Submission to Treasury on design options for the annual superannuation performance test

19 April 2024

David Bell & Geoff Warren

About The Conexus Institute

The Conexus Institute is an independent, not-for-profit research institution focused on improving retirement outcomes for Australian consumers. Philanthropically funded, the Institute is supported by the insights of a high-quality advisory board, who work on a pro-bono basis. The Institute adopts a research-for-impact model and frequently collaborates with researchers from academia, associations, and industry. Where possible research is made open source to assist industry and create transparency and accountability. Further information here.

About David Bell

Dr David Bell is Executive Director of The Conexus Institute. Bell's career has been dedicated to the investment and retirement sector. He has worked with both commercial and profit-for-member firms, and ran his own consulting firm. Bell worked with APRA in the development of the APRA heatmap. Academically, Bell taught for 12 years at Macquarie University and in 2020 completed his PhD at UNSW which focused on retirement investment problems. Full bio here.

About Geoff Warren

Dr Geoff Warren is a Research Fellow with the Conexus Institute and an Honorary Associate Professor at the Australian National University, as well as a member of various investment and research advisory boards. Warren's research focuses on investment-related areas specially including superannuation and retirement, and is widely published in leading journals. He has a prior career in the investment industry spanning over 20 years. Full bio here.

***** The authors are willing and able to participate in further consultation. *****

Overview and summary

The Conexus Institute has historically expressed reservations about the value of a backwards-looking quantitative performance test (the 'test') and whether it provides net benefits to members. Further, the design of the current test has always raised concerns. The test has now been operating for three years, and the learnings inform this submission. We are pleased that the design of the test is being reviewed by Treasury. We also acknowledge Treasury for producing a high quality consultation paper.

We broadly agree with the principles outlined in the consultation paper, i.e. that the test should improve member outcomes, be effective and efficient, widely applicable and transparent, and enduring. We focus most strongly on member outcomes. The test should be designed for the benefit of members (not the industry). Indeed, we believe that the other principles listed are secondary and may need to be challenged in order to maximise member outcomes.

Nevertheless, we adopt a more targeted approach in our submission. We start with the purpose and objectives in evaluating performance, which relate to both assessment and incentives. We trace these concepts through to seven criteria, which lead us to a proposal for a three-metric test. Two of the metrics are based around assessing total portfolio returns as they are most aligned with member outcomes.

The performance test performs the function of assessing past performance, but also creates incentives that have behavioural impacts. There is substantial debate as to what degree past performance informs future performance and thus the right of a fund to continue to operate. We choose not to add to this debate other than saying it is far from clear, and that many factors inform future performance. Our primary aim is to propose a more effective test of past performance, when no perfect design exists. A performance test with strong consequences for failure also creates strong incentives to pass. It is important that the test aligns with investing for good member outcomes and limits unintended consequences. Further, the industry is now actively managing the current test making it quite possible that no MySuper option will fail in the future. As such, the current test may have become ineffectual yet nevertheless has some adverse behavioural impacts.

We strongly believe that it is possible to design a better, stronger test that is better aligned to member outcomes and less exposed to active management and associated adverse unintended consequences.

Our primary proposal is a three-metric performance test applied to all multi-asset portfolios, with the need to pass at least two of the three metrics. The three metrics consist of the current test alongside the peer-based and simple reference portfolio approaches (2b and 2c in the consultation paper). The test we propose better aligns with member outcomes through introducing two metrics that assess total portfolio returns. A three-metric test limits reliance on any single metric when all metrics have shortcomings, and will be more difficult to actively manage. Our hope is that funds will become less focused on managing to the test and conclude that the best way to pass is to work towards delivering good member outcomes.

Other aspects we raise in this submission include:

- While a two-metric approach involving a new metric may also benefit members to a degree, any requirement to pass one of two metrics would provide more options and incentive to manage the test(s). This could actually weaken consumer protections.
- Technical details should be determined later with the aid of a Technical Advisory Group.
- Consideration should be given to introducing a process of reviewing the test results before a fund is declared as 'underperforming'. The aim would be to capture situations where the results are misleading due to shortcomings in the test itself. Any review would occur under the presumption that the test results are correct unless there is clear evidence to the contrary.

- We do not recommend assessment of single sector products. If these products are tested, dedicated tests should be developed, i.e. tailored metrics, threshold levels and consequences.
- Care is needed in applying performance testing to retirement products, where the focus should be overall evaluation of retirement income strategies ideally under a member outcomes assessment framework. While performance testing might be applied to account-based pension products as part of this broader assessment, we recommend against the using existing performance test and view the three-metric test as more fit-for-purpose.
- With respect to ESG and sustainability products, the starting point needs to be the Government establishing its position on two matters. First, the ESG and sustainability activities it wants to facilitate, e.g. dedicated investments, exclusions, or both. Second, whether members have the right to invest based on their values on the understanding that it could result in lower returns. From here a testing framework, possibly dedicated, could be designed.

Submission structure

SUBMISSION

Section 1: Proposal for 'three-metric' test – Outlines the objectives and criteria for designing a better test, our proposed test structure and how it meets the criteria, each proposed metric, and implementation and review considerations.

Section 2: Alternatives – Discusses the metrics we rule out, candidate industry solutions, the case for a review process, single-asset and choice products, and application to retirement.

Section 3: Socially responsible and ESG investing – Summarises our past research on the impacts of the current test on sustainability activities, and discusses future possibilities in this area.

APPENDICES

Section 4: Working group note '*Improving the YFYS performance test*' – Reproduces a note documenting the outcomes from consulting with an industry working group that was able to reach broad agreement on the key design features for a revised test.

Section 5: Summary of Conexus Institute research on the performance test – High level summary of selected pieces of research undertaken on the performance test that may be of use to Treasury, with links and references.

Section 6: Responses to the consultation questions test – Brief response to each consultation question, with links back to our proposal and commentary in this submission.

Submission

1. Proposal for 'three-metric' test

This section sets out our proposal for a re-designing of the YFYS performance test as a 'three-metric' test. We outline key objectives and criteria, how the proposed test structure meets the criteria, issues with each proposed metric, and implementation and review considerations. We write this section from the perspective of developing an effective *quantitative test of past performance for multi-asset portfolios*, with MySuper options in particular in mind. Section 2 addresses issues such as testing of choice, ESG and retirement options, and whether to create leeway around the 'bright lines' nature of the test.

1.1. Key objectives

We agree with the broad principles set out in the Treasury consultation paper. Nevertheless, we have applied a different organising structure in forming our proposal. Our starting point is that any evaluation of past performance has two functions that in turn suggest two overarching objectives for the test:

(a) **Assessing performance** – Any assessment of performance supports identifying good performance so it can be rewarded, and poor performance so it can be addressed. The purpose and objective of the YFYS test is the latter, with the aim of weeding out funds that are chronically poor performers.

(b) **Incentivising behaviour** – Here the overarching objective for the YFYS test should be to encourage funds to focus on improving member outcomes, or at very least not divert them from doing so.

The current test has significant shortcomings when measured against both objectives, most of which are acknowledged by the consultation paper. To call out the major shortcomings, assessing implementation only tests one component of total portfolio returns, which are what matters for member outcomes. The design also creates incentives to manage the test as a priority, which may lead to adverse incentives and so-called 'unintended consequences' in some situations. We see significant scope to improve the test from both these perspectives.

In addition, using the test as an arbiter of the 'right to continue' implicitly assumes that the test is predictive of future performance, i.e. poor performers in the past will remain poor performers into the future. This is a *very* tenuous assumption. However, this shortcoming applies to *ANY* backward-looking measure of investment performance, and cannot be easily addressed through test design.

1.2. Our criteria

We have applied the following criteria in considering alternatives to the current test, many of which overlap with the principles set out in the consultation paper:

Assessing past performance:

1. **Member outcome focus** – This requires an assessment of total portfolio returns.
2. **Risk adjustment incorporated** – Risk adjustment facilitates comparison across multi-asset funds with differing risk positions, thus making the test more widely applicable.
3. **Effective and efficient** – The test should identify genuine underperformance, be relatively straightforward to administer, and the basis of assessment should be transparent and arrive at a clear result.
4. **Stronger, not weaker, test** – There should be no weakening of member protections.

Incentivising behaviour:

5. **Incentivises funds to focus on member outcomes** – Funds should ideally form the opinion that maximising member outcomes maximises the chance of passing the test.

6. **Difficult to manage and 'game'** – The capacity and incentive to 'manage' the test should be minimalised, leaving funds focused on maximising member outcomes rather than the implications of activities for their test results.
7. **Does not dictate industry behaviour** – The test should not direct the industry on where or how to invest. We note that the current test does so through the imposition of index benchmarks. Also, framing around strategic asset allocation (SAA) does not accommodate other approaches to portfolio formation, e.g. total portfolio approach, managing exposure to economic factors.

The ability of members to understand the test itself is not an important criteria in our view. The starting point is that the current test is not easy for members to understand in any event. We believe community expectation is that the best test will be applied, rather than apply a flawed test due to imposing that constraint that it is understandable by members. We also suspect that many members will take any pass/fail at face value, rather than try to understand how the result is derived.

1.3. General structure of the three-metric test

We propose moving to a three-metric test (option 3b in the consultation paper) where two additional tests evaluating total portfolio returns are added to the current test, along with the requirement to pass two out of three. The two additional tests we propose are presented in the consultation paper as options 2b and 2c, and are respectively denoted as 'peer comparison of risk-adjusted returns' and 'risk-adjusted returns relative to simple reference portfolio (SRP) frontier'.

1.3.1. Why the three-metric test

The table over assesses the proposed three-metric test against our criteria. Our analysis suggests the expanded test offers significant improvements with three areas standing out. First is much better alignment with member outcomes through introducing two tests focusing on total portfolio returns. Second, multiple metrics reduces reliance on any single metric when every metric has shortcomings. Third, capacity and incentive to manage the test should be much diluted under three metrics, thus encouraging funds to focus primarily on improving member outcomes. Nevertheless, some issues will remain as identified in the last column. The aim is to design a better test: perfection is impossible.

1.3.2. Why two out of three to pass

We favour a two out of three or 'on-balance' approach in determining pass or failure. As well as providing a result with a clear basis, this approach implicitly affords a higher weight to total portfolio returns, which are more directly linked to member outcomes than the current 'implementation' test. We are wary of a hierarchical approach that anoints one test as most important, as this could increase incentive to game the test sitting at the top of the hierarchy. Averaging may be acceptable. However, averaging would dilute visibility around the source of the failure, and detailed parameterisation would be required to the extent that the various tests will have differing distributions.

We see the failure criteria as a matter that could be referred to a technical advisory group (TAG).

Assessment of three-metric test against the criteria

Criteria	How criteria met	Issues
Assessing past performance		
1. Member outcome focus	• Two total return tests place two-thirds weight on primary driver of member outcomes	• Does not assess overall option design, including how much risk that is appropriate for members to take. An issue for MySuper in particular. • Backwards-looking nature of test remains a problem, i.e. it is not reliably predictive future outcomes.
2. Risk adjustment incorporated	• Two new tests as proposed entail an element of risk adjustment	• Standard deviation and growth / defensive mix both have shortcomings as risk measures (see Section 1.4)
3. Effective and efficient	• Multiple metrics diversify exposure to the shortcomings of any specific test, noting that any single metric will be flawed. 'Spotlights on performance from differing angles'. • Significantly dilutes reliance on indices, and their efficacy as benchmarks • Two new tests will be easy to administer as they draw on readily available data	• Three tests more complex than one • Introducing new tests imposes some implementation burden on the super industry and regulators • Still reliant on indices to a degree • Some significant implementation issues will need to be addressed (see Section 1.4)
4. Stronger, not weaker, test	• Adding perspective on total portfolio returns strengthens the test • Likelihood of detecting underperformance improved relative to current situation where funds have learnt how to manage current test	• Scope of strengthening depends on how the three tests interact, i.e. extent to which more 'spotlights' better highlight genuine underperformance
Incentivising behaviour		
5. Incentivises funds to focus on member outcomes	• Encourages maximising total portfolio returns while managing risk, both of which are relevant to member outcomes • Asset diversification encouraged by SRP test • Dilutes incentive to focus on tracking error versus index benchmarks • New tests provide additional room to invest in off-benchmark assets, e.g. sustainability	• The very existence of a test will still influence behaviours to some extent • Two tests (SRP, current) refer to index benchmarks, and hence do not totally remove incentive to manage benchmark-relative risk • Peer risk is made more relevant, albeit minor (see Section 1.4.2)
6. Difficult to manage and 'game'	• Three tests with differing benchmarks much harder to manage and game than a single test	• Some opportunity and incentive to manage the tests would remain
7. Does not dictate industry behaviour	• Impact of benchmark index selection on investment behaviour is much diluted • The two new tests accommodate non-SAA approaches to portfolio construction	• Benchmarks remain of some relevance, and hence will influence industry behaviour to a degree

Source: Conexus Institute

1.4 Comments on the three metrics

We comment on each of the three metrics individually including their nature, what the test brings to the party, and any major issues. An important point is that the metrics *differ* along all these lines, bearing in mind that a key motivation in proposing multiple tests is that any single metric will have shortcomings and there is benefit in shining spotlights on performance.

1.4.1 Current test

The nature of the current test is well understood and needs no comment. The main reasons to retain the current test include to maintain some continuity, improve acceptance and aid transition. A sudden jump to an entirely new testing regime creates uncertainty and will probably not be well-received by an industry that has developed processes around the current test. Retaining the current test also assists in constructing a three-metric test, which has advantages as outlined in Section 1.3.

The current test also casts a different spotlight on performance to the two new tests we propose. Assessing performance versus benchmark indices in accordance with a fund's reported SAA means that the test assesses the breadth of implementation activities being undertaken by a fund across asset classes. Failure across a breadth of activities might indicate systemic problems within the organisation. By contrast, the two proposed tests that focus on total portfolio returns may be more heavily influenced by SAA activities that are typically low breath. For instance, a single SAA position may lead to failure of the overall test, but could reflect a risk-based decision where the risk event did not transpire that does not reflect systemic problems[1].

Nevertheless, we suggest that the role for the current test metric be reviewed at a later date. This matter is discussed in Section 1.6.

1.4.2 Peer comparisons of risk-adjusted returns

This metric examines total portfolio returns and thus links to member outcomes. It benchmarks performance against the expected return for a hypothetical fund with an equivalent growth/defensive (G/D) mix. Using G/D weights by fitting a regression line amounts to an implicit form of risk adjustment through establishing the additional return required for additional units of exposure to growth (i.e. risker) assets. By considering the realised performance of the industry this metric captures a range of practical challenges (e.g. portfolio re-balancing costs) not identified by more theoretical benchmark-based approaches. This metric is market based: it acknowledges that each fund is navigating the same investment market challenges that theoretical benchmark-based strategies may struggle to reflect. The main implication of this method is that the test is *NOT* a direct peer assessment, as there is no distinct peer group against which performance is being compared. Essentially a fund is being benchmarked against the full sample of funds with a range of differing asset mixes and approaches. This much dilutes the incentive for peer-focused behaviour relative to an alternative where a sub-group of peer funds is identified and performance evaluated against the group. The regression-based method thus strengthens the value of this test from the perspective of avoiding adverse behaviours as it creates limited incentive to engage in herding behaviour.

This test gives rise to the following issues:

- There is heavy reliance on the declared G/D mix, as this determines where a fund is positioned along the regression line and hence the benchmark rate of return. This has a number of implications:
  - Unfortunately, there is currently no consistency in the G/D categorisation of assets across the industry for peer group analysis. APRA has a standardised G/D categorisation approach for regulatory reporting, but it is quite basic in nature and was formed without industry consultation.

[1] Whereas member outcomes may have been impaired under this situation, the right to continue operating should arguably not be taken away due to one well-considered position not happening to work out.

The approach to G/D categorisation should be addressed before the test is brought into effect (see Section 1.5)[2].

  - There is a large incentive to 'game' the G/D mix of the portfolio. The industry will aim to exploit assets that offer high expected returns relative to their categorisation.
  - The G/D mix would need to either be set by an independent party or closely reviewed by APRA. Funds should not be allowed to self-declare their own G/D mix.

- G/D is an imperfect risk measure. Indeed, it is an asset categorisation rather than a risk measure. The primary concern for members should be the risk of their balance ending up lower at the point of retirement. The implicit presumption is that exposure to growth assets increases this risk. This is only partly true, especially during the accumulation phase when most members continue to experience contributions. While growth assets heighten the potential for *very* poor outcomes over long horizons (relative to defensive assets where returns are lower but relatively reliable), growth assets tend to deliver higher expected returns that increases the probability of better outcomes[3].
- G/D categorisation does not directly recognise the benefit of diversification.
- Peer-based tests are concerned with relative performance and hence are not perfectly connected to actual member outcomes. The test only reveals whether a fund had done significantly worse than the peer group, not whether it has destroyed value for members. It is feasible that every fund adds value and deserves to continue, yet some funds may fail a peer test due to adding relatively less value and falling below the threshold. (The reverse occurs if the industry is destroying value – not enough funds will fail.) The underlying presumption of this test is that there must be something wrong with funds that significantly trail their peers. This will often hold, but not always.
- We feel that this test metric might be given a better name to avoid it being interpreted as a traditional peer-group test. Perhaps 'peer universe test' may be a more appropriate framing.

1.4.3 Risk-adjusted returns relative to SRP

This test metric also examines overall portfolio returns and thus links to member outcomes. The SRP benchmark represents the return that is hypothetically accessible to the member at equivalent risk at low cost, and without any investment skill. Effectively it benchmarks against outcomes that members could access for themselves. The metric rewards any activity that improves the risk-return trade-off of the portfolio, including identifying attractive assets not included in the SRP, successful dynamic asset allocation, value-adding active management and diversification. The test adopts standard deviation (SD) as a risk measure, raising questions around whether members are concerned with shorter-term volatility or whether SD is a relevant proxy for risk over the long term. The former is possible, the latter is questionable[4]. In any event, we consider SD as a reasonable risk proxy for the purposes of normalising risk and comparing returns against the risk-adjusted returns of a SRP. The fact that SD has shortcomings

[2] G/D categorisation is an area where the Conexus Institute has undertaken significant research in an effort to try and establish an industry standard. Materials are found on our *Growth/defensive asset categorisation* webpage.
[3] We are making a distinction between the *probability* of achieving a better outcome and the potential for losses of larger *magnitude* in the lower tail of the distribution. More detail on these concepts can be provided on request.
[4] This shortcoming is most relevant to trustee-directed product risk decisions, such as the appropriate risk target for a MySuper default, an area which believe should be covered in APRA's member outcome assessments and not in performance testing.

is a healthy reminder of the flaws inherent in any single metric and hence the merits of a multi-metric approach.

This test gives rise to the following issues:

- SD is not as straightforward to estimate as it may first seem. The estimation methods should be carefully considered by any TAG. Considerations include:
  - Smoothed returns for unlisted assets may understate risk. The question arises of whether measurement adjustments should be made, or if the implicit encouragement to invest in such assets is viewed as an acceptable consequence.
  - Long-term risk will be mis-stated if there is serial correlation in returns, be it positive (i.e. momentum) or negative (i.e. mean revision).
  - The data interval over which SD is estimated is not innocuous, e.g. whether say monthly, quarterly or yearly data is used. There are trade-offs, e.g. lengthening the measurement interval can limit the impact of serial correlation but reduces the number of observations.
- SD is an imperfect risk measure when investing for the long-term, as outlined above, but we consider it reasonable for this particular test. The issue raised above with respect to G/D applies to SD to the extent that more volatile assets also deliver higher returns, i.e. higher SD can indicate a higher probability of better outcomes but increased risk of particularly poor outcomes.
- Formation of the SRP is fundamental and should be considered by any TAG. For example, what Australian and international equity indices should be used, and what should be the local/international weight that is imposed? Should listed forms of other assets like real estate be included in the SRP? What defensive assets might be employed? APRA has a standard SRP used in the heatmap, but it hasn't been updated over time and there was no industry consultation during its development.
- There will remain some incentive to hug the benchmarks in the SRP for funds at risk of failure, i.e. when there is a low 'buffer'.

1.5 Implementation

1.5.1 Sort the details out later with assistance of a technical advisory group (TAG)

We recommend that an initial decision be made on the broad design of the test, with the intent of then sorting out the implementation details with the assistance of a TAG. The TAG should comprise parties with limited direct conflicts as far as practicable.

Below are some implementation details that might be considered with assistance from TAG.

- Metric design, particularly benchmark formulation including:
  - G/D categorisation and peer group universe selection
  - Structure of the SRP
  - Potential adjustments to indices used under the current test
- Estimation of SD for use in the SRP test
- Treatment of administration fees, i.e. continue with current year, or introduce a lookback period
- Calibration of failure thresholds, i.e. appropriateness of a 0.5% margin
- Failure criteria across three tests, i.e. is pass two out of three appropriate?
- Value of using multiple look-back periods in testing performance
- Scope of application, i.e. what products are assessed using the three metrics
- General review of investment market dynamics (e.g. emerging uptake of a 'new' asset class)
- *Note*: Other matters raised in Section 2 may also be considered.

1.5.2 Pathway to introduction

We suggest delayed introduction of any new testing regime, possibly targeting commencement with respect to FY2026-27 (i.e. over three years before the first assessment results are announced). This allows time for policymakers to formulate the tests and for industry to prepare. It also dents the retrospectivity of any new test. Recommended activities include formation of the TAG, and initiation of a G/D categorisation project.

Nevertheless, the general design and format for the test should be settled and announced as soon as possible, ideally as a result of the current consultation. Doing so will immediately influence behaviours, and hence help reduce the unintended consequences flowing from the current test. Knowledge of the nature of the impending test regime should reduce the reluctance of funds to take actions that may improve the risk/return profile but increase tracking error versus the current test as they start to focus on the implications for their test results 3-years down the track.

1.6 Review

The YFYS test should be subject to regular review, say every 3 years. This would help ensure that the test remains fit-for-purpose by leveraging off the learnings from implementing the test in practice, and adjusting in response to industry developments (e.g. investment approaches, benchmarking, retirement solution design) or evolving community expectations.

Responsibility

Our preference is that responsibility for review, maintenance and any reformulation of the test be taken out of the legislation and into the regulatory responsibilities of APRA under delegated legislation. We would like to see APRA undertake reviews with the assistance of a TAG.

Potential removal of current test

We suggest that the role for the current test be reconsidered as part of the initial review, possibly 2-3 years after any three-metric test has been in operation. Reasons include:

- Implementation performance is not directly connected to member outcomes
- Any contribution of implementation to member outcomes is implicitly embedded in tests that assesses total portfolio returns (i.e. the other two proposed metrics), reducing the call for a separate metric.
- The reliance on benchmark indices in the current test is problematic. The indices are often imperfect, they direct industry behaviour, and force the industry to rely on index providers who are placed in a monopoly position. Reducing the test to reliance on only a handful of broadly available indices would be helpful. (Indeed, using ETFs would remove the need for any indices.)
- The industry would have had ample time to adjust to the new tests by the time of the review.

Offsetting considerations include whether reducing to two tests might increase reliance on the efficacy of what are two similar tests, and whether it increases the scope to game the test if there are two metrics.

In any event, delaying the review of the current test would provide practical experience to gauge whether it is contributing anything meaningful to an expanded testing regime.

2 Alternatives

2.1 Metrics we ruled out

In assessing the design options, we ruled out the following metrics:

- **CPI-plus** – The outcome is largely driven by market outcomes (e.g. how equity markets and other risk assets perform). For instance, a strong bull market could see all funds pass while a bear market could result in large-scale systemic failures, regardless of the quality of investment decisions. Further, CPI-plus objectives are set by the fund itself, which raises agency issues.
- **Sharpe ratio** – Similar to CPI-plus, the outcome is largely driven by market outcomes and could result in wide-scale passes or fails regardless of the quality of investment decisions. We gave some consideration to a 'relative' Sharpe ratio, framed versus an SRP or peers. However, we ruled this out in favour of the two proposed tests for technical reasons[5].
- **APRA heatmap** – The heatmap involves a collection of metrics and assessment timeframes with differing relevance for member outcomes. This adds to complexity and makes it difficult to clearly define the criteria for failure. The heatmap might be used as a 'dashboard' if a party (e.g. APRA) was charged with determining whether a fund has failed. However, it would provide a poor foundation for an 'effective and efficient' quantitative test.
- **Promises to members** – Another possibility might be to hold funds to account for delivering on the promises published in the PDS. Generally we consider this approach would be undermined by a collection of agency, objectivity and measurement challenges. We note that one notable promise for multi-asset funds is to deliver a real return, which runs into the issue raised above for CPI-plus. There might be a stronger case to use promises to members for choice options where the implicit promise is to deliver to a well-defined benchmark, depending on how the promise is framed.

2.2 Analysis candidate industry solutions: comparing one, two, and three metric tests

Through general engagement, including sharing of the informal working group paper (see Section 4), we believe the super industry may be landing on two different positions of either retaining the existing test or moving to a two-metric test. The table below describes and analyses these positions alongside the three-metric test design proposed by the Conexus Institute and the informal working group.

[5] Adjustment for the risk-free rate is not innocuous as it influences the magnitude of the risk premium across assets. For example, assuming SD is consistent, it will favour defensive or growth assets depending on the realised spread over the risk-free rate. The Sharpe ratio also has an ill-defined distribution, making it difficult to calibrate the failure threshold.

Candidate industry solutions vs. the three-metric proposal

	Industry position 1: Maintain existing test	Industry position 2: A two metric test	Three-metric test
Description	Preserve the existing test (perhaps with minor modifications).	Existing test metric combined with another metric, around which there is no clear consensus. Metrics mentioned range from achieving CPI-plus through to the SRP metric. General position is that funds must pass one of two metrics.	Existing test metric combined with peer-referencing and SRP metrics, i.e. options 2b and 2c from the consultation paper. Funds must pass at least two metrics.
Assessment and member protection	No change. The test provides some member protections, with respect to admin fees and implementation performance. Total returns not tested hence does not align strongly with member outcomes.	Arguably a weakening of member protections due to the pass one of two element. Questionable efficacy of some of the proposed additional metrics (see Section 2.1) and the link to member outcomes.	Strengthens member protections by creating stronger alignment with member outcomes.
Capacity and incentives to manage and 'game' the test	Current test is being actively managed by industry. We see a high probability that very few and possibly no funds will fail in future as a consequence.	Incentive to manage the existing performance test metric is preserved under this test, as this would be sufficient to pass the test. Under a 'pass one of two' design, adding a second test would open up an additional avenue to manage or game the outcome.	A three-metric test with the requirement to pass two metrics that differ in design will be harder to manage and game. It should encourage funds to stop actively managing the test and focus more on member outcomes.
Comment on unintended consequences	The test creates a range of unintended consequences, largely captured in the consultation paper.	The unintended consequences should remain as funds are likely to anchor on the existing metric if it is sufficient to pass the test on its own.	A three-metric test places focus on multiple factors, which in aggregate align with member outcomes if two tests assess total returns. Combined with less ability to actively manage the test, this should create greater alignment with member outcomes and reduce unintended consequences.

2.3 Should there be a review process?

Concerns have been expressed about the 'bright lines' nature of the performance test. Two issues stand out. The first relates to the effectiveness of the test as an arbiter of the 'right to continue', and the implicit assumption that the test result is predictive of future performance. One concern here is that the test can generate Type I-style errors (i.e. a 'good' fund failing the test) due to shortcomings in the test itself. Another concern is that the source of the underperformance may have been addressed by the trustee. For instance, trustees may have taken some positive actions (for example, improving investment governance) that the test fails to recognise. The second issue relates to adverse incentives. The fact that failure is existential for funds leads them to view passing the test as an over-riding priority. This amplifies any unintended consequences that arise from attempts to manage the test, even if doing so is detrimental to member outcomes.

In previous submissions the Conexus Institute has suggested that a qualitative overlay would help counter both these issues. It would provide an avenue to identify where a Type I-style error has occurred as a consequence of shortcomings in the test design. It could also reduce the strength of the incentives

for industry to actively manage the test and hence dilute some of the unintended consequences by making a test failure less like a 'death sentence'.

We work on the basis that APRA assessing all super fund investment options via quantitative and qualitative methods is not a candidate option. (If so, there would be no need for a formal performance test!) A more limited alternative worthy of consideration would be to allow for some kind of **review process** of the initial test results with a view to identifying instances where the assessment may be misleading. Possibilities include:

- **Room for interpretation of results** – APRA or an assessment body would investigate the context of specific failures, and make a determination of whether to endorse the failure and declare the fund as 'underperforming'. This could be done on the basis of presumption of 'guilty, unless there is clear evidence to the contrary'. This approach may assist with specific product types (e.g. sustainability products) or where failure can be attributed to a specific performance event that occurred a long time ago. The primary benefit is that this may result in fewer instances of 'rough justice'.
- **Scope to appeal / show cause** – This is a variation on the above point where the fund presents its case to APRA or assessment body for consideration. The fund could be required to 'show cause' why failure should not be declared, imposing a high burden of evidence. It is likely that this would introduce a strong degree of contestability.
- **Reference to an established secondary metric set** – Secondary metrics may be examined upon failure of the primary test with a view to investigating whether a Type I-style error may have occurred. The additional metrics could include other relevant performance measures and assessment timeframes, e.g. investigating whether the underperformance is concentrated earlier in the evaluation period followed by improvement. A review process might also consider a broader set of indices linked to ESG and sustainability activities. Nevertheless, any additional information would require interpretation thus raising issues around resourcing, subjectivity, and possibly contestability.

Overall, we believe that it is possible to introduce a manageable review process to improve the efficacy of the assessment while not significantly diluting the bright lines aspect of the performance test. It is an area that we encourage policymakers to consider further.

2.4 Single-asset and choice products

The Conexus Institute views consumer protections as providing the greatest benefit when applied to products where engagement is weakest. We are thus supportive of applying a well-designed performance test to MySuper options and multi-sector options where the member is relying on the trustee to construct a balanced portfolio in accordance with their risk preferences.

The case for performance testing of single sector options is not as strong. Here the member (or their adviser) has made an explicit choice to invest in the particular product, likely as part of a diversified portfolio. In this situation, the multi-metric approach that we have advocated will not readily apply. These products may have objectives that do not link to traditional market-based benchmarks, e.g. maximising equity income, delivering an absolute return, etc.

As a general rule, we are not advocating for the test to be extended to single sector products. There are other avenues for creating an appropriate level of accountability, e.g. APRA assessment, possibly via heatmaps; trustee governance requirements. If performance testing is to be extended to single sector options, we would advocate for more tailored performance tests that might be aligned with the promise that the member has bought into. For example, an SRI equity option designed around a certain index might be benchmarked against that index. Thought should be given to designing test thresholds appropriate to the targeted tracking error of the product. For instance a high conviction equity fund targets far greater tracking error than an index fund, warranting different test threshold levels. We also recommend that the consequences of failure be given further consideration. There is an argument that consequences of failure for a choice product should be less harsh on the basis that the member has made

an explicit choice to invest in the product (likely as part of a diversified portfolio) and is more likely to be engaged in reviewing their portfolio.

2.5 Retirement

The Conexus Institute has written extensively of the large range of factors that contribute to a quality retirement solution[6]. Our key message is that assessment should focus on the overall retirement income strategy and the likelihood that it will deliver good member outcomes *looking forward*. Investment performance is only one component in delivering these outcomes, which will comprise income arising from a range of sources.[7] In retirement, investment performance does not play the centrepiece role that it does in the accumulation phase. Performance testing in the retirement phase of super hence would comment on only one of many important factors in delivering member outcomes.

It may be feasible to apply performance testing to the return-generating components of retirement solutions, e.g. account-based pensions. However, we have two reservations. First, the existing test does not apply particularly well to the way that investment portfolios should be managed for the retirement phase. For instance, retirement portfolios may be formed with a view to managing sequencing risk or maximising franking credit capture. Second, extending performance testing to retirement portfolios may act as a distraction: trustees may direct their focus towards one specific retirement activity (i.e. generating investment outcomes), when there is a need to develop a whole range of activities. The aim is to get the industry focused on delivering income in retirement, not just returns! We would prefer other policy and regulatory measures be used to assess retirement offerings, including a dedicated member outcomes assessment framework for the retirement phase.

If an improved performance test was developed in line with our proposal, we can see logic in extending this test to account-based pension products. However, the test should be positioned as one component that feeds into a broader assessment of retirement solutions. Under these circumstances, the test design may need to be revisited to ensure it is suitable for retirement products, and adjusted if appropriate. Another possibility is to develop a secondary analysis process of the type outlined in Section 2.3.

[6] For instance *"Assessing retirement income strategies… when outcomes are but a promise"*, *"How to Approach Quantitative Assessment of Retirement Income Strategies"*, and *"Investing for retirement"*.
[7] Income delivered by a retirement solution may arise from the Age Pension, possibly a lifetime income product, and a drawdown strategy that governs how income is drawn from accessible funds.

3 Socially responsible and ESG investing

Substantial concerns have been raised around the implications of the current performance test for investment activities associated with ESG, sustainability, climate, transition, and socially responsible investing (henceforth 'ESG/SI')[8]. The consultation paper has identified this theme. We first summarise our past research on the impacts from the current test on ESG/SI activities before considering the future under a revised testing regime.

3.1 Impact of the current test on ESG/SI activities

The Conexus Institute has explored the issue of ESG/SI activities in the presence of the performance test in detail. We summarise our research in the table below. The conclusion is that the performance test creates varying restrictions across the spectrum of ESG/SI activities that can be significant in some instances.

Summary of Conexus Institute research on the implications of the current test for ESG/SI activities

	1. Impact investing	2. Opportunistic investing	3. Exclusions	4. Engagement
Description	Investments targeted at a specific ESG/SI outcome, assumed to be in private markets.	Investments targeted to participate in ESG/SI themes, assumed to be in public markets.	Excluding specific investments based on values-based principles. Predominantly in public markets, but may also be in private markets.	Implemented investment strategy is supplemented with a range of engagement strategies to drive positive change
Impact of current performance test	While creates some tracking error constraints, there remains reasonable portfolio 'space' for allocations, albeit competing with other portfolio activities for part of the overall 'tracking error budget'.	Creates tracking error constraints, but these are measurable and similar in nature to other forms of active risk. Opportunistic investment competes with other portfolio activities for part of the overall 'tracking error budget'.	Creates *significant* tracking error constraints. For a typical growth portfolio, largest impact (by far) arises from Australian public equities.	This activity incurs no tracking error.

To summarise, the varying impacts of the current test on differing ESG/SI activities suggest a hierarchy of constraints:

- *Engagement* activities are not directly constrained by the performance test[9].
- *Impact investing and opportunistic investing* both incur tracking error that is broadly proportional to other active management investment activities. These ESG/SI activities are hence hindered by the degree to which super funds are prepared to allocate their constrained tracking error budget to these activities relative to other competing priorities (e.g. return seeking via off-benchmark assets, diversification, risk management, and so on).

[8] Faith-based investment activities face similar challenges. While we don't address faith-based challenges directly, we would propose a similar solution pathway to the one detailed.

[9] One caveat is that super funds may need to have the option to exclude companies through exiting the position in order to give force to their engagement attempts, i.e. exclusion can act as a threat.

- *Exclusions* can create an unsustainably high level of tracking error, especially with respect to Australian public equities. Thus this activity is most constrained by the current test. It is important to note that climate-transition aligned portfolios (e.g. Paris-aligned portfolios) require exclusions. Further, many investors in ESG/SI options are choosing these investment options based on personal values, and may be willing to sacrifice some performance.

3.2 Future considerations

Our proposed three-metric approach should help reduce some of the constraints around ESG/SI by focusing on risk-adjusted total returns under the two new metrics we propose. As long as an investment (or an exclusion) does not sacrifice returns proportionate to any change in SD or G/D exposure, then it will not raise the risk of underperforming either of these tests. Our proposed three-metric test is hence a partial solution to the hurdles created for ESG/SI investing under the current test.

Nevertheless, there is a much wider question around the role of investment activities with respect to ESG/SI that needs to be considered. The Conexus Institute has engaged with a variety of industry bodies and networks as well as funds that undertake ESG/SI investment activities in this space. Our understanding is that mandates and views differ significantly around ESG and sustainability goals. The inclusion of one or two additional benchmark indices will not resolve all of the issues faced by this part of the industry, as the issues go much broader. ESG/SI is a difficult area to resolve with respect to performance testing. We believe a solution needs to start with Government taking a position on two matters:

1. What ESG and sustainability investment activities are to be facilitated (and, conversely, what activities effectively restricted).
2. Whether consumers have the right to choose to invest their super in dedicated ESG/SI options on the understanding that investment could result in lower financial outcomes when measured over different timeframes.

With clarity around these matters, policymaker attention could then switch to creating testing environments that account for both MySuper defaults and dedicated ESG/SI choice options, and whether dedicated arrangements are required for the latter[10]. There are many candidate mechanisms that could be incorporated into a testing environment for ESG/SI options. There is no need to hypothesis on these mechanisms until the Government decides on the principles governing SG/SI investments.

[10] For instance, it could entail assessment against ESG/SI indices that align with the promise stated in the PDS, along with an accreditation framework.

Appendix

4 Working group note: Improving the YFYS performance test

The Conexus Institute was recently part of informal discussions with representatives of the asset consulting and research house communities to garner opinions on the Your-Future-Your-Super (YFYS) performance test ('the test'). The aim was to see if broad consensus could be reached around how the test should be redesigned. We hope that releasing this short paper might help in facilitating some consistency in submissions made to the Treasury in response to its current review. We want to avert a situation where Treasury is confronted with divergent views that supply no clear direction and might lead to inaction.

The Conexus Institute has penned this note as an account of where the group landed. The intent is to share the note more broadly in the hope of building wider consensus around the **broad design** of the test (e.g. which of the Treasury's options are preferred). There is a view that the **implementation details** might be best dealt with separately.

The discussion group is listed at the end of this note, with each providing input on a personal basis. The views and opinions expressed in this note do not necessarily reflect the views or positions of the entities represented by the members of the discussion group nor the views of any individual member.

Key areas of clear agreement

We received broad agreement on the following matters:

1. **Member outcomes lens** – The test should be designed based on what is best for members while being cognisant that impacts on industry cost impact member outcomes.
2. **The current test should be improved** – The question at hand is how to best design a backward-looking performance test. From this perspective, a major problem with the current test is that it is not well-connected to member outcomes as it assesses only implementation and ignores overall portfolio returns as well as risk. Further, it can encourage behaviours that are detrimental to member outcomes in some situations. There is ample scope to limit these shortcomings and hence improve the test, even if perfection is impossible.
3. **Key considerations** – Of a range of potential considerations, three are most important:
   a. Assessment should include total portfolio performance as it ultimately determines member outcomes.
   b. Disincentives to taking actions to improve member outcomes due to fear of failing the test should be mitigated as far as possible.
   c. Opportunity and incentive to manage (i.e. game) the test should be limited, leaving funds to focus on member outcomes.
4. **Three-metric approach preferred** – A test comprising multiple metrics incorporating a measure of overall portfolio returns is strongly preferred. Multiple metrics will help to address the fact that any individual metric is flawed, and reduces the capacity and incentive to game the test. The three-metric test (Treasury's option 3b) with the requirement to pass two-out-of-three is supported. Details on how the three-metric test might operate appear below.
5. **Technical advisory group** – Treasury should decide on the broad design of the test at this stage and appoint a technical advisory group to help frame up the implementation details. The group should comprise parties with limited conflicts as far as practicable.

Other notable points

- **Current test has become ineffectual** – There was a sense within the consultation group that it is highly probable that, beyond the second application of the performance test to trustee directed products (TDPs), it is very unlikely that any funds will fail the current test for MySuper going forward as it is now being managed.
- **Retain the current test, nevertheless** – While there are reasons to be concerned over its efficacy[11], it is probably best to retain the current test for now (preferably with enhancements) to help with acceptance and transition. The role of the current test should be reviewed at a later stage, e.g. 2-3 years after any new test is brought in.
- **Transition arrangements** – There was some concern within the consultation group over the potential impact on the industry of retrospectively introducing an entirely new test. Delayed or phased introduction and retention of the current test may help in this regard.
- **Confirming a failure** – Many groups, including the Conexus Institute, have argued in the past for a qualitative overlay, e.g. review by APRA, or a right of appeal. Most of the consultation group agreed that some sort of qualitative review of the test results would be beneficial. However the focus of the working group was whether an improved metric-based test could be developed.
- **Administration fees** – These should remain folded into the return metric rather than being a separate 'leg' of a multi metric test.
- **Test metrics that were not supported** – Members of the discussion group did not support the use of CPI-plus, Sharpe ratios or the APRA heatmap.
- **Addressing the details** – Below are examples of details that might be addressed with assistance from a technical working group:
  - Test metric design, including benchmark formulation
  - Use of multiple look-back periods
  - Calibration of failure thresholds, i.e. appropriateness of a 0.5% margin

Three-metric test design

The three preferred metrics are:

(a) **Risk-adjusted returns relative to a simple reference portfolio (SRP)** – This is Treasury's option 2c. This metric examines overall portfolio returns and thus links to member outcomes while including a risk adjustment for standard deviation. The SRP benchmark represents the return that is hypothetically accessible to the member at equivalent risk and low cost, with little or no investment skill.

(b) **Peer comparisons of risk-adjusted returns** – This is Treasury's option 2b. This metric also examines overall portfolio returns and thus links to member outcomes, but benchmarks against the estimated return for peer funds with an equivalent growth/defensive mix. The use of growth/defensive weights amounts to an implicit but imperfect form of risk adjustment.

(c) **Current test** – Retaining the current test would maintain some continuity and aid transition. Assessing performance relative to benchmarks across a range of asset classes also provides a different perspective by measuring the value add to members through implementation.

[11] Our consultation brought out some interesting comments around the ordaining of benchmark indices that places index suppliers like MSCI in a monopoly situation and forces funds to pay up. A suggestion that this situation might be brought to the attention of the ACCC received support from the consultation group.

A range of test metrics provide 'spotlights' on performance from differing angles. Doing so makes the test less exposed to the vagaries of a single metric. It also diminishes the capacity to manage (and game) the test, in a large part due to use of differing benchmarks. Inclusion of two new metrics that examine overall performance should shift the balance towards evaluating and hence encouraging actions that enhance total portfolio performance and hence member outcomes. Risk adjustments would provide credit for managing risk exposure, which can also benefit members.

All three metrics listed above have shortcomings. However, the existence of shortcomings for any single metric only strengthens the argument for multiple metrics. The aim is to design a better overall test that benefits members, while recognising that there is no perfect solution. We believe the outlined three-metric test is the best realistic way forward.

Authors

David Bell
Geoff Warren
Conexus Institute

Convening support

Aleks Vickovich
Conexus Financial

Consultation group

Andrew Boal
Nick Callil
David Carruthers
Richard Dunn
Ian Fryer
Kirby Rappell

Note date: 8 April 2024

5 Summary of Conexus Institute research on the performance test

We summarise various pieces of research undertaken by the Conexus Institute on the performance test that may be of use to Treasury. We are happy to explain any of this research in further detail.

Summary of research on the performance test by the Conexus Institute

Area of research	Description, summary of findings, reference
1. Type I / Type II error calculations	• We produced calculations for estimates of Type I errors ('good' fund failing the performance test) and Type II errors ('bad' fund passing the performance test). • Our research considers the impact of benchmarks, timeframes, and an additional return source (asset allocation decisions) that is not captured by the current test. • Reference: Section 2.5.2) of https://theconexusinstitute.org.au/wp-content/uploads/2023/10/20221017-YFYS-Submission-Conexus-Institute.pdf • Model available on request
2. Through-time performance management incentives	• Estimates the amount of tracking error required to avoid failure under different realised performance scenarios. • Identifies different scenarios that motivate tracking error reduction and, concerningly, tracking error increase. • Reference: Sections 3.1.2 – 3.1.3 of https://theconexusinstitute.org.au/wp-content/uploads/2020/11/YFYS-Detailed-Paper-20201127.pdf
3. Brief literature review on performance persistence	• Performance persistence is a subset of the active management debate, a highly contested part of the academic literature. Much of the analysis is based on US mutual fund performance, but provides some useful insights nevertheless. • Summary: (1) management fees are a drag on performance (Note: mutual funds charge retail fees); (2) some evidence of persistence amongst the poorest performing funds, partly due to high fee loads; (3) mixed evidence of sustained elevated levels of outperformance by individual investment managers over the long-term. • Reference: Appendix 1: https://theconexusinstitute.org.au/wp-content/uploads/2020/11/YFYS-Detailed-Paper-20201127.pdf
4. Sustainable tracking error and opportunity cost	• Estimates the rational sustainable level of tracking error that funds will end up taking. • Accounts for overlapping performance series, and the ongoing risk management of each performance series with a view toward potential future test results. • Sustainable tracking error is estimated to be around 1%. • Reference: https://theconexusinstitute.org.au/wp-content/uploads/2022/10/YFYS-Sustainable-tracking-error-re-visited-20221012-final.pdf
5. Active risk in super fund industry - sources and associated degrees	• Uses APRA heatmap data to analyse dispersion amongst funds in implementation performance and SAA performance (i.e. choice of markets, not overall level of risk). • Finds implementation risk is largest source of risk but SAA risk is also sizable. • Reference: Diagram 5: https://theconexusinstitute.org.au/wp-content/uploads/2021/05/20210525-Your-Future-Your-Super-Regulations-and-associated-measures-Submission-by-The-Conexus-Institute.pdf
6. Impact on industry as reported by fund chief investment officers (CIOs)	• Confidential interviews with ten super fund CIOs. • Majority will take less active risk as a consequence of the test, and believe that this will cost members. A view emerged that investment horizons have been shortened. • Conservative options and ESG/SI options were identified as particularly problematic. • Reference: https://theconexusinstitute.org.au/wp-content/uploads/2022/07/Final-survey-paper-20220726-Conexus-IM-Final.pdf
7. YFYS – Constraint on ESG, sustainability and carbon transition activities	• Collaborative research undertaken with FTSE Russell, ASFI and RIAA. • Estimated varying performance test tracking errors for different ESG/SI activities. • Portfolios that utilise exclusions (e.g. Paris-aligned portfolios) experience the greatest tracking error, exceeding the levels we identify as sustainable. • Reference: https://theconexusinstitute.org.au/wp-content/uploads/2022/11/YFYS-Performance-Test-Constraint-on-ESG-Sustainability-and-Carbon-Transition-Activities-20221109-Final.pdf

6 Responses to the consultation questions

The table below lists the consultation questions and provides a brief response to each, including linking back to our proposal and commentary in this submission. Our responses are not intended to be comprehensive, but hopefully are helpful nevertheless.

Responses to consultation questions

Consultation question	Response
Topic area 1: Options for reform	
1. Do you agree with these principles? Are there any other principles that should be considered?	• We broadly agree with the design principles, but decided to take another direction that was more in accord with our perspectives on the issue. • We found ourselves motivated by two purposes of performance evaluation (assessing performance, creating incentives) that suggest two objectives. We designed our criteria from this perspective. See 1.1. and 1.2.
Topic area 2: Status quo – SAA benchmark portfolio	
2. Is assessing the implementation of a strategy, as opposed to assessing the choice of strategy itself, a strength or weakness of the current framework?	• A central theme of this submission is that the performance test should better align with member outcomes. This means capturing as much relevant investment activity as possible. • We support 2b and 2c because they capture a broader set of activities; see 1.3 for further detail.
3. Can the existing methodology be materially improved, such as by further calibrating benchmarks, to largely address unintended consequences? How could these improvements overcome the incentive to benchmark hug, and remove barriers to invest in emerging asset classes?	• Adding more benchmarks improves the accuracy of the performance test as a measure of implementation performance. Additional indices will also result in less opportunities to manage or game the test and reduced unintended tracking error. • However, adding further indices comes with extra cost and complexity. • In previous research Conexus Institute estimated that over 50 benchmarks would be required to create an accurate benchmark-based approach.
4. What asset classes do you consider require better coverage in the test? What asset classes are covered well by the existing test?	• Many areas could be improved. We are sure that other submissions will make suggestions around unlisted assets and ESG/SI activities. • In terms of improving accountability and reducing portfolio constraints we advise additional indices in the credit spectrum and inflation-linked bonds. We would also consider changing the benchmarks for selected categories of alternatives (defensive, 'standard' and growth) to cash + benchmarks rather than stock / bond combinations.
5. Do you consider additional indices covering additional asset classes should be added to the test? If so, please provide the following details for each of your recommendations: a. Description of asset class b. Name of recommended index covering the above asset class, including the length of time data is available on the index c. Details of appropriate fee and tax assumptions for such an asset class d. Explanation of why you consider this index is appropriate for inclusion	• See response to Q4 and our reference in 1.5.1 to a Technical Advisory Group (TAG) to assist with implementation details.

6. How should the test cater for new asset classes in the future?	• See our recommendations re: TAG 1.5.1
7. Should the threshold for failure be recalibrated for some products? What evidence supports the need for a different threshold? How could a different threshold deliver better long term returns to members?	• While we do not recommend that performance testing is extended to single sector products • If performance testing is expanded to single sector products we recommend that testing methodologies, thresholds, and consequences are explored • See 2.4 for discussion.
8. Would retaining the current framework but moving to a simpler structure, such as a simple-reference portfolio of only bonds and equities, address some of the concerns with the current test?	• We are supportive of the SRP approach as part of broadening out the number of test metrics (see 1.3 and 1.4.3). We view any individual metric as a flawed approach as it will inevitably have shortcomings and motivates active test management.
Topic area 2: Alternative single-metric test – Risk-adjusted returns	
9. Would the Sharpe ratio be a more appropriate testing approach than the current framework? Would this lead to better member outcomes?	• We believe the Sharpe ratio has significant flaws, as discussed in 2.1.
10. How should the benchmark for performance be calibrated?	• We believe the Sharpe ratio has significant flaws, as discussed in 2.1.
11. What data should be used to estimate the Sharpe ratio, and how frequently?	• We believe the Sharpe ratio has significant flaws, as discussed in 2.1.
12. Are either of these approaches better than the existing test methodology (Option 1) or a simple Sharpe ratio (Option 2a)? Are there any other considerations that make this a better or worse option?	• We view 2b and 2c as stronger assessment metrics than the Sharpe ratio. However, we view a multi-metric approach as superior to a single metric as per 1.3.
13. Are there any other alternative single-metrics that would be superior in addressing the principles set out in this paper? How would they provide a better testing framework? What net benefits do they provide over other proposed metrics?	• We consider 2b and 2c as having good merit, as outlined in 1.4.2 and 1.4.3. Both metrics are superior to the current metric (see 1.4.1). As any individual metric has flaws, we recommend a multi-metric (see 1.3).
14. What incentives would these alternative single-metric options provide trustees, and what would be the consequence of this for member outcomes?	• Both 2b and 2c motivate a focus on delivering good outcomes for members via asset class selection and implementation (as explored in 1.3.4 and 1.3.5).
Topic area 3: Multi-metric test	
15. Would greater alignment to the APRA heatmaps improve the sophistication of the test?	• We believe the test should not be aligned with the APRA heatmap for reasons outlined in 2.1
16. Would it reduce incentives to benchmark hug and improve member outcomes?	• Being a multi-metric test, the APRA heatmap would be much harder to actively manage, thus limiting unintended consequences. However, we believe a better set of metrics is achievable, as outlined in 1.
17. Is correlation between metrics an issue? If so, how should this be addressed?	• Our working group discussions explored alignment between different metrics, concluding that there is a strong positive correlation partly because activities like implementation are subsets of broader activities. For instance, total portfolio returns reflect both asset class selection and implementation. • Focusing on correlation is asking the wrong question. This is a common issue in finance and investments. For example, when valuing a publicly-listed company it is common to use multiple valuation metrics. This approach is valuable when the measures DO NOT agree as it helps diversify away the shortcomings of individual metrics. When the measures agree it is typically a strong signal.

	• Individual metrics are not only flawed but can be actively managed, resulting in unintended consequences. For these reasons we recommend a multi-metric approach detailed in 1.
18. Should the test capture all the metrics in the heatmap? If not, what metrics?	• We recommend 2b (peer) and 2c (SRP), for reasons detailed in 1.3 and 1.4.2 and 1.4.3
19. How would the benchmark for performance be calibrated for chosen metrics? How would these metrics combine to determine overall pass/failure of the test?	• In our submission we recommend a three-metric approach with the requirement to pass two metrics. • Other approaches are either difficult (but not impossible) to calibrate or may increase opportunity and hence incentive to manage the test metric • We recommend that a TAG be appointed to assist with benchmark design and calibration (see 1.5.1)
20. What costs would be associated with aligning the test to the heatmap? What would be the benefits?	• We believe the test should not be aligned with the APRA heatmap for reasons outlined in 2.1
21. Would this framework improve the sophistication of the test? Would it reduce incentives to hug benchmarks and improve member outcomes?	• We recommend a three-metric approach as detailed in 1. We believe it creates a well-rounded assessment and reduces the opportunity and incentive to actively manage the test, which can create unintended consequences
22. Would this approach be more, or less, favourable than the heatmap approach?	• We recommend a three-metric approach (see 1) as having sizable net benefits compared with the heatmap approach, where we have concerns with (detailed in 2.1)
23. What would the costs of implementing this approach be? What would the benefits be?	• Implementation costs of this approach apply to industry and regulators. We note that the metrics in aggregate create better alignment with what funds *should be* targeting in their existing practices. • Implementation benefits include a better assessment (benefits members and industry) and reduced opportunity and incentive to manage the test, which may create unintended consequences
24. Are these the right measures of performance or are there other more important indicators of performance that should be measured in addition to or instead of those outlined? What metric should be used to assess these indicators?	• We recommend 2b (peer) and 2c (SRP), for reasons detailed in 1.3, 1.4.2 and 1.4.3
25. How should the benchmark for performance be calibrated?	• We recommend a TAG be formed to advise on technical details, as per 1.5.1
Topic area 4: Alternative frameworks	
26. How would an alternative framework be constructed according to the elements outlined above? Please provide specific details.	• There could be considerable benefit in making the consequences of failure less absolute. This might entail a review process. Possibilities include a limited qualitative overlay, right of appeal or secondary metrics. See 2.3. • Any review might be undertaken by APRA. The efficacy of this could be managed by establishing situations that will be considered, and information that will be used.
27. How would this framework more effectively advance the principles outlined in this paper?	• A review could limit the instances of 'rough justice', and help to reduce trustee focus on actively managing the test, which motivates unintended consequences
28. What would be the costs and benefits associated with this framework, compared to the current test and any other alternatives?	• A review process would introduce additional cost. • A review process could undermine member confidence in the test if not communicated well. However. we believe is would ultimately strengthen the test and the message around failing funds more definitive

Topic area 5: Broader considerations for reform	
29. What are the most important considerations for performance of retirement products?	• We discuss the challenges of assessing retirement strategies in 2.5. Investment performance is only one component in delivering these outcomes, which comprise of income arising from a range of sources. In retirement, investment performance does not play the centrepiece role that it does in the accumulation phase. Performance testing in retirement would comment on only one of many important factors in delivering member outcomes.
30. If the test were to expand to retirement products, would they require a different test to the accumulation phase? Would the test differ for different retirement products?	• As explored in 2.5 it may be feasible to apply performance testing to the return-generating components of retirement solutions, e.g. account-based pensions. However, we have reservations around applicability given the need to manage retirement portfolios differently, and the test acting as a distraction from the need to develop a whole range of activities.
31. How could longevity products be most appropriately assessed? How could the products be compared?	• We believe it is difficult to extend performance testing to longevity products. It is important to assess the overall retirement strategy being offered, an area we explored in detail in *"Assessing retirement income strategies… when outcomes are but a promise"*, *"How to Approach Quantitative Assessment of Retirement Income Strategies"*
32. Do you agree that retirement phase, single-sector and externally-managed products are suitable for testing? Why or why not?	• We have reservations about testing account-based pensions, especially under the current performance test (see 2.5) • We do not recommend performance testing be extended to single-sector products. There are other fiduciary and regulatory governance mechanisms that better complement this end of the choice spectrum (see 2.4). • We do not consider externally-managed products. Our working view is that the test should be member-focused, hence cannot see how these products should be treated differently.
33. Should different assessment methods be applied to different cohorts of products?	• Yes, see discussion in Q32 • There are additional challenges for ESG/SI options, as detailed in 3
34. Do you agree that the 'other products' outlined above are unsuitable for testing? If you think the 'other products' (or a sub-section of these products) are suitable for testing, how could they be appropriately tested?	• See discussion in Q32 • There are additional challenges for ESG/SI options, as detailed in 3
35. Under each design option, how could the test accommodate cohorts that are suitable for testing? For example, using different metrics or benchmarks for performance for different cohorts.	• See discussion in Q32 • There are additional challenges for ESG /SI options, as detailed in 3
36. How should fees be measured under each design option?	• We believe all metrics should be holistic in nature, incorporating investment fees and administrative fees
37. Should fees be measured at the current option level, or should they be measured on a different level? How would this be achieved?	• We have no response to this question.
38. Are the current assumptions made in comparing fees acceptable? For example, should the $50,000 representative member balance be adjusted based on the median member balance for a product cohort?	• We don't like this assessment approach as (1) many members with higher balances may not be experiencing the best outcome; and (2) we are anecdotally aware of funds actively managing administration fee pricing to achieve a favourable performance test outcome.

	• Nevertheless, it is hard to come up with an alternative framework that can be implemented efficiently.
39. Is a peer comparison of fees the best way to measure fees? Is there a better approach to benchmarking fees? If so, how should this work?	• This approach is flawed because it doesn't consider the breadth and quality of services provided. • However, we do not have a superior approach to recommend.
40. What product cohorts should be considered? How should different cohorts be defined where products could meet multiple cohort definitions, such as single-sector retirement products?	• We have not considered this question in detail and so do not offer a response.
41. How many years of fees data is appropriate to test? Should a greater weighting be given to certain years?	• We believe forward-looking administration fees are a better indicator of future performance, which should be the priority. • However, it is important that these fees are sustainable. While APRA undertake sustainability analysis, we are not certain that it effectively 'polices' fee settings. • An alternative view is measure a shorter timeframe (say three years) including the current year. We believe this approach has merit.
42. Should the consequences be adjusted to improve outcomes for members? How would this need to be tailored for the different options for performance testing?	• We consider this specifically for the test in general in 2.3 (where we suggest considering a review process) and for single sector products in 2.4
43. How should the consequences be amended to better account for edge cases or different cohorts that fail the test for reasons beyond the trustee's control?	• We recommend a review process be considered in 2.3
44. How could these provisions be effectively ring-fenced so that it applies only to the edge cases and not failures at large?	• We recommend a review process be considered in 2.3
45. How could this be achieved without subjecting the regulator to undue challenge and impacting the efficiency of the regime?	• We recommend a review process be considered in 2.3
46. What other remediation processes could occur?	• We recommend a review process be considered in 2.3
47. Are there any key barriers to consolidating closed and underperforming products? What quantitative evidence is there of these barriers? How do these weigh against other reasons a person may choose to remain in a product?	• We encourage Treasury / APRA to analyse the net flows experience of funds being consolidated post-announcement, pre-closure. Our initial analysis (Alcoa, AvSuper and CBA Officers Super) suggests instances of sizable outflows. This makes the business case for consolidation weaker for the 'acquiring' fund.
48. What evidence do trustees use to demonstrate that remaining in a closed and underperforming product is in the best financial interests of members, compared to moving to a performing product?	• We have no comment.
49. What is the process or criteria that trustees use when deciding on what product they will transfer members to when consolidating underperforming products?	• We have no comment.
50. Should APRA receive increased regulatory powers to direct superannuation trustees to consolidate underperforming products?	• We have not considered this issue in detail, but we broadly support this where a clear case can be made by the regulator.