Chant West research paper

November 2022

YFYS performance test: a better way

In our corresponding article a year ago we argued that the Your Future, Your Super performance test was flawed, that it failed to measure what was important for member outcomes and that its application had adverse, unintended consequences. We vowed to work to improve it, and our ideas for improvement are the subject of this article.

The problems we highlighted with the current test firstly had to do with the sole metric against which funds were evaluated, and secondly with the consequences of the test, not only for the funds that failed but also for the subsequent behaviour of those that passed.

Regarding the test itself, we argued that:

- A test based on just one measure over one time period could never produce a representative picture of a fund's ability to deliver long-term value to its members
- The test fails to measure the value added by what is widely accepted as the most important driver of a fund's long-term performance – its strategic asset allocation (SAA)
- Instead, it measures how well a fund has implemented its SAA (e.g. through asset class strategies, manager selection and active asset allocation tilts) against a set of asset sector indices that (a) do not necessarily reflect how funds actually invest in those sectors and (b) in some cases are uninvestible

And as for the consequences, we predicted that:

- The dire consequences of failing the test (public shaming, member anxiety and existential threat to the fund) so far outweigh the rewards for passing, that some funds would make passing the test their prime investment focus
- This in turn would change the way many funds invest – even some of those that passed the test by a comfortable margin.

Based on our observations over the past year and with the publishing of the second round of test results, it is clear that our fears were warranted. Of the 13 funds that failed the first test in 2021, only two now survive. But funds have a strong survival instinct and quickly worked out what they needed to do to pass the test, with the result that no fund failed in 2022 that had not already failed in 2021.

Funds are concerned that they can deliver a strong return for members through a well-constructed SAA but still fail the test. While failure might be unlikely for any single fund in any given year, the way the test is constructed means that failure is a distinct possibility for the majority of funds at some time over the next 5 to 10 years. Even those funds that comfortably passed the first two tests are monitoring their YFYS tracking error to ensure they don't introduce the risk of failure down the track. Funds with a smaller buffer, of course, are particularly fixated on the test and its consequences.

On the ground, we are seeing clear evidence of how the test is influencing funds' investment behaviour. We forecast that some funds' 'best ideas' might be compromised, and we know of several instances where funds have not taken up investment opportunities that they would have taken up eagerly before the test. Perhaps of more concern, we know of several funds that terminated diversifying strategies in 2021 because they had relatively high YFYS tracking error, when those strategies (e.g. alternatives, portfolio protection, lower volatility equities, shorter duration fixed interest) would have provided very effective protection in the FY22 investment environment when major asset sectors fell significantly.

While the test has identified and eliminated some funds that were clearly sub-standard and destined to fail, it has caused collateral damage to others that were seeking to do the right thing by their members but were somewhat unlucky in the timing of the first test (i.e. over seven years of strong returns). More ominously, it has driven the behaviour of the surviving funds in ways that, in too many cases, have resulted in diminished returns for their members. Those are the unintended, and unacceptable, consequences of the current test regime.

Which metrics, and how many are enough?

One of our criticisms of the current test is that it is so myopic, relying on just one measure over one period. Evaluating super funds properly is complex and multi-faceted, and ideally requires a whole range of measures to be taken into consideration. These might include the current test together, perhaps, with: **return versus peers, return adjusted for risk, and return versus a simple reference portfolio**. Each of these measures has merits, but each also has its own drawbacks.

The current test, as we have seen, is a blunt instrument in correctly recognising underperforming funds. Adding more measurement periods – maybe 3 years and 5 years as well as the current 8 years – would at least introduce a dynamic element and pick up instances where funds have identified past problems and taken steps to address them. But that still wouldn't fix the inadequacies of the test or the way funds' fear of failure constrains and distracts them from investing in their members' best interests. Other measures should also be utilised.

Peer comparison would involve measuring each fund's performance against other investment options with similar levels of risk. Those peer universes would have to be quite tight in terms of the risk band. For example, using our Growth category (61–80% growth assets) would be unfair because funds towards the top of the growth assets range would tend to outperform over time. A very tight risk band, however, might result in a universe so small that comparisons become meaningless.

Another problem with a peer comparison is the perennial one about the classification of growth and defensive assets, where there is still no accepted industry standard. Also, peer comparison is relative and so provides no insight into whether the peer group as a whole is performing well or badly. And if any funds that lie, say, in the bottom 10% in each category fail the test, then some funds will fail every year no matter how well they are objectively doing for their members.

A metric incorporating risk-adjusted returns would indicate whether fund members are receiving a reward commensurate with the risk being taken on their behalf. Volatility is the most commonly accepted measure of risk, and would most likely be used because it is simple to measure over any time period. It does, however, only reflect one type of risk.

The risk-adjusted comparison could be with a peer group or with a benchmark portfolio, or both. One issue with this approach is that it favours funds with a significant exposure to unlisted assets. Those assets exhibit lower volatility because of the infrequency with which they are valued (typically quarterly rather than daily as is the case for listed assets) so for any nominal rate of return they generate a higher risk-adjusted return. A significant exposure to unlisted assets is a hallmark of the most successful wealth funds in the world and is widely accepted as global best practice. If the adoption of this metric were to lead to superior risk-adjusted returns over the long term, even recognising the issues with volatility, that is probably a good thing for fund members, as long as funds appropriately manage liquidity.

The simple reference portfolio (SRP) approach was first aired in the Productivity Commission report and is already in use in APRA's heatmap. In its simplest form, it measures what value a fund has added relative to a naïve portfolio with the same growth/defensive mix but comprising nothing more than listed equities, bonds and cash, all using common market indices as shown in Table 1.

Table 1: Simple Reference Portfolio – allocations and indices

Growth	Benchmark	%	Defensive	Benchmark	%
Australian Equities	S&P/ASX 300	50	Australian fixed interest	Bloomberg AusBond Composite Index	40
Int' Equity (hedged)	MSCI ACWI ex-Aust Special Tax (hedged)	25	Int'l fixed interest	Bloomberg Global Aggregate Index (hedged)	40
Int' Equity (unhedged)	MSCI ACWI ex-Aust Special Tax (unhedged)	25	Australian Cash	Bloomberg AusBond Bank Bill Index	20

Source: Chant West

As with the peer comparison, this approach entails a determination about what constitutes a growth asset. And, if it were to stand alongside the existing test, it would introduce another benchmark for which funds would need to manage their tracking error. With those caveats, this approach has some logical appeal because the SRP represents what members could theoretically achieve for themselves with index funds and no professional input whatsoever.

These, then, are some of the other measures that might be considered for the test. Each of them would provide a different perspective on how a fund has performed. None of them are perfect, but used together and over multiple timeframes they would provide much greater efficacy in identifying the underperforming funds.

When we contemplate these multiple metrics, together with measurement over multiple timeframes, we start to see how complicated the ideal test might become. It may be technically superior, but even a combination of two or three metrics over three time periods could be seen as 'over-egging the pudding'. Treasury may prefer to continue with a single 'bright line' measure for simplicity and clarity of message, at least for MySuper - a fund fails the test, and there are consequences. With that in mind, we have considered all the available options and now propose what we believe to be the best and simplest single alternative metric.

Chant West's proposal for an alternative test

Our proposal is a composite of two of the approaches outlined above: risk-adjusted returns and the simple reference portfolio (SRP). By combining these two approaches we end up with a metric that captures their merits and minimises their shortcomings. It also addresses the two key factors that determine fund members' investment outcomes, namely risk and return.

Under our proposal, a fund's performance would be compared with the performance of a SRP with the same volatility. Say, for example, that Fund A and Fund B both exhibited a volatility of returns of 6% per annum over the past 8 years. Their returns would be compared with those of a SRP that also had a volatility of 6% per annum. If that SRP had returned 8% per annum over the period, then that would be the pass mark. Fund A returned 9% per annum and so passed, while Fund B's 7% per annum return would see it fail.

This approach has some key advantages:

- The benchmark is simple, and avoids all the technical arguments that are still being debated over the current model (like which, and how many, asset class benchmarks to use)
- It is not reliant on a particular methodology to classify what are growth assets
- The calculations are also simpler than at present. The performance and volatility of both the fund and the SRP are calculated over the entire period under consideration (say 8 years) rather than being stitched together from multiple quarterly data points
- Volatility is simple to measure, and APRA already collects quarterly performance data from funds so no additional inputs would be required
- Most importantly, perhaps, our test incentivises funds to strive for strong risk-adjusted returns, so it is absolutely aligned with desired member outcomes
- Because the test aligns with members' best interests, funds are less likely to be distracted from their proper long-term goal rather than on passing an annual test.

Our preference would be for our proposed test to replace the current test at the earliest possible opportunity. We do recognise, however, that this would be too abrupt a change, in that funds would have to switch their focus from passing one test to passing a different test.

It may be, therefore, that a compromise can be achieved by running both the current and new tests in parallel for a short time. If that is the case, we would argue that passing at least one of the tests should be sufficient for survival during this period. A multiple test scenario is not ideal (because funds have already learned how to pass the current test and might continue to concentrate on that) but it would give funds a little time to re-focus if they knew that the replacement test would soon be the only one that matters.

Our proposed new test – the results

To assess how our proposed test would have worked out, and how it compared with the current test, we applied it initially to the 42 distinct MySuper products covered in our Multi-Manager Survey. That amounts to 94 portfolios in total, spanning both single option strategies and lifecycle cohort strategies. The test period was 8 years to 30 June 2022. For this analysis, we included a 50 bps tolerance in the same way as the current test.

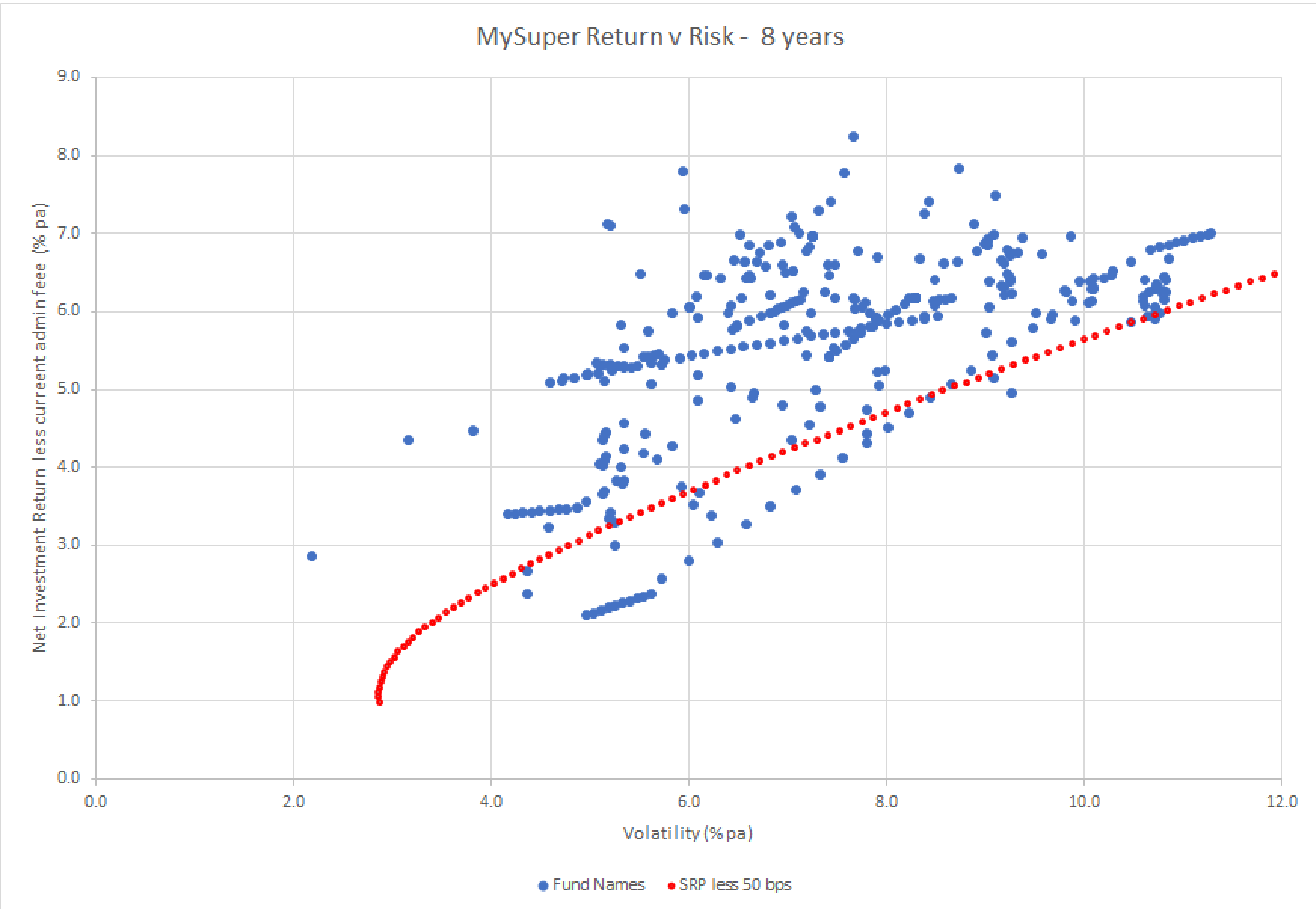
Of the five funds that failed the 2022 YFYS test, only three are in our survey universe because the other two don't provide us with monthly performance data. Of those three, only one would have failed our test. Another fund fell within the 50 bps tolerance band, so would have failed if that buffer was removed.

These initial results seemed reasonable and indicated the approach was worthy of further consideration. One factor we wanted to investigate further was volatility, given that a lower volatility SRP produces a lower 'pass mark' return, so the lower volatility of unlisted assets favoured funds that had significant allocations to them.

All the funds in our survey provide us with monthly returns so we were able to calculate volatility based on monthly data, and that is what we used for our initial analysis. But APRA doesn't have access to reliable monthly performance data, so we then re-ran our model using quarterly volatility instead. What we found was that, although absolute volatility changed, the ranking of volatility between funds remained very consistent. This gave us confidence that using quarterly data would make little difference to the ranking of funds, while it would suppress the unlisted asset advantage to some degree.

Our next step was to apply our test to the APRA data. This is a wider universe of 65 distinct MySuper products comprising 361 single sector and lifecycle portfolios. The results for each portfolio are plotted in Chart 1, together with a dotted line representing the performance of the risk-adjusted SRP less a 50 bps buffer.

Chart 1: Return vs risk compared with Simple Reference Portfolio to Jun 22 – 50 bp buffer

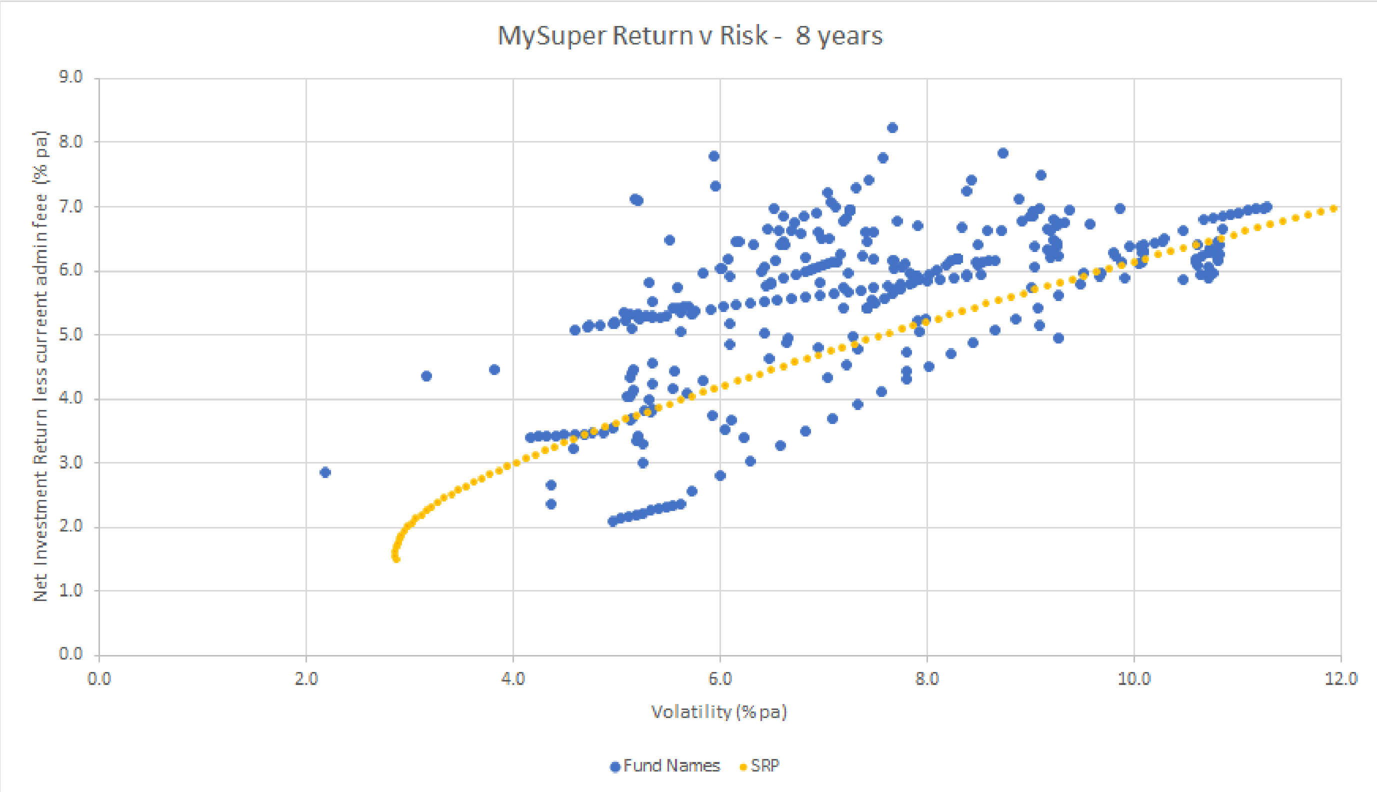


Note: Performance is net of representative administration fees and expenses (RAFE).

Under this test, only one very small product would have failed, AMG Corporate Super. The string of data points below the SRP-50 bps line all relate to older cohorts of lifecycle products whose younger cohort results beat the benchmark, and the weight of younger members would have seen all these products pass the test.

We then considered what would happen if the 50 bps tolerance was removed. The outcome is shown in Chart 2, and the result is that eight funds would have failed. These include all the funds that failed the current test, with the exception of Australian Catholic Super, and another four products whose portfolios have minimal exposure to unlisted assets.

Chart 2: Return vs risk compared with Simple Reference Portfolio to Jun 22 – no buffer



Note: Performance is net of representative administration fees and expenses (RAFE).

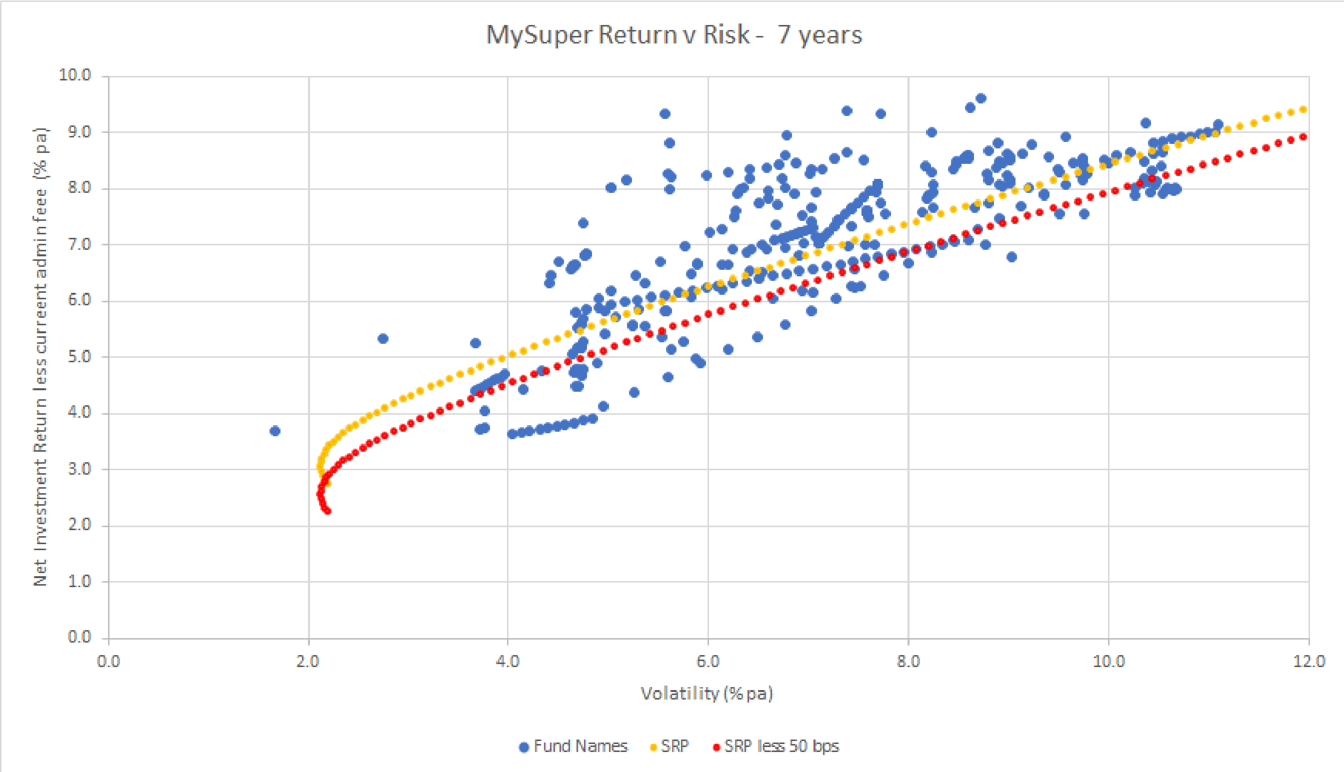
The 50 bps tolerance may be justified for the current test, but there may be an argument to significantly reduce it for our proposed test. It was introduced in the current test to provide some leeway for funds, most likely as a nod towards the limitations of the test model in identifying true underperformance.

In contrast, our test already has a built-in advantage for funds that are well-diversified and so demonstrate lower volatility. So a lower tolerance such as 25 bps (as used in the Productivity Commission report in relation to the SRP) or even 10 bps could be used.

2021 test revisited

Our test on the 2022 data achieved the sort of results we were expecting, so we thought it would be interesting to take a look back to 2021. Readers will recall that the initial YFYS test resulted in 13 failures, so what would the results have been if the test had been run using our proposed metric? The answers are quite illuminating as shown in Chart 3, which includes benchmarks of both the SRP return and the SRP return less 50 bps.

Chart 3: Return vs risk compared with Simple Reference Portfolio to Jun 21



Under our test, still allowing the 50bps tolerance, nine funds would have failed in 2021. These comprise seven MySuper products that failed the 2021 YFYS test and another two that passed the test (i.e. Commonwealth Essential Super and Westpac Group Plan) but are closely related to products that failed in 2021.

Importantly, four funds that failed the YFYS test in 2021 but passed it in 2022 (AvSuper, Christian Super, Commonwealth Bank Group Super and Maritime Super) would have passed our test in both years. Yet that first year failure meant they had to endure the consequences of the test of (a) being named as an underperforming fund and (b) having to notify their members of that fact and suggest they seek an alternative fund. In addition, there are another two funds that failed the YFYS test in 2021 and have since closed, but they would have passed our 2021 test (BOC Gases Super and LUCRF). Since these two funds have some similarities to the four funds mentioned above, they may well have passed the YFYS test this year, given the changed investment environment. While some of these six funds were smaller and there were concerns about their scale, we believe it would be more appropriate to deal with the scale issues rather than remove them from the system due to a flawed performance test.

If the 50 bps tolerance was replaced with a tolerance of say 10 bps, no additional funds would have failed our test, but if the tolerance was removed altogether, a further three funds would have failed.

In summary, our analysis shows that our alternative test is less susceptible to the 'false positives' that are a recognised problem with the current test. These 'false positives' constitute the sort of collateral damage we wrote about last year and have feared all along. They stem from the mismatch between the efficacy of the existing test and its consequences.

Efficacy and consequences

The efficacy of any test should have a direct bearing on the severity of the consequences of failure. At present we have a test that has low efficacy in terms of identifying underperforming funds, yet as we have seen the consequences of failure are dire.

We commented last year that we believed there were funds that failed the test that deserved to pass, and that seems to be borne out when we re-test using our metric. Also, we continue to see good funds curtail some elements of their investment strategy due to the very rational fear of failing the test at some point, not because they are poor funds but because of the methodology currently applied. We need to address that unhelpful distraction, and the best way to do so is to replace the current test with one capable of gaining broad acceptance as a reasonable basis for taking significant remedial action against a fund.

Our proposed test helps to rectify that asymmetry between efficacy and consequences. The test is simple, robust and aligned with member interests. A fund that fails our test really does identify itself as an underperforming fund, and probably needs to bear the consequences.

Application to choice and retirement income products

We support the current pause on including choice products in the test regime, but we still believe that choice options should be included at some point. Indeed, we believe the test should be extended not just to trustee-directed products but to all diversified choice options whether trustee-directed or not. Otherwise, the same investment strategy that is subject to the test (and its consequences) in the manager's own product offering would not be tested when offered by a different product provider. That is a recipe for confusion. Further, we believe that since choice products have been selected by a member and/or their adviser, the consequences for failing the test should be different to those that apply for MySuper.

The question then is: which test (or tests) would work best for choice products? Even if we exclude single sector options (which we don't believe should be included in the test regime at all) the range of diversified options is so broad – including variants such as responsible investment, real return and income-focused strategies – that the shortcomings of the current test would only be magnified and the difficulties compounded. And also, how do we deal with a product where some of the investment options pass the test but others fail, potentially leading to the closure of conservative options? Our view is that choice products should be included only when a more appropriate test, such as ours, is in place.

Further, we believe that testing of retirement income products should not be introduced any time soon. These portfolios are grossly ill-suited to the current test, which is so focused on implementation. We have seen that already, where funds did the right thing by their members by taking defensive positions to protect against downside risk when markets were strong (as they were for the first seven years of the look-back period), only to be penalised under the current test. The same problem would inevitably occur for retirement income products. Their members are in the drawdown phase and particularly vulnerable to volatility and downside risk. Funds need to look after these retirees' needs, and should not be penalised for taking positions between and within asset sectors to reduce risk when they consider it prudent to do so. Also, the Retirement Income Covenant encourages funds to develop new products that deal with the well-known risks in retirement – the last thing we want to do is stifle this innovation and for funds to simply offer carbon-copies of benchmark-aware accumulation products for their retired members.

So we believe it should be one step at a time. First, transition to a new test for MySuper products as soon as practicable and, once that is bedded down, extend its scope to include diversified choice products but not retirement income products.