectation 100% from risk / asset allocation
Fund properties	Implementation: - Exp. excess return: -0% pa - Volatility: 0.7% ann. Risk / Asset Allocation: - Exp. excess return: 0% pa - Volatility: 0.7% ann.	Implementation: - Exp. excess return: -0.75% pa - Volatility: 0.7% ann. Risk / Asset Allocation: - Exp. excess return: 0% pa - Volatility: 0.7% ann.	Implementation: - Exp. excess ret.: -0.375% pa - Volatility: 0.7% ann. Risk / Asset Allocation: - Exp. excess ret.: -0.375% pa - Volatility: 0.7% ann.	Implementation: - Exp. excess return: 0% pa - Volatility: 0.7% ann. Risk / Asset Allocation: - Exp. excess return: -0.75% pa - Volatility: 0.7% ann.
Type of error	Falsely identifying a fund as a poor performer when it is not	Effectiveness: failing to identify a 'poor' performer	Effectiveness: failing to identify a 'poor' performer	Effectiveness: failing to identify a 'poor' performer
Likelihood of error (est.)	2%	16%	69%	98%
Interpretation	For every 50 'good' funds, this test will falsely identify 1 as a poor performer.	For every 6 'poor' funds, this test will likely fail to identify 1.	For every 10 'poor' funds, this test will likely fail to identify around 7.	Nearly all 'poor' funds will fail to be identified.
Attribution	Variability & timeframe: 8% Implementation focus: -6% Total: 2%	Variability & timeframe: 23% Implementation focus: -7% Total: 16%	Variability & timeframe: 23% Implementation focus: 46% Total: 69%	Variability & timeframe: 23% Implementation focus: 75% Total: 98%

Table 3: Assessment of the effectiveness of the YFYS performance test in assessing total investment performance in a controlled environment with an eight-year window. To explain attribution: (1) variability and timeframe: the general statistical reliability of the test (the Effectiveness Test in Table 2), (2) implementation focus: by using a test which only focuses on implementation do we miss identifying total underperformers? Note if this number is negative it reflects coincidence: the narrow implementation test identified negative implementation performance, ignoring the variability from risk / asset allocation.

We can see from Table 3 that, regardless of the source of underperformance, the test has low reliability of identifying underperforming funds over an 8yr timeframe. The reason is two-fold:

1. The existence of multiple sources of performance generate additional variability to the process.
2. The YFYS performance metric is designed to identify poor implementers, so it is intuitive that it will find it difficult to identify poor performers driven by their risk / asset allocation process. The small success (approximately 2%) in the final column of Table 3 is purely coincidence.

2.2.4. Accounting for benchmarking 'noise'

The YFYS performance test compares the performance of a portfolio against a tailored collection of public market benchmarks. Unfortunately, there are a range of investments which are not accurately benchmarked. Example 2 explains our analysis, Table 4 presents a short summary of our findings (which are not exhaustive), while Appendix 3 contains the full analysis.

Example 2

In the YFYS performance test unlisted property is benchmarked against a listed property index. There can be large variations between the performance of the two, independent of implementation performance. Historically, over rolling 8-year periods we found the difference to be as large as 9.3% pa. This means that a 10% allocation to unlisted property could have had an impact on the performance test of >90bp pa.

Investment	Historical Size of Impact
Private Equity	5% allocated to private equity 3yrs ago could have an impact of 20bp pa once spread over the 8yr period[5].
Unlisted Property	A 10% allocation to unlisted property could have resulted in a fund performance difference of >90bp pa relative to the YFYS benchmark.
Unlisted Infrastructure	A 10% allocation to unlisted infrastructure could have resulted in a difference of >100bp pa relative to the YFYS benchmark.
Credit (including high yield, direct lending activities, and emerging markets debt)	A 10% allocation to high yield bonds could have resulted in a difference of >90bp pa relative to the YFYS benchmark.
Australian inflation-linked bonds	A 10% allocation to inflation-linked bonds could have resulted in a difference of >20bp pa relative to the YFYS benchmark.
Other (including alternative assets)	A 10% allocation to low risk rather than high risk alternatives could have resulted in a tailored benchmark difference of approx. 20bp pa.

Table 4: Summary of the impact of some of the benchmarking challenges in the YFYS performance test.

Based on Example 2 and Table 4 we make the following observations:

- At times the historical impact from many of these exposures would individually have exceeded the YFYS performance test threshold, independent of implementation performance.

[5] Calculated as three years of listed equity returns while early stage private equity returns were flat.

- In recent times exposures to unlisted property, unlisted infrastructure and credit would have added significantly to performance. This creates the risk that the YFYS performance test inadvertently favours one cohort of investor-type over another.

- The treatment of inflation-linked bonds is concerning. These are widely considered the most appropriate asset for managing the liability risk yet funds would incur 'benchmark noise' under the YFYS test for using these bonds.

Given the identification of significant degrees of 'benchmark noise' it is important to reassess the effectiveness of the YFYS performance test. We focus only on the test's ability to assess implementation performance (i.e. we leave out the other components of investment performance explored in 2.2.3). We make the following assumption:

- 'Benchmark noise' brings 3.6% pa volatility into the test. This is based off a portfolio with allocations of 10% to unlisted property, 10% to unlisted infrastructure and 10% to high yield. Full details of the calculation are included in Appendix 4.

Table 5 details our results.

Mistakes: falsely identifying a fund as a poor performer when it is not	**Effectiveness: failing to identify a fund as a poor performer when it is**
'Good fund' properties: - Expected excess return: 0% pa - Volatility of implementation alpha: 0.7% ann. - Volatility of noise: 3.6% pa	'Poor fund' properties: - Expected excess return: -0.75% pa - Volatility of implementation alpha: 0.7% ann. - Volatility of noise: 3.6% pa
Probability of error: 35%	Probability of error: 42%
Interpretation: For every 3 'good' funds, this test will likely mistakenly identify 1 as a poor performer because of the noise introduced by the benchmarking process.	Interpretation: For every 5 'poor' funds, this test will likely mis-identify 2 as good performers because of the noise introduced by the benchmarking process.

Table 5: Assessment of the effectiveness of the YFYS performance test in assessing implementation performance where we incorporate the noise impact due to benchmarking shortcomings.

Table 5 details the impact of the noise created by benchmarking issues inherent in the YFYS performance test. The results in Table 5 can be compared directly with those in Table 2, thereby being able to isolate the reduction in test performance due to benchmarking noise. We can identify that benchmarking noise raises the likelihood of mistakes (falsely identifying a 'good' fund as a poor performer) by 33% while making it 26% more likely that the test will fail to identify a 'poor' fund as underperforming. In the presence of substantial benchmarking noise, the test begins to approach the same degree of effectiveness as a coin toss.

2.2.5. Incorporating all of the issues and assessing overall effectiveness

Bringing together the three issues identified in Diagram 6 we can assess the overall effectiveness of the YFYS performance test at identifying historically 'good' and 'bad' funds. We account for the following:

- The 8-year timeframe
- Total investment performance (i.e. accounting for risk / asset allocation)
- The benchmarking noise inherent in the YFYS performance test

Table 6 details our assessment.

	Case 1 A 'good' performing fund	Case 2 Underperformance expectation 100% from implementation	Case 3 Underperformance expectation split between implementation and risk / asset allocation	Case 4 Underperformance expectation 100% from risk / asset allocation
Fund properties	Implementation - Exp. excess return: 0% pa - Volatility: 0.7% ann. - Noise volatility: 3.6% ann. Risk / Asset Allocation - Exp. excess return: 0% pa - Volatility: 0.7% ann.	Implementation - Exp. excess return: -0.75% pa - Volatility: 0.7% ann. - Noise volatility: 3.6% ann. Risk / Asset Allocation - Exp. excess return: 0% pa - Volatility: 0.7% ann.	Implementation - Exp. excess ret.: -0.375% pa - Volatility: 0.7% ann. - Noise volatility: 3.6% ann. Risk / Asset Allocation - Exp. excess ret.: -0.375% pa - Volatility: 0.7% ann.	Implementation - Exp. excess return: 0% pa - Volatility: 0.7% ann. - Noise volatility: 3.6% ann. Risk / Asset Allocation - Exp. excess return: -0.75% pa - Volatility: 0.7% ann.
Type of error	Falsely identifying a fund as a poor performer when it is not	Failing to identify a poor performer	Failing to identify a poor performer	Failing to identify a poor performer
Likelihood of error (est.)	35%	42%	54%	65%
Interpretation	For every 3 'good' funds, this test will falsely identify 1 as a poor performer.	For every 5 'poor' funds, this test will likely mis-identify 2 as good performers.	For every 2 'poor' funds, this test will likely mis-identify 1 as a good performer.	For every 3 'poor' funds, this test will likely mis-identify 2 as a good performer.

Attribution	Variability & timeframe: 35% Implementation focus: 0% Total: 35%	Variability & timeframe: 42% Implementation focus: 0% Total: 42%	Variability & timeframe: 42% Implementation focus: 12% Total: 54%	Variability & timeframe: 42% Implementation focus: 23% Total: 65%

Table 6: Assessment of the overall effectiveness of the YFYS performance test in assessing total investment performance where we account for the eight-year timeframe, multiple sources of investment performance and noise from ineffective benchmarks. To explain attribution: (1) variability and timeframe: the general statistical reliability of the test (the Effectiveness Test in Table 2), (2) implementation focus: by using a test which only focuses on implementation do we miss identifying total underperformers? Note if this number is negative it reflects coincidence: the narrow implementation test identified negative implementation performance, ignoring the variability from risk / asset allocation.

Table 6 produces results which reinforce the observation made in 2.2.4: a test with weak effectiveness and a high mistake rate. By comparing the results in Table 6 with those in Table 3 we can see that all results trend towards 50% (the expected result of a coin-toss), implying the test has little effectiveness.

2.2.6. Summary: YFYS test effectiveness

Diagram 6 details three areas of concern relating to the effectiveness of the YFYS performance test:

1. Whether 8 years is an appropriate length of time for the YFYS test to be effectively assess implementation performance
2. Whether the YFYS test can provide insight into the assessment of total investment performance which accounts for risk / asset allocation
3. Whether the listed benchmark approach in the YFYS performance test creates benchmarking noise which impairs the effectiveness of the YFYS performance test

We found that, when we aggregate these issues, the YFYS performance test is likely to have weak effectiveness at identifying poor total investment performers across a range of different types of super funds. Even if policymakers decide to focus solely on implementation the existence of significant benchmarking noise renders the test highly unreliable.

Timeframe proved to be only a minor issue, with the other two issues (total investment performance and benchmarking noise) proving significant.

We have additional concerns that, if applied today, the test favours certain styles of funds, likely to a multiple of the threshold level (50bp). Further the benchmarking issues create a cohort risk where certain types of funds may all fail at the same time, independent of their implementation performance.

3. Undesirable outcomes

We believe the YFYS performance test will result in undesirable outcomes relating to how funds invest, direct impact on consumers, and industry structure. Our concerns are summarised in Table 7 and we explore each of these issues in further detail.

Concern 1: How funds will invest	Concern 2: Direct impact on consumers	Concern 3: Impact on industry structure
- Distortion to portfolio management behaviour. Focus on managing tracking error to performance benchmark and through-time performance management will make trustees more short-term focused and make a range of 'noisy' investment strategies less attractive. We expect this to increase portfolio turnover costs and reduce portfolio quality. - Dangerous incentive for funds which are well behind on the performance test to 'swing for home runs' and take high tracking error relative to benchmark. - Actively managing the YFYS performance test by taking advantage of benchmark shortcomings. - Poor alignment with portfolio management approaches such as total portfolio approach (TPA).	- Given the low effectiveness of the test, super funds may 'contest' the result with their members, creating confusion. - The YFYS performance test result may create confusion for consumers when placed alongside total performance on the YFYS Comparison Tool. - Potential for a large cohort of funds to fail the YFYS test concurrently (due to benchmarking noise), reducing system confidence. - Doesn't remove consumers from investment products with assessed high administration fees. - Penalises the heavily disengaged who may remain in a fund which becomes more impaired.	- A deterrent to consolidation as funds will be hesitant to merge with other funds which may dilute their portfolio quality, impair their inflow profile, or distract management focus. - Potential 'zombie' funds which are impaired, making them an unattractive merger partner.

- Deterrent to strategies which reduce risk and provide diversification. - Management of ESG risk creates more benchmark "tracking error". - Potential for reduced investment in Australian unlisted assets. - Features of the YFYS performance test do not match up well with future portfolio management challenges.		

Table 7: Summary of undesirable outcomes likely to result from the YFYS performance test. For more details see the Long Paper.

3.1. How will funds invest in the presence of the YFYS performance test?

It is reasonable to assume that most trustees will seek to avoid failing the test at all costs. We expect a range of impacts on the way that super fund portfolios are managed.

3.1.1. Manage tracking error relative to YFYS performance benchmark

Funds are likely to manage the tracking error of their portfolios relative to the YFYS performance benchmark. This means reducing exposure to sectors where there exists benchmark noise. We explore this in Table 8 where we extend the analysis performed in 2.2.4 to explore the trade-off between allocations to sectors with large benchmark noise and the resulting risk of being mistakenly identified as a poor performer.

	Case Study 1	Case Study 2	Case Study 3	Case Study 4
Portfolio allocations to unlisted property, unlisted infrastructure and high yield	2% to each	5% to each	10% to each	15% to each
Estimated benchmarking noise	0.7% ann.	1.8% ann.	3.6% ann.	5.4% ann.
Likelihood of being mistakenly identified as a poor performer (independent of actual implementation performance)	8%	23%	35%	40%

Table 8: Possibility of being mistakenly identified as a poor performer by the YFYS performance test due to benchmarking noise across portfolios with varying asset allocations to inaccurately benchmarked assets. We assume in each case study that the fund is a 'good' performer with an expected excess return of 0% pa.

Table 8 raises the prospect that funds may significantly reduce their exposures to assets not accurately benchmarked. We make no claims about the direct return potential of these assets but note that this potentially constrains the opportunity set of portfolios and reduces the potential to improve portfolio diversification.

Other industry participants raise issues around passive investing. The active / passive debate is ongoing and subjective and we don't re-visit it here.

3.1.2. Through-time performance management

Given anticipated trustee concerns of failing the test, it is not unreasonable to expect that some trustees may adopt through-time portfolio management strategies which take account of their accrued under or over-performance relative to the YFYS benchmark. This has the potential to constrain portfolio management decisions independent of the outlook for returns. In Table 9 we provide examples to define this. In both case studies the trustee wants to be 95% certain that they will not fail the performance test assuming they have 'good' investment capabilities.

	Case Study 1 – poor past performance	Case Study 2 – good past performance
Scenario: excess performance over the last 6 years	-0.5% pa	0% pa
Performance (excess) required over the next 2 years to fail the 8-year test	-0.5% pa	-2% pa
Tracking error to the YFYS benchmark to be 95% certain of not failing the performance test (assuming 'good' investment capability)	0.4%	1.7%

Table 9: Tracking error that affords a 'good' fund 95% certainty of not failing performance test given 6 years of known performance. Tracking error accounts for both active management and the benchmarking noise which comes from benchmarking noise. We assume a 'good' performer has an expected excess return of 0% pa.

Table 9 illustrates the difference in operating risk budgets that may exist across funds. If all funds undertook activities to actively managed their risk of failing the YFYS performance test, then:

- At any point in time different funds will have varying degrees of ability to diversify their portfolios and best reflect their convictions
- For any fund the ability to diversify and reflect convictions will vary through time (as they adjust to their realised performance)

We believe that such through-time portfolio management activities will increase portfolio turnover costs and result in lower quality portfolios.

3.1.3. Dangerous incentive to 'swing for home runs'

There is a dangerous incentive for funds which are well behind on the performance test to 'swing for home runs', meaning they target a large degree of benchmark-relative risk in the hope that they generate sufficiently strong short-term performance against the YFYS criteria to pass the rolling 8-year test. We explore this in Table 10.

	Case Study 1 – low YFYS benchmark risk	Case Study 2 – high YFYS benchmark risk	Case Study 3 – very high YFYS benchmark risk
Scenario: performance over the last 6 years	-1% pa	-1% pa	-1% pa
Performance required over the next 2 years to pass the 8-year test	1% pa	1% pa	1% pa
Tracking error target	0.5%	5%	10%
Likelihood of a 'poor' fund failing the YFYS performance test	0%	31%	40%

Table 10: Likelihood of a 'poor' fund gaining achieving sufficient performance in the last two years to pass the YFYS performance test. We assume the expected return of a 'poor' fund is -0.75% pa.

In Table 10 we observe that for a 'poor' fund which is well behind, a low-risk strategy makes failing the YFYS performance test almost certain. However, if the same 'poor' fund increases tracking error significantly then a reasonable probability emerges that, through chance alone, the fund will pass the YFYS performance test (the limit is 50% as the tracking error target approaches a very high number). We have large concerns that such risk taking is far from consumer best interests.

3.1.4. Actively managing the performance test

While perhaps 'gaming' is an unfair term, we believe that the YFYS performance test will incentivise funds to actively manage the performance test. We identify a number of opportunities where funds are incentivised to focus on the impact of investment decisions on the YFYS performance test rather than total performance outcomes to consumers. Most of these opportunities emanate from two issues: (1) a performance test which only looks at implementation performance rather than total performance, and (2) a number of benchmarks which are inappropriate and create noise. Exposures which embed market exposure while being assessed against a benchmark which reflects a lower degree of market exposure will be expected to outperform over the long-term.

We consider it important to acknowledge these active management opportunities as the alternative is that they are known to only a few who can then exploit them. This provides some chance to develop a more effective test or a policing process (though there is nothing illegal about these strategies – they are an incentive generated by the flaws in the YFYS performance test).

The essence of the opportunity is to devise techniques that achieve extra non-benchmarked exposure to assets with high expected returns (such as equities). Such approaches bring a trade-off: higher long-term expected outperformance but significant tracking error (benchmark noise). Table 11 details some of the specific ways that the YFYS performance metric could be actively managed.

Active management opportunity	Description	Impact
Equities – introduce a small degree of shorting	By introducing a small degree of shorting to equity mandates, they could be re-categorised as hedge funds which sit in alternatives and be assessed against a 50/50 mix of blended equities and global fixed income, a lower hurdle over the long-term.	Over the long-term we would expect running equity mandates with 90% exposure to equities would generate a relative-to-benchmark benefit of 1.6% pa. Applying this to 40% of a super fund portfolio would be expected to improve long-term benchmark-relative performance by 64bp pa.
Credit – 'free ride'	Allocate to credit, even as an overlay on top of existing more traditional mandates. Net credit spread performance is assessed against a zero benchmark.	Higher yielding credit strategies may generate 2.5% pa higher returns than a composite fixed income benchmark over the long-term. Allocating 10% to higher yielding credit would be expected to improve long-term benchmark-relative performance by 25bp pa.
Alternatives – create "highly flexible fixed income mandates"	For lower risk alternatives, formalise their benchmark to be against cash or bonds and benchmark them against global bond performance, expected to be a lower hurdle over the long-term.	Benchmarking lower-risk alternatives against fixed income would lower the expected long-term benchmark return benchmark by 2% pa, effectively increasing performance relative to benchmark by the same degree. Switching 10% in alternatives to "highly flexible fixed income mandates" would be expected to improve the performance test by 20bp pa.

Table 11: Examples of opportunities to actively manage the YFYS performance tests. Our long-term assumptions: that equities outperform bonds by 4% pa, and that higher-yielding credit outperforms composite bonds by 2.5% pa.

3.1.5. Interferes with total portfolio approach (TPA)

The approach interferes with a total portfolio approach (TPA) to portfolio management, whereby funds focus on portfolio outcomes with less focus on benchmark-relative returns (SAA approach). Diagram 8 explains characteristic differences between the two approaches.

Performance assessed vs.	Benchmarks	Fund goals
Success measured by:	Relative value added	Total fund return
Opportunities for investment defined by:	Asset classes	Contribution to total portfolio outcome
Diversification principally via:	Asset classes	Risk factors
Asset allocation determined by a:	Board-centric process	CIO-centric process
Portfolio implemented by:	Multiple teams competing for capital	One team collaborating together

Diagram 8: Differentiating focus between SAA (strategic asset allocation) and TPA (total portfolio approach) portfolio management approaches. Source: *Total Portfolio Approach (TPA) by the Thinking Ahead Institute*.

How does the YFYS performance test interfere with TPA? Simply, because it looks at one component of performance (implementation) and not the total investment performance delivered for consumers. Further there is the benchmarking noise challenge. For context, recall Example 1 where we identified that high yield bonds could result in good investment outcomes for consumers but may entail accepting poor implementation performance. In addition to those assets identified in Table 4 we flag low volatility equity strategies, low-duration bond strategies and even government bonds themselves (they are assessed against a composite bond index and investment grade credit will likely outperform over the long-term) as components of possible portfolio strategies designed to deliver total outcomes amidst, possibly, negative implementation performance.

It is difficult to estimate the performance benefits of TPA; it is a subjective issue. A global study of asset owners undertaken by The Thinking Ahead Institute in conjunction with NSW Treasury Corporation suggested that the average expected uplift in annual long-term performance of a TPA-based approach relative to an SAA approach amongst those surveyed to be 0.55% pa.

3.1.6. Deterrent to strategies which reduce risk and provide diversification

Dedicated risk management strategies such as portfolio protection are punitively treated by the performance test as:

- Their impact (which is really a component of risk / asset allocation) is part of investment performance but no benchmark is included in the YFYS performance metric. In this regard, the YFYS performance test uniquely assesses the investment performance of portfolio protection decisions whereas for every other portfolio decision it focuses on implementation.

- The YFYS performance test places no value on any risk reduction benefits provided by dedicated portfolio protection strategies.

The second point also applies to the benefits of diversification. In Table 1 we established the case that diversification benefits will differ across funds and that lower risk is a benefit to consumers. But risk reduction is not acknowledged in the YFYS performance test.

While trustees have a duty to manage risk and return on behalf of their members, we believe that, due to the YFYS performance test failing to acknowledge risk reduction, trustees may place less focus on risk reduction strategies and portfolio diversification.

3.1.7. Management of ESG risks creates benchmark risk

Any form of risk management activity which creates active decisions within investment sectors or leads to allocations to strategies with benchmark noise creates YFYS-benchmark risk for trustees. Managing for ESG risks is a case in point. To appropriately manage for ESG risks may mean significantly departing from benchmark portfolios within all major investment sectors. Trustees, concerned about the risk of failing the YFYS performance test may be hesitant to take the required degree of YFYS-benchmark risk to appropriately manage ESG risk.

3.1.8. Potential for reduced investment in Australian unlisted assets

We acknowledge that the role of superannuation is to efficiently manage contributions into an accumulation balance and ultimately income in retirement. One outcome of the size of superannuation has been the sizable investments into Australian unlisted assets such as infrastructure and

property. The YFYS performance test will create a disincentive to invest in these assets. As illustrated in Table 8 even a small allocation to such assets increases benchmark noise significantly, greatly increasing the chances of a 'good' fund being mistakenly identified as a 'poor' fund.

3.1.9. YFYS performance metric doesn't match up well with future challenges

Some of the future investment challenges facing trustees of super funds include:

- Low expected returns from fixed income
- How to protect against equity volatility – will bonds provide the protection that they have in the past?
- Retirement investment solutions

Some of the potential solutions to these investment challenges include:

- Greater use of low volatility alternatives
- Dedicated portfolio protection strategies
- Greater use of credit
- More heavily diversified portfolios

Each of the potential solutions detailed above (acknowledging that none are certain solutions) are unfairly treated under the YFYS performance test, either by not being scored appropriately, creating high benchmark noise, or not having the benefits of risk reduction recognised.

3.2. How will the YFYS performance test directly impact consumers?

The YFYS performance test requires a degree of member engagement for existing members to take action. The level of engagement will determine the success of the YFYS performance test (assuming it successfully identifies poor performing funds). In this regard we have a number of concerns which we detail below.

3.2.1. Funds will have a reasonable right to contest the result with their members

Given the number of concerns around the accuracy of the YFYS performance metric as an effective identifier of underperformance (outlined in this paper), we think super funds will have a valid basis on which to respond to the underperformance notification they will be required to provide members. An example may be that their prior APRA Heatmap score wasn't red or that their research house rating was "high"/positive. Both approaches consider more information than the YFYS performance metric. Funds could also detail the changes they have made to improve through time (something not addressed by the YFYS performance test).

In aggregate we think consumers will likely be confused by two sets of (likely conflicting) information. This may not prove to be a healthy basis for improving member engagement and undermine consumer confidence in the industry.

3.2.2. Confusing interaction with Comparison Tool

The YFYS package also detailed creation of a Comparison Tool. From the detail provided this will focus on fees and performance and highlight funds which have failed the YFYS performance test. As identified in section 2, there is a reasonable possibility of funds failing (passing) the YFYS performance test and delivering high (low) performance. This has the potential to create large confusion for consumers.

3.2.3. The disengaged may be punished even further

We have a concern that those who do not engage may be adversely impacted. Consider a fund that experiences loss of inflows and member outflows as a result of poor performance on the YFYS performance test. This fund may have a diminished cashflow position whereby any illiquid impaired positions become a larger part of the portfolio, potentially resulting in subsequent underperformance.

3.2.4. Fails to direct consumers away from funds with high assessed administration fees

High assessed administration fees are an acknowledged predictor of future underperformance in the sense that persistent high fees will erode member balances with certainty. By 'assessed', we mean administration fees relative to the services provided. By failing to incorporate administration fees the YFYS performance metric does nothing to move consumers away from expensive funds.

3.2.5. Consumer confidence could be damaged by the potential scenario of a large cohort of funds failing the YFYS performance test concurrently

As identified through section 2.2 there is a large cohort of funds which have some similar characteristics (sizable allocations to unlisted property, unlisted infrastructure and credit) that generate sizable benchmarking noise. In some market environments this noise may work against funds creating a risk that a sizable number of funds fail at the same time. We would expect that this would reduce consumer confidence in superannuation.

3.3. How will the YFYS performance test impact industry structure?

Despite APRA being of the view that there is a merger partner for every fund, we are concerned that this may not be the case. Reasons include impact on performance, impact on cashflow and firmwide distraction from other strategic activities. We are anecdotally aware of funds which are being overlooked for these reasons.

There is a risk that the YFYS performance test may further distort industry structure.

3.3.1. The YFYS performance test will be an additional deterrent to consolidation

A super fund which has underperformed or is carrying some illiquid impaired assets will be an unattractive partner for another fund through the lens of the YFYS performance test. The underperforming fund may also may be faced with structural outflow, which would dilute the inflow profile of the senior fund. Even if the merger has some longer-term benefits, the YFYS performance test is likely to generate a heightened focus on the shorter-term impacts of mergers along with a raised bar for taking on activities which create distraction.

3.3.2. The YFYS performance test will potentially create 'zombie' funds

Funds which are unattractive as merger partners may become zombie funds. They are likely to be experiencing net outflows and may have some illiquid assets which are impaired. In the absence of mergers, it is difficult to wind-up a super fund. At present there appears no strategy for how the tail of the industry will be consolidated. The YFYS performance test potentially further hinders industry consolidation by making some mergers less attractive.

4. Solutions

There is merit in providing consumer protection via a performance test. However, it needs to be an effective performance test, and one with limited undesirable outcomes. In this section we consider a range of different solutions, each of which we believe would improve consumer outcomes relative to the YFYS performance test. We analyse each solution through a set of principles centred on consumer outcomes.

In section 2.1 we explored what drives consumer outcomes in superannuation. We detailed in Diagram 1 a complex interaction between investment outcomes and product design, through the experience of a unique individual. While in our view this should be the focus, we acknowledge that Government wants an investment performance test and is following the recommendation of the Productivity Commission. Accordingly, we focus on developing a high-quality investment performance test. We also consider an alternative, a consumer investment outcome test, which additionally accounts for administration fees.

4.1. Principles of an investment performance test

To develop a set of principles which guide our recommended solutions, we reflect on our learnings thus far.

First, we identified that there are three components which contribute to investment outcomes in superannuation (risk, asset allocation and implementation – see Diagram 2). Each can have a significant impact on total investment outcomes and so should be accounted for. And in Diagram 3 we recognised that realised return and risk are both outcomes of an investment process. We acknowledged that risk is important and identified that gross risk may be an unreliable proxy for the realised net risk.

Given the intent of the performance test is to protect consumers from experiencing poor performance in the future we investigated what informs future performance. Here we found that past return provides little insight into future returns. Evidence does exist that total fees informs future performance (especially underperformance) as does good governance. While fees can be objectively identified, governance requires a degree of qualitative assessment.

Our design principles for any performance test are simple and are summarised in Diagram 9. We make three reflections:

1. Under the first principle we account for the non-stationary issues in section 2.2.1. The impact of changes made by funds through time may directly impact future consumer outcomes and it is important that they are accounted for

2. The second set of principles ("Assesses total investment performance") provides little insight into future performance. Effectively it is a principle of accuracy: any assessment of investment performance should fully acknowledge all aspects of investment outcomes.

3. Total fees, incorporating investment fees and administration fees, are only considered in a consumer outcome test, not a pure investment test.

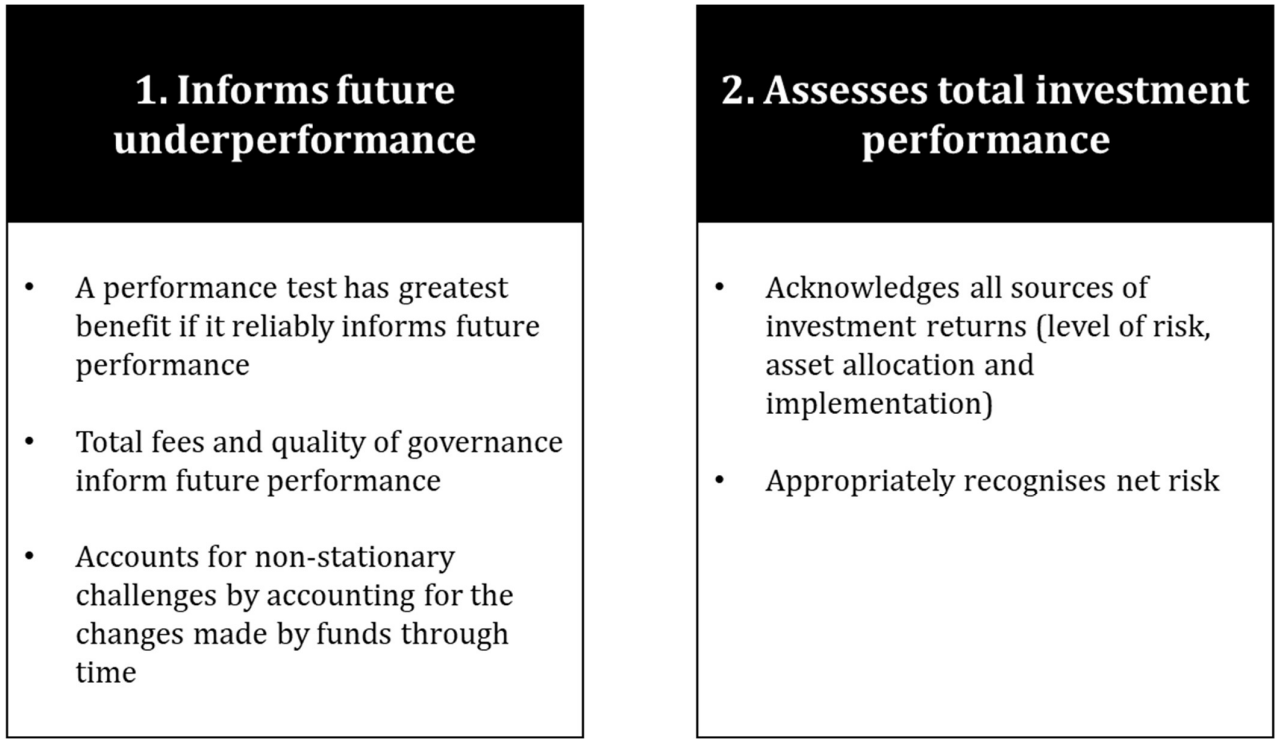

Diagram 9: Principles for designing a performance test which will protect consumers from experiencing poor performance in the future.

4.2. Addressing the principles to develop an effective performance test

To address the principles outlined in Diagram 9 we need to resolve a number of issues. Table 12 details solution techniques to address the principles.

Assessment challenge	Assessment solution techniques
Total fees (informs future returns)	- Total fees account for investment and administration fees (adjusted for services provided). Administration fees, hence total fees, would only be considered in a consumer outcomes test - Past returns can be calculated which account for administration fees - 'Adjusted' past returns can be calculated which account for past net investment returns and current 'assessed' administration fees, thereby accounting for changes made through time (i.e. accounting for administration fee changes made through time)
Governance (informs future returns)	- Some systematic approaches have been developed to enable a high-level assessment of governance - Nonetheless we view a qualitative assessment technique as more effective for assessing quality of governance - A qualitative approach can account for changes made through time
Acknowledges all sources of return (level of risk, asset allocation and implementation)	- This can be achieved by analysing total returns (as per the YFYS performance metric), but requires consideration of how results are normalised: o For example, the YFYS performance metric normalises by the asset allocation of the fund. The outcome of this approach is that the value of asset allocation decisions is not acknowledged o Normalising realised returns by measures of total net risk accounts for all investment activities

Acknowledges the net risk profile (gross risk measures are not a reliable proxy for net risk outcomes)	- Possible solutions include:
	  o Estimating ex-ante portfolio diversification benefits (this is difficult as it involves the subjective estimation of correlations)
	  o Using realised volatility means we are accounting for realised asset volatility and diversification benefits in a similar manner as to how realised returns are assessed

Table 12: Solution techniques to address the principles for an effective performance test outlined in Diagram 9.

4.3. A family of solutions

In this section we assess a range of candidate solutions, detailed below:

1. A solution which incorporates regulator assessment: incorporating metric-based analysis, deeper insights produced by the regulator, and deliberation on self-assessment produced by funds themselves. This permits consideration of many challenges such as those created by the through-time changes made by super funds, as well as the opportunity to normalise administration fees for services provided. We note that the Productivity Commission suggested APRA perform a role of this nature in certain situations.

2. Multiple metrics: a collection of metrics designed to form a high-quality investment performance test.

3. Single metric: the best that can be achieved with a single metric as an investment performance test.

4. YFYS-reviewed: includes adjustments to recognise:
   - Tailored benchmarks to account for unlisted assets (as per the Productivity Commission's BP2)
   - A risk scaling approach as proposed by the Growth / Defensive Working Group to determine an appropriate benchmark for assets in "Other"
   - Remove dedicated portfolio protection strategies from fund performance
   - Incorporation of administration fees if the focus is consumer outcomes (preferably current administration fees assessed for the services provided)

5. YFYS (current form).

We assess each candidate solution through a range of criteria, namely:

- Recognises member outcomes over lifetime – i.e. the interaction described in Diagram 1
- Could be designed to be a pure investment metric or a consumer outcome metric (which incorporates administration fees)
- Provides insight into two areas which inform future underperformance:
    1. High total fees, particularly administration fees (relative to the services provided)
    2. Weak governance
- Focuses on total investment outcomes delivered to consumers
- Acknowledges the net risk outcome
- Accounts for non-stationarity of funds – the issue represented in Diagram 7
- Results in statistical effectiveness at identifying historic under or outperformers
- The potential for, and size of, undesirable outcomes (as explored in section 3)

Our assessment is detailed in Table 13.

	Solution incorporating regulator assessment	**Multiple metrics**	**Single metric**	**YFYS (adjusted)**	**YFYS (proposed)**
Recognise member outcomes over lifetime	Possibility, but currently does not consider	Possible to develop	Not in a single metric	Does not consider	Does not consider
Could be designed to be a pure investment test or a consumer outcome test	Either	Either	Either	Either	Pure investment test
Acknowledges high current fees	A regulator would have insight	Could be incorporated	Could be incorporated	Could be incorporated	Does not consider

Acknowledges weak governance	A regulator would have insight	Would not consider	Would not consider	Would not consider	Does not consider
Focus on total investment outcomes	A regulator would have insight	Would address	Would address	Does not consider	Does not consider
Acknowledges net risk outcomes	Possibly, but regulators do not currently assess net risk	Would address	Would address	Would not address	Does not consider
Accounts for non-stationarity	A regulator would have insight	Could only account for fee changes	Could only account for fee changes	Could be incorporated	Does not consider
Statistical effectiveness	Not applicable	Reasonable effectiveness – similar to Table 2	Reasonable effectiveness – similar to Table 2	Weak effectiveness in some situations – similar to Table 3.	Weak effectiveness – similar to Table 6.
Potential for undesirable outcomes	Risk of regulatory capture	A smaller range of issues will still exist relating to investment management, impact on consumers, and industry structure	A range of issues will still exist relating to investment management, impact on consumers, and industry structure	Large range of issues will still exist relating to investment management, impact on consumers, and industry structure	Very large range of issues will still exist relating to investment management, impact on consumers, and industry structure

Table 13: Assessment of candidate solutions.

When we consider the assessment provided in Table 13 we make a number of observations:

- A solution which incorporates regulator assessment has the greatest potential to be effective. It affords the opportunity to recognise total outcomes to consumers, assess total investment performance, and account for the through-time changes made by funds. It allows the integration of both multi-metric-based and qualitative assessment techniques (in areas such as governance).

- A collection of metrics, all else equal, is superior to an individual metric. Any individual metric will have shortcomings and these can be reduced through the judicious use of additional metrics. However, metrics can only go so far: they cannot always account for changes made through time while some areas such as governance are best assessed qualitatively.

- There are alternative individual metrics which we believe are superior to the YFYS performance test. These alternative metrics better account for total investment outcomes delivered to consumers and a smaller degree of undesirable outcomes

- The YFYS performance metric can be improved. At best it can be more stable and assess all fund-types consistently. Nonetheless, it will remain statistically ineffective at assessing implementation performance. It fails to focus on total investment performance to consumers. The YFYS metric, even if improved, will create a large range of undesirable outcomes.

Designing a single metric to assess performance is a controversial and highly subjective area. We have undertaken some initial work which is detailed in Appendix 5. Further research is required before we label this as a 'recommended solution'. Indeed, given our above analysis, our recommended solution would not be a single metric. Accordingly, it simply provides a point of comparison with the YFYS performance test.

5. Conclusion

This Detailed Paper presents a substantial analysis of the YFYS performance test. There is merit in protecting consumers with a performance test. However, it needs to be an effective performance test with limited undesirable outcomes. Unfortunately, our analysis suggests the YFYS performance test does not meet these goals: it will be ineffective at identifying poor performing funds while introducing a range of undesirable outcomes. We are concerned that the detriments of the YFYS performance test may outweigh the benefits.

To summarise our analysis:

- An investment performance metric has shortcomings: it doesn't account for the total consumer outcome (by considering product design), fails to account for changes made through time by super funds, while the evidence that past performance is a predictor of future performance (performance persistence) is modest (total fees and governance appear informers of future outcomes).

- Placing these issues aside we assessed the YFYS performance metric. The statistical effectiveness of the YFYS performance metric at identifying poor performing funds is found to be weak. The YFYS performance metric faces three major challenges: (1) timeframe (8 years may not be sufficient), (2) it focuses on one (likely minor) component of performance (implementation) rather than investment performance in total, and (3) the benchmarking process generates inaccuracies which create benchmark 'noise'. Our statistical analysis reveals that, in its current form, the performance metric will be ineffective at identifying poor performers and have a high likelihood of falsely identifying good performers as poor.

- We believe the YFYS performance test will result in undesirable outcomes relating to how funds invest, may have some adverse impacts on consumers, and create a distorted industry structure (the potential for 'zombie' funds). We expect there to be a detrimental effect on both industry performance and individual consumer outcomes.

We detail a range of alternative solutions and self-assess through the lens of consumer outcomes. To summarise:

- A solution which involves regulator assessment is, as it stands, the only way to acknowledge the qualitative issues and the through-time changes made by funds. It can also incorporate metric(s).

- Better metrics exist than the YFYS performance metric and a well-designed collection of metrics has advantages over a single metric.

- The YFYS performance metric can be improved. At best it can be more stable and assess all fund-types consistently. Nonetheless, it will remain statistically ineffective at assessing implementation performance, which itself is a small component of total investment performance.

We are positive about the opportunity to improve consumer outcomes. There is a great opportunity to implement an effective performance test which protects consumers from being exposed to funds which are likely to underperform in the future, whilst limiting undesirable outcomes. We are happy to share models and discuss this work and solutions in further detail.

Appendix 1 – Brief literature review on performance persistence

Paper	Setting	Findings
Performance Persistence The Journal of Finance Stephen J. Brown and William N. Goetzmann 1995	US mutual funds	- Our sample, largely free of survivorship bias, indicates that relative risk-adjusted performance of mutual funds persists; however, persistence is mostly due to funds that lag the S&P 500.
The Persistence of Risk-Adjusted Mutual Fund Performance The Journal of Business Edwin J. Elton, Martin J. Gruber and Christopher R. Blake 1996	US mutual funds	- We find that past performance is predictive of future risk-adjusted performance. Applying modern portfolio theory techniques to past data improves selection and allows us to construct a portfolio of funds that significantly outperforms a rule based on past rank alone. In addition, we can form a combination of actively managed portfolios with the same risk as a portfolio of index funds but with higher mean return. The portfolios selected have small but statistically significant positive risk-adjusted returns during a period where mutual funds in general had negative risk-adjusted returns.
Mutual Fund Performance: An Empirical Decomposition into Stock-Picking Talent, Style, Transactions Costs, and Expenses The Journal of Finance Russ Wermers 2002	US mutual funds	- We find that funds hold stocks that outperform the market by 1.3 percent per year, but their net returns underperform by one percent. Of the 2.3 percent difference between these results, 0.7 percent is due to the underperformance of nonstock holdings, whereas 1.6 percent is due to expenses and transactions costs. Thus, funds pick stocks well enough to cover their costs. Also, high-turnover funds beat the Vanguard Index 500 fund on a net return basis. Our evidence supports the value of active mutual fund management.

Short-Term Persistence in Mutual Fund Performance The Review of Financial Studies Nicolas P. B. Bollen, Jeffrey A. Busse 2005	US mutual funds	- The post-ranking abnormal return disappears when funds are evaluated over longer periods. These results suggest that superior performance is a short-lived phenomenon that is observable only when funds are evaluated several times a year.
Luck versus Skill in the Cross-Section of Mutual Fund Returns The Journal of Finance Eugene F. Fama, Kenneth R. French 2010	US mutual funds	- The aggregate portfolio of actively managed U.S. equity mutual funds is close to the market portfolio, but the high costs of active management show up intact as lower returns to investors. Bootstrap simulations suggest that few funds produce benchmark-adjusted expected returns sufficient to cover their costs. If we add back the costs in fund expense ratios, there is evidence of inferior and superior performance (nonzero true α) in the extreme tails of the cross-section of mutual fund α estimates.
Performance Persistence of Dutch Pension Funds De Economist Xiaohong Huang, Ronald J. Mahieu 2011	Dutch pension funds	- We find that pension funds as a group cannot beat their self-selected benchmarks consistently. Applying a cross-sectional portfolio approach, we find evidence that the largest pension funds outperform the smallest funds.
On Persistence in Mutual Fund Performance The Journal of Finance Mark M. Carhart 2012	US mutual funds	- Persistence in mutual fund performance does not reflect superior stock-picking skill. Rather, common factors in stock returns and persistent differences in mutual fund expenses and transaction costs explain almost all of the predictability in mutual fund returns. Only the strong, persistent underperformance by the worst-return mutual funds remains anomalous.
Active Investing as a Negative Sum Game: A Critical Review Journal of Investment Management, Forthcoming Geoff Warren 2020		- The case for active versus passive management is sometimes discussed or reported as an adversarial debate, and portrayed as a ubiquitous either/or choice. The mantra that 'active management adds no value' seems to have been embraced in some quarters, and is helping to fuel the current switch to passive management. A key point to emerge from this review is that approaching active versus

		passive as an either/or choice is not only unhelpful, but also does not accord with the broader body of evidence. The discussion would be much healthier if it was framed around seeing the choice between active and passive as a conditional one, including giving consideration to whether both might be used in tandem.

Appendix 2 – Approximation of implementation volatility

The implementation volatility of each super fund will be a unique number. What is important for our analysis is to have a representative number that is plausible. Our workings are briefly detailed below. We note that this simple model is available to be manipulated and our statistical tests allow for different implementation volatility assumptions.

A snapshot of the calculation is provided in Diagram A2.1.

		# Mgrs	Gross TE	Net TE	TE Cont.
Australian Equities	25%	4	3.0%	1.5%	0.38%
Global Equities	25%	4	3.0%	1.5%	0.38%
Unlisted Property	10%	1	3.0%	3.0%	0.30%
Unlisted Infrastructure	10%	1	3.0%	3.0%	0.30%
Credit	5%	1	2.0%	2.0%	0.10%
Fixed Income	20%	2	1.0%	0.7%	0.14%
Cash	5%	1	0.0%	0.0%	0.00%
		Gross Volatility:		1.6%	
		Net Volatility:		0.7%	

Diagram A2.1: Forming a sensible estimate for implementation volatility.

To explain Diagram A2.1:

- We defined a simple asset allocation of a growth-oriented super fund.
- We assume that a number of managers for each sector and the consistent level of tracking error (TE) targeted by each manager.
- To generate the net tracking error across for each asset sector we assume that the correlation of active returns amongst each manager is zero.
- We assume the correlation of net tracking error between sectors is also zero.

Appendix 3 – Benchmarking 'noise' created by YFYS performance test

Asset	How Benchmarked	Mis-treatment	Impact on Super Fund Assessment	Potential Size of Impact	Est. Tracking Error
Private Equity	Benchmarked against listed equity performance	The relative performance will depend on how 'seasoned' the private equity portfolio is to smooth out J-Curve effects.	Funds with less seasoned portfolios are likely to underperform equities in a strong market environment.	5% allocated to private equity 3yrs ago could have an impact of 20bp pa once spread over the 8yr period[6].	Complex to estimate (but high).
Unlisted Property	Benchmarked against listed property performance	There can be large variations in the performance of listed and unlisted property.	At different points in time funds will experience positive or negative performance which is independent of how well they implemented their unlisted property portfolio.	Over rolling 8yr periods the difference in returns between unlisted and listed property have been as large as 9.3% pa. A 10% allocation to unlisted property could have resulted in a fund performance difference	20.5% pa.

[6] Calculated as three years of equity returns while early stage private equity returns were flat.

				of >90bp pa relative to the YFYS benchmark.	
Unlisted Infrastructure	Benchmarked against listed infrastructure performance	There can be large variations in the performance of listed and unlisted infrastructure.	At different points in time funds will experience positive or negative performance which is independent of how well they implemented their unlisted infrastructure portfolio.	Over rolling 8yr periods the difference in returns between unlisted and listed infrastructure have been as large as 10.2% pa. A 10% allocation to unlisted infrastructure could have resulted in a difference of >100bp pa relative to the YFYS benchmark.	20.2% pa.
Credit (including high yield, direct lending activities, and emerging markets debt)	Benchmarked against composite bond indices	There can be large variations in the performance of various forms of credit and respective composite bond indices. Generally, credit brings higher yield and, adjusted for realised credit losses,	At different points in time funds will experience positive or negative performance which is independent of how well they implemented their credit portfolio. Over the longer-term funds which had more	Over rolling 8yr periods the difference in returns between high yield and the global composite bond index has been as large as 9.1% pa. A 10% allocation to high yield bonds could have resulted in a difference of	9.6% pa.

		would be expected to outperform.	invested in credit are likely to experience a favourable uplift in their performance, independent of how well they implemented their credit exposure.	>90bp pa relative to the YFYS benchmark.	
Australian inflation-linked bonds	Benchmarked against the Australian Composite Bond Index	There can be large variations in the performance of inflation-linked bonds and the bonds which comprise the Australian Composite Bond Index.	At different points in time funds will experience positive or negative performance which is independent of how well they implemented their unlisted inflation-linked bond portfolio.	Over rolling 8yr periods the difference in returns between Australian inflation-linked bonds and the Australian composite bond index has been as large as 2.1% pa. A 10% allocation to inflation-linked bonds could have resulted in a difference of >20bp pa relative to the YFYS benchmark.	4.3% pa.
Other (including alternative assets)	Benchmarked against an equal-weighted mix of blended equities and global fixed income.	There can be large dispersion in the risk profile of different investments in the 'other' category.	The risk level of investments in the 'other' category can vary greatly. By having a fixed benchmark, funds with lower (higher) risk exposures to alternatives	Over rolling 8yr periods lower or higher risk portfolios of alternatives could have experienced 'benchmark	3.1% pa.

			are set an arbitrarily high (low) benchmark[7].	differentials'[8] as large as 1.9% pa. A 10% allocation to low risk rather than high risk alts could have resulted in a tailored benchmark difference of approx. 20bp pa.	

Table A3.1: Exploration of asset class benchmark 'noise' created by the YFYS performance metric.

[7] This is based on the expectation that equities outperform bonds over the long-term.
[8] This is based on the difference between the benchmark performance for the proposed (50% growth, 50% defensive) benchmark and a more appropriate benchmark of 25% growth / 75% defensive.

Appendix 4 – Estimating 'implementation noise' which comes from YFYS benchmarking approach

The implementation noise of each super fund will be a unique number. What is important for our analysis is to have a representative number that is plausible. Our workings are briefly detailed below. We note that this simple model is available to be manipulated and our statistical tests allow for different implementation volatility assumptions.

A snapshot of the calculation is provided in Diagram A4.1.

	Allocation	TE	Gross TE contribution			Unlisted Property	Unlisted Infrastructure	High Yield		
Unlisted Property	10%	20%	2.0%		Unlisted Property	1	0.5	0		0.03
Unlisted Infrastructure	10%	20%	2.0%		Unlisted Infrastructure	0.5	1	0		0.03
High Yield	10%	9%	0.9%		High Yield	0	0	1		0.009
		Gross Total TE	4.9%							3.6%
		Net Total TE	3.6%							

Diagram A2.1: Forming a sensible estimate for implementation volatility.

To explain Diagram A2.1:

- We define allocations to three representative sectors with noted YFYS-related benchmarking issues.
- The tracking error estimates (tracking error for each asset to its public market index) is detailed in Table A3.1.
- To generate the net tracking error, we apply a variance-co-variance approach assuming a degree of positive correlation between the benchmark tracking error of unlisted property and unlisted infrastructure.

Appendix 5 - Proposing a single metric (working version)

Designing a single metric to assess performance is a controversial and highly subjective area. On the assumption that this may be what is ultimately required, we "stick out our chin" with a working version.

A5.1. Workings of proposed metric

Description:

Performance versus a volatility-matched reference portfolio

How it works:

The steps can be interpreted by reference to Diagram A5.1.

Step 1: A fund has realised (net of investment fees) return (A) and realised volatility (B) measured over 8 years

Step 2: Construct a reference portfolio which has realised volatility equal to (B)

- The reference portfolio is a combination of two portfolios:
    1. 'Growth' portfolio: 50% Australian shares and 50% global shares (50% hedged)
    2. 'Defensive' portfolio: 20% cash, 40% Australian bonds, 40% global bonds (hedged)

  Calculate the realised performance of the reference portfolio, net of benchmark fees and benchmark tax for each asset class: (C)

Step 3: Calculate fund performance including historical administration fees: (D)

Step 4: Calculate fund performance including current administration fees: (E)

Step 5: Calculate relative performance based on historical administration fees: (F1) = (D) – (C)

Step 6: Calculate relative performance based on current administration fees: (F2) = (E) – (C)

Step 7: Compare (F1) or (F2) against an appropriately determined hurdle

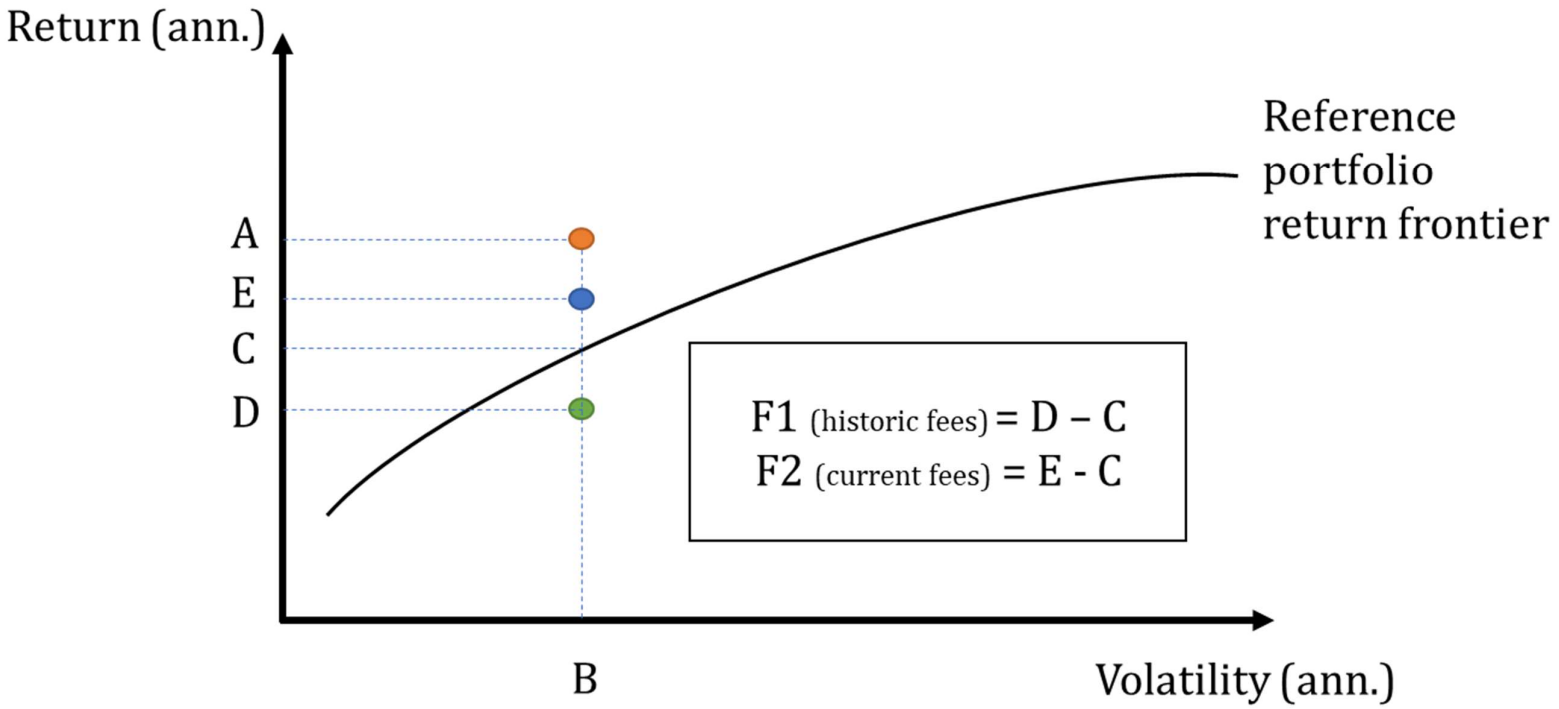

Diagram A5.1: Representation of "working version" performance metric.

All that is required to calculate the proposed metric is a time series of performance (recommend monthly). We also note that by comparing to a reference portfolio frontier we normalise for the scenario that long-term equity returns are modest or weak and the frontier is downward sloping at a particular point in time (a Sharpe Ratio approach fails to capture this).

A5.2. Assessing the proposed metric

In Table A5.1 we self-assess using the same criteria considered in Table 13.

	Proposed metric
Recognise member outcomes over lifetime	The metric is a performance metric and doesn't consider lifetime outcomes
Could be designed to be a pure investment test or a consumer outcome test	Either
Acknowledges high current fees	Yes – if the F2 measurement is applied
Acknowledges weak governance	No – not considered
Focus on total investment outcomes	Yes
Acknowledges net risk outcomes	Yes – realised volatility recognises the performance of how well portfolio protection and diversification strategies.
Accounts for non-stationarity	Only accounts for fee changes
Statistical effectiveness	Reasonable effectiveness – similar to Table 2
Potential for undesirable outcomes	Maximising risk-adjusted returns to consumers is well-aligned with the focus of trustees. One major potential distort is the enhanced attraction of unlisted assets which appear to have understated volatility on a period-to-period basis (see further notes below). However, a large range of undesirable outcomes are likely to result from a performance test.

Table A5.1: Assessment of proposed metric across multiple criteria.

We make some further comments on unlisted assets:

- There are often claims made that the volatility of unlisted assets is understated and that public market equivalents represent true insight into prices, hence the volatility of unlisted assets should be the same as listed assets. An opposing argument can be made that public market prices

include an element of speculation which heightens volatility. In effect heightened volatility is the 'cost' of liquidity. We are not taking a view either way, rather acknowledging that there are opposing, strongly held views.

- Research into unlisted property reveals that quarterly and semi-annual returns have high autocorrelation. This doesn't mean that the volatility of quarterly or semi-annual returns are understated. Rather it means that returns in consecutive periods are not independent. This means that the volatility of unlisted property over longer time horizons is understated. An adjustment could be approximated (something in the order of adding 0.25% to total annual realised volatility is a rough estimate). This would likely diffuse any concerns raised by parts of the industry which do not invest in unlisted assets.