Treasury consultation: Strengthening the superannuation performance test

Submission by The Conexus Institute

26 June 2026

David Bell and 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 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 Research Fellow at The Conexus Institute and an Honorary Associate Professor with the Australian National University. Warren's career has contained distinct industry and academic phases, with investment and superannuation being two particular areas of focus. He has worked with commercial investment firms and government centres, while his academic career has predominantly been at Australian National University. Warren is a member of various advisory boards, including the ASIC Consultative Panel. Full bio here.

***** The authors are willing and able to participate in further consultation. Please treat this as a public submission. *****

1. Our main recommendations

1. **Building a more sustainable performance test** – We strongly recommend focusing on what is required for a more sustainable test for an evolving industry, rather than tinker with the existing test to address particular issues. Our preference would be to introduce either of the following options, with (A) preferred and (B) being our recommended fall-back if a bright line test regime is to continue:
   (A) Empower and resource an independent authority with discretion to oversee performance assessment. This authority could be either APRA or a newly established body.
   (B) Introduce a multi-metric test regime (as recommended in our 2024 submission).
2. **Option 1: Adjusting for emerging and alternative asset classes** – We are not against either introducing a new emerging covered asset class (i.e. option 1.1, the 'side-pocket') or broadening the existing Alternatives covered asset class (i.e. option 1.2). However, we doubt either would spur any significant uplift in investment into emerging and alternative assets classes.
3. **Option 2: Introducing assessment of risk-adjusted returns** – While we support the intent of introducing assessment of total portfolio returns with a risk adjustment, we recommend against moving to a metric based solely around a simple reference portfolio (SRP) or indeed any single-metric test, which will create its own unintended consequences. If a multi-metric test is not deemed to be an option for assessing risk-adjusted returns, we would recommend retaining the existing test.
4. **Option 3: Introducing a routine review of the benchmarks** – We support the proposal to establish a routine review of benchmarks. We consider it important that reviews be conducted by a body that is structurally independent from the political process and the superannuation (super) industry.
5. **Option 4: Expanding the performance test** – Here we are hesitant to make strong recommendations. We see a case for applying the test to externally-directed products (EDPs) and account-based pensions (ABPs) where they are clones of existing tested accumulation options and not embedded within integrated retirement solutions, but are equivocal over testing EDPs. We do not support expanding the test to separately managed accounts (SMAs), single sector products or ABPs where they are explicitly designed to meet the needs of retirees.
6. **Apply a multi-phase process for any adjustments to test design** – We suggest that the current review focus on the broad nature of the test, with any recommended changes brought into effect through a multi-phase process involving: (a) mandating broad adjustments following the current review; (b) scoping out implementation details with the assistance of a technical advisory group (TAG) or groups; (c) further consultation over the implementation details; and then, (d) finalising through specification, legislation if required, and introduction of any re-designed testing regime.

2. About this submission

This submission is written from the position that the most important policy decisions at this stage involve the broad design features of the Your-Future-Your-Super performance test (which we refer to as either the YFYS test, or just 'the test'), specifically the test structure and its management. As a consequence, our submission focuses on higher-level issues. While we suggest that design and implementation matters be settled later with input from expert working groups and further consultation, we nevertheless selectively discuss design and implementation issues. This is done to highlight the complexities involved and the potential implications, which serves to underline the need to tread cautiously and seek further expert input before proceeding to adjust the test structure.

From this position, the subsequent sections address the following:

- *Section 3* provides broad reflections on the (limited) scope of the current consultation, the need for a bright line test, and two directions for creating a more sustainable testing regime.
- *Section 4 through to Section 7* discuss the four options put forward in the consultation paper, focusing on the merits or otherwise of each option and offering comments around the issues that need to be addressed if an option is to be pursued.
- *Section 8* recommends bringing any adjustments to the test into effect through a multi-phase process.

- *Appendices* present supplementary material on broad reference portfolios and the trade-off faced by funds regarding mid-risk assets under a simple reference portfolio.

We also thank Treasury for referencing the Conexus Institute three times in the consultation paper (on pages 8, 9 and 21). The acknowledgement is appreciated, and it is good to know that our research may have proved of some value.

3. Broad reflections

We commence with overarching comments on the current consultation and how we would like to see the review of the test proceed. Section 3.1 appeals for the scope of the review to be broadened. Section 3.2 reflects on the original reasons for introducing a bright line test regime and whether it remains appropriate. Section 3.3 suggests two directions for creating a more sustainable testing regime going forwards, both of which go beyond the scope of the current consultation.

3.1. Scope of the review

Our overarching position is that the most important consideration is the broad nature of the test, rather than the specifics of the test metrics and their design. From this perspective, we consider the scope of the current review to be too narrow. In our view, focusing more on 'the forest rather than trees' is critical at this juncture with the intent of building a more sustainable testing regime.

Our sense is that the current review and its proposed options are directed towards the following:

1. Addressing, or at least limiting, the extent to which the test constrains investment in certain 'nation building' areas including climate transition, housing and innovation. This aim is reflected in option 1.
2. Limiting the extent to which the test design encourages 'benchmark-hugging' around specified asset class benchmarks and thus implicitly directs and constrains the investment activities of super funds. This aim is reflected in option 1 and (we surmise) also option 2.
3. Broadening the scope of the test where prudent to do so. This aim is reflected in option 4.
4. Threading the needle with palatable options in the face of strong and disparate lobbying efforts from a range of stakeholders. This is reflected in the particular suite of options being proposed, which appear limited to those that seem 'achievable'. It also seems reflected in option 3 under which a formal and pre-defined process for review of benchmarks would be established that may be less prone to lobbying influences.

While we see merit in the implied aims and some of the proposed options, to be frank, we do not consider them to amount to a major set of reforms that will place the test on a sustainable footing. This seems at odds with the more expansive intent of the *Annual Superannuation Performance Test – design options* consultation of March 2024, and the Treasurer's comment that "*any changes must be enduring to set the test up for long-term stability*" (covered by Investment Magazine here). Without major reform we expect that, in time, it will become necessary for the test to be reviewed yet again. Our suggestions for putting the test on a sustainable footing are outlined in Section 3.3. They would require policymakers to take a stand rather than trying to trace out a route to reform that encounters limited resistance.

Further, we sense that the proposed options may not achieve much in working towards the implied aims as listed above. We expect option 1 to have at best a modest impact on increasing investment in emerging and alternative asset classes. Option 2 would merely change how super funds manage the risk of test failure and hence potential nature of benchmark-hugging, which could unfold in a way that could run counter to supporting investment in emerging and alternative asset classes under certain circumstances. Option 2 will also be difficult to design in a way that delivers an effective test while limiting the potential for unintended consequences, which could have weightier implications than those stemming from the existing test. Although option 3 and option 4 will have beneficial impacts, these should be marginal rather than profound. Our reasoning is set out in Sections 4, 5, 6 and 7.

Finally, all four options require careful design as each involves addressing some challenging issues that are often complex and technical. We doubt that these issues will be adequately resolved through the current review process. Accordingly, we recommend that the current review largely focus on deciding

on how to take the test forward, with implementation of any decided adjustments involving input from experts and further consultation. Our thoughts on this matter are outlined in Section 8.

3.2. Reflections on the need for a bright line performance test

In exploring the options for revising the test, it is useful to reflect on some of the pre-conditions that led to the Productivity Commission (PC) recommendation to implement a bright line performance test following its 2018 review. These include:

(i) **Performance dispersion** – The PC had uncovered statistical evidence of a large dispersion in performance across super fund MySuper options, including a tail of funds that were delivering significant underperformance versus their reference portfolio benchmarks. The initial introduction of the test had much to do with clearing up this 'tail of underperforming funds'.

(ii) **Fee dispersion** – Evidence also existed of sizable fee differences across funds, with the implication that certain funds may have been 'over-charging' their members.

(iii) **Comparative data** – Data provision by the sector was poor, making informed comparisons between funds difficult to undertake. The test provided a vehicle to introduce a source of consistent performance data across the industry.

(iv) **Regulatory capabilities** – There was limited confidence in the ability of regulators to take action to ensure consumers did not experience prolonged exposure to underperforming and/or high-fee funds. Introducing a formal, bright line performance test helped to fill this gap.

A number of these pre-conditions have dropped away, particularly with regard to MySuper and trustee-directed options (TDPs). (*Note*: Section 7 discusses these pre-conditions in the context of candidate areas for test expansion.) Since the introduction of the test, we have seen funds exit the industry, convergence in performance and fees across the sector, and APRA collecting far more data on fund performance. Further, the future protective benefits of the test have diminished as super funds have worked out how to 'manage' their test outcomes to limit the risk of failure. Meanwhile, many of the 'unintended consequences' that were mooted at the time – such as incentives for benchmark-hugging and constraints on investing in innovative, emerging and alternative assets – have become more evident. In short, much of the potential upsides from introducing a bright line test with existential consequences have been secured, while many of the downsides remain prevalent.

This leads us to the existing review, which as discussed in Section 3.1, aims to address some of the specific shortcomings of the test but has pulled back from whether the very nature of the test should be reconsidered. *We maintain that the nature of the test itself is the core issue to be addressed.*

3.3. Two possible directions for establishing a sustainable test

We see two ways forward for establishing a sustainable test, being (A) empower an independent authority with discretion to oversee the test and/or (B) move to a multi-metric test regime. Both offer scope to address some of the issues that are motivating the current review, including index-hugging, constraints on investment in emerging and alternative asset classes, accommodating a broadening of the scope of the test in a coherent manner and limiting potential for unintended consequences that can arise from a single test metric. Our preference would be to empower an independent authority that draws on a range of quantitative and qualitative assessment approaches in making determinations. A multi-metric test regime would be our second choice for retaining a form of bright line test, and superior to the options currently on the table all of which are built around a single test metric.

3.3.1. Independent authority overseeing the test with discretion

Our preferred approach would be to replace the bright line test based around a single backward-looking metric with an 'independent authority'. This authority would be empowered and resourced to assess performance and then make determinations on fund performance, ideally with limited scope for contestability. We envisage the authority would be responsible for determining how performance is to be assessed (i.e. test design) and would combine quantitative and qualitative assessment techniques with scope for interpreting the evidence in order to arrive at a determination. APRA would be a strong

candidate to undertake this role. However, if for various reasons APRA was considered unsuitable, then a separate independent entity could be established. This approach offers a number of benefits:

(i) **Forward-looking** – An independent authority with discretion can introduce a forward-looking assessment regime that accounts for strategic changes in how investment portfolios are managed, perhaps made to address the sources of poor past performance. For instance, changes in investment philosophy, fund governance, resourcing including staffing, passive versus active investing, public versus private asset exposure, degree of internal management and so on can be far more relevant for future member outcomes than past performance. In effect, an independent authority can refer to the current state and not just examine the past state as reflected in observed returns. For example, the authority could overlook poor performance sourced from 5-10 years ago that has led to significant changes in investment operations that subsequently deliver solid returns.

(ii) **'Fairer' test** – An independent authority can deliver a fairer test by considering and interpreting the available evidence around performance within the context that it was delivered. We envisage an independent authority considering multiple quantitative performance metrics supplemented by qualitative analysis. This might be similar to APRA reviewing and interpreting the body of metrics under what was previously the heatmaps and is now known as the APRA's *Comprehensive Product Performance Package* (CPPP). For example, there should be more room to make allowance for off-benchmark investments, e.g. sustainable investment options that accord with member choices, or investments that have not yet matured causing returns to fall marginally below a cut-off.

(iii) **Capacity to update the assessment process as and when required** – An independent authority with responsibility for test design can adjust the performance assessment process as the super system evolves through time. For instance, changes may occur in available assets, benchmarks and the range of investment activities being undertaken, e.g. some large funds are now undertaking a sizable degree of collateralisation activities. A number of funds will be managing close to a trillion dollars by 2035, which undoubtedly will impact on industry practices.

(iv) **Potentially to inform other APRA activities** – Another benefit if APRA was the responsible authority is that the act of undertaking detailed performance assessment would inform other aspects of their industry engagement activities, such as oversight of fund governance and investment activities. Management of performance assessment may also support consideration of the implications of the assessment regime for the systemic risk profile posed by the super system.

(v) **Extracting from the political process** – Establishing an independent authority to undertake assessment would take the maintenance of the testing regime outside of the (fraught) policy review process involving engagement through Treasury and inevitable lobbying, which can potentially require legislative change.

An independent authority with discretion is, of course, no panacea. There are many issues with allowing scope for interpretation, including opening up opportunities for regulatory capture and potentially appeal of any determinations. Nevertheless, we feel this should give rise to fewer problems than a bright line test based on a single metric provided that the independent authority is unbiased and appropriately skilled to interpret the analysis of fund performance with deep understanding.

While we see merit in APRA acting as the independent authority that oversees the test, there may be reluctance for APRA to enter the realms of making determinations around product suitability. APRA may view a performance assessment function as sitting outside their role as a prudential regulator. A newly established body would overcome this hurdle and may have the advantage of being better placed to appoint a targeted group of experts.

We realise that this proposal could involve a radical change in the operation of test, especially if it entails interpretation of metrics or a new established entity. However, the benefits could be substantial, including extracting the test from what has proved to be a fraught political process once and for all.

3.3.2. Introduce a multi-metric test regime

We recommended a multi-metric test regime in our submission to the consultation on *Design options for the annual superannuation performance test* in 2024. Our proposal was for a three-metric test comprising the SRP test with standard deviation (SD) as a risk proxy, a peer-relative test using growth-

defensive (G/D) weights as risk proxy and the existing test, requiring a pass on two out of three metrics to avoid failure. Our (substantial) discussions of the multi-metric test can be found not only in our 2024 submission but also subsequent notes we provided to Treasury on a private basis during December 2025 (some of which has been reproduced and updated in this submission).

We believe that a multi-metric test would be a far more appropriate way forward than any single metric if a backward-looking bright line test is to be retained. Our position is that *all* performance metrics have issues that can result in misleading assessment and particular incentives. However, the issues differ in nature across various test metrics thus diluting exposure to the nuances of any particular metric. Our proposal for a multi-metric test revolves around the following core ideas:

(i) **Differing spotlights** – A multi-metric test might be conceptualised as placing spotlights on performance from differing directions, i.e. examining performance from various perspectives. Ideally the performance metrics would be designed to capture a range of considerations that matter for member outcomes, e.g. differing types of risks that matter to members.

(ii) **More robust to flaws** – A multi-metric test would be more robust to the issues that arise with any single test, as it would tend to 'diversify' the differing flaws of individual tests.

(iii) **Diluted incentives for benchmark-hugging** – Benchmark-hugging incentives should be much diluted under multiple test metrics that adopt benchmarks which differ in the behaviours they reward. The degree to which this would occur may depend on the extent to which trade-offs between the metrics are involved. Funds would be incentivised to pursue investments that look attractive when viewed from all angles, and limit exposure to those that improve the prospect of passing one test while worsening the prospect of passing another.

(iv) **Examining total portfolio performance using differing risk proxies** – Through including a focus on total portfolio returns with differing proxies for portfolio risk, a multi-metric test offers the scope to assess value created or lost by asset allocation decisions along with risk adjustment from multiple perspectives. For instance, risk might be proxied by SD as a measure of portfolio volatility and G/D weights as a measure of underlying economic exposures. (Section 5.1.3 discusses how these two proxies capture differing and complementary perspectives on risk.) By contrast, the existing test largely assesses implementation of a fund's declared strategic asset allocation (SAA) and does not account for the risk taken in pursuing returns.

(v) **Reducing incentive to focus on 'managing' the test** – Making the test harder to manage (i.e. game) would incentivise funds to focus more on delivering the best possible overall portfolio outcomes and less on managing the test itself.

Notably many of the above features of a multi-metric test would help address some of the issues raised in the consultation paper, including easing the constraint on investing in emerging and alternative assets and incentives for benchmark-hugging, both of which arise from concern with tracking error (TE) relative to a single performance benchmark. A multi-metric test should also be more sustainable. It would not be as exposed to the unintended consequences that may arise from imposing a single metric, and so should be more robust and thus less prone to ongoing calls for reviewing the structure of the test.

The main disadvantages of a multi-metric test largely relate to the increased complexity involved in designing and managing the test, and lower understandability for members (although we doubt most members understand the existing test in any event). We acknowledge that this complexity will make it more difficult for funds to characterise and manage exposure to the risk of test failure. However, we view making the test more complex for funds to manage as a strength rather than a weakness.

In short, we consider a multi-metric test as far superior to both option 1 and option 2 as presented in the consultation paper, and more likely to achieve what those reform options appear to be setting out to do. Multiple metrics are well worth the additional complexity in our view.

4. Option 1: Adjust for emerging and alternative asset classes

While we are not against adjusting for emerging and alternative assets classes, we doubt it would spur any significant uplift in investment into these areas. Nevertheless, proceeding with this option may be useful to provide industry with some additional flexibility to invest. We have no strong preference for

option 1.1 versus option 1.2. Considerations under either approach include maintaining visibility around the type of assets that are included in any emerging and alternative asset class category based on their 'nation-building' credentials, formulating any cap on these assets, and the importance placed on maintaining consistency with the existing test structure based around market-based benchmarks.

4.1. Option 1.1: A new emerging covered asset class ('side-pocket')

We start by making some observations on CPI+ benchmarks and then proceed to comment on implementation considerations. We also offer an alternative idea of benchmarking against prevailing interest rates through a cash+ benchmark or perhaps market-based benchmarks rather than CPI+.

4.1.1. Observations on CPI benchmarks

The Conexus Institute examined CPI+ benchmarks in a report titled *What is the role of CPI+ investment objectives?* in February 2025. While this research focused on the use of CPI+ as an overall portfolio objective, many of the observations made also carry over to the side-pocket proposal. Our position is that CPI+ is an inappropriate benchmark for assessment of how much a super fund has contributed to member outcomes through its own actions. A core principle is that performance evaluation should be based on elements that are controllable by those being assessed. However, super funds have no control over the absolute real returns delivered by markets at large. Any portfolio of assets (including potentially those included in any side-pocket) is highly exposed to market (i.e. systematic) risk, such that broad-based weakness in asset markets should generate under-performance versus a CPI+ benchmark. This would be especially the case if the weakness occurs in the context of higher inflation. Conversely, ongoing market strength would make the CPI+ benchmark easy to beat. In sum, the side-pocket is likely to carry exposure to the real returns that markets deliver over which a super fund has limited control. Section 4.1.2 presents analysis to illustrate this point.

This issue may be mitigated to some extent where individual assets provide an effective inflation hedge, so that the assessment reflects the real return delivered by those assets thus diluting the impact of inflation shifts on realised returns. However, such assets are rare. Further, many purported 'inflation hedges' (such as inflation-linked bonds and certain types of infrastructure) tend to offer inflation-linked cash flows but remain exposed to changes in market pricing parameters (i.e. discount rates and valuation multiples) and how they respond to inflation innovations. See our CPI+ report referred to above for further discussion on the challenges of hedging against inflation.

Against this background, three main issues emerge with respect to CPI+ benchmarking in the context of the YFYS test, to the extent that real returns on investments within the side-pocket reflect broader market returns and/or inflation developments rather than idiosyncratic real returns:

1. A side-pocket assessed against a CPI+ benchmark has potential to be a significant source of TE, albeit depending on the weight it is assigned;
2. CPI+ performance assessment can be a poor measure of the value added by super funds through asset selection within the side-pocket to the extent that performance reflects broader market returns; and,
3. By assessing absolute real returns rather than relative return versus a counterfactual investment in an index, the conceptual foundations of a CPI+ test sit at odds with those behind the existing test.

4.1.2. Illustrative analysis of the side-pocket as a source of TE

The TE issue is particularly pertinent given the motivation behind option 1, i.e. to ease the constraints that the test places on investing in certain emerging and alternative asset classes. To the extent that a super fund cares about TE, a side-pocket that is benchmarked against CPI+ could be viewed as a 'risky' investment from the perspective of a super fund. But how risky?

We conduct some analysis to help gauge the degree of TE risk at the total portfolio level that might be associated with a side-pocket benchmarked against CPI+. The analysis entails constructing a notional historical simulation comparing returns on a balanced portfolio comprised of 70% equities and 30% fixed income against the same portfolio combined with a side-pocket that is benchmarked against CPI+3%. We assume that the balanced portfolio and the side-pocket both earn the same return which in turn aligns with the YFYS test benchmark, such that all performance deviations and hence TE arises

from benchmarking the side-pocket against CPI+3%. The simulation uses quarterly total return indices commencing in December 1972 to span as many investment cycles as feasible[1]. We examine portfolios with side-pockets weighted at 5%, 10%, 15% and 20%. The breakout box details the data and method.

Side-pocket historical simulation analysis – Data and method

Data

Quarterly total return indices for December 1972 to March 2026 were created from the following data series, for conversion into rolling compound returns over time frames of interest:

- ***World equities, unhedged*** – US$ total return index translated at A$/US$ exchange rate. *Data sources: LSEG (Datastream) for world equities and RBA for A$/US$.*
- ***World equities, hedged*** – Quarterly hedged return series was generated as: *US$ world equity return less US T-bill yield plus Australian 90-day bank bill yield*, where the yield differential is being used as a proxy from the cost of hedging and is observed at beginning of each quarter. The return series were accumulated to generate a total return index. *Data sources: LSEG (Datastream) for world equities, St Louis Fed for US T-bill yields, RBA for 90-day bank bill yields.*
- ***Cash*** – Returns based on 90-day bank bill yields observed at beginning of quarter were converted to quarterly returns and accumulated to form a total return series. *Source of bank bill yields: RBA*.
- ***Bonds*** – A quarterly return proxy was formed based on Australian 10-year government bond yields, assuming a bond priced at par and paying yearly coupons equal to the redemption yield is purchased at beginning of each quarter and then revalued at the end of the quarter as a 9.75-year bond using the prevailing bond yield (i.e. entire return comprises bond price change, without any coupons). Bond returns were accumulated to form a total return index. *Source of bond yields: RBA.*

Portfolio construction

- 70/30 balanced portfolio was created with weights of: 20% in world equities, unhedged; 20% in world equities, hedged; 30% in Australian equities; 20% in bonds; and 10% in cash.
- Quarterly rebalancing was assumed

Side-pocket analysis

- Side-pockets of 5%, 10%, 15% and 20% were examined
- Both the balanced portfolio and the side-pocket were assumed to earn the same return as the balanced portfolio and YFYS test benchmark, such that all deviations from the benchmark return and hence TE arise from the side-pocket being benchmarked differently.
- Side-pocket is benchmarked against the Australian CPI+3% accumulated over the period of interest. *Source of CPI data: RBA*.
- Returns over 10-year rolling periods were estimated and benchmarked, with 5-years also examined for a guide of the path that might be experienced over the interim.

Output:

- ***TE*** - Estimated as standard deviation of return differences between portfolio with and portfolio without side-pocket across all 10-year (and 5-year) rolling periods during the analysis period[2].
- ***% of periods of test failure*** - Portion of 10-year (and 5-year) rolling period where returns including side-pocket are less than -0.50% below the return on the balanced portfolio.
- ***Charts*** - Plot differences in rolling returns between the portfolios with and without the side-pocket, and while quoting TE at various side-pocket weights in the legend.

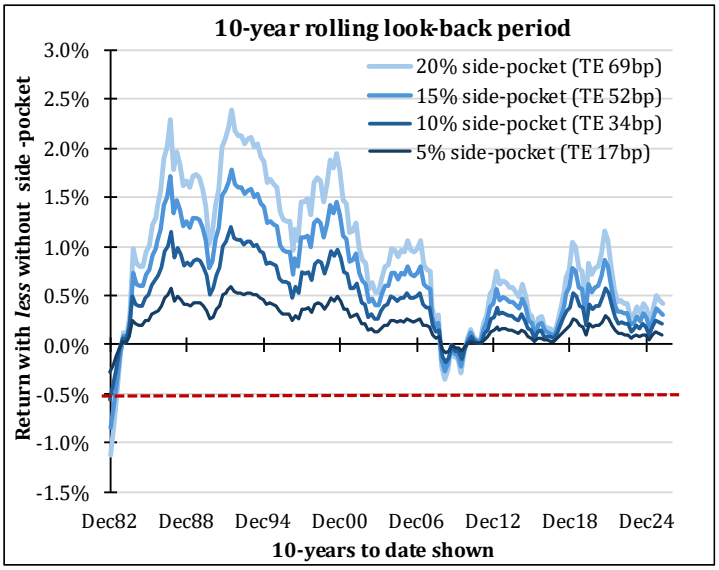
[1] We were constrained by lack of data availability from going back further in time.

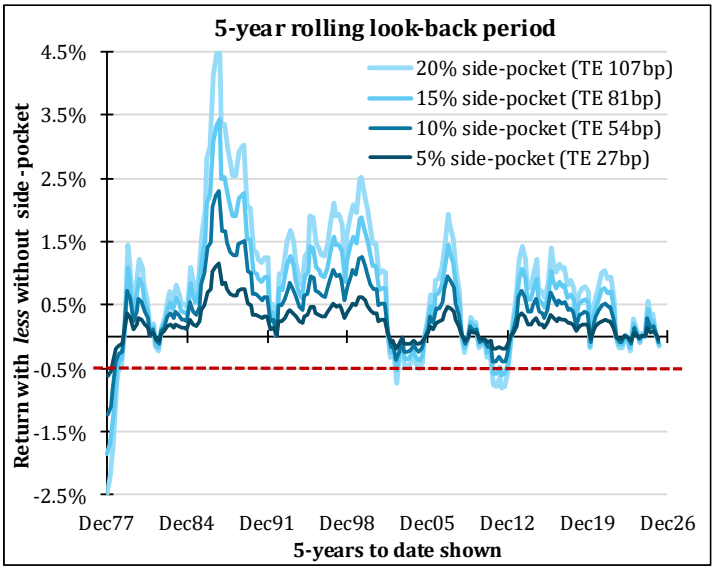
[2] *Note:* This estimates TE with respect to standard deviation of portfolio outcomes at the end of long holding periods. TE estimated over short horizon is much higher. For example, the standard deviation of quarterly return differences between the portfolio with and without the side pocket is 5.65% per quarter, which equates to 11.3% p.a. This is one small example of the significant impacts of complex yet seemingly minor design details.

Side-pocket historical simulation analysis – Results

Side-pocket weight	5%	10%	15%	20%
Tracking error, p.a.				
10-year rolling periods	0.17%	0.34%	0.52%	0.69%
5-year rolling periods	0.27%	0.54%	0.81%	1.07%
Failure %: periods <50bps				
10-year rolling periods	0.00%	0.57%	1.15%	1.15%
5-year rolling periods	0.00%	0.00%	2.87%	4.60%

Contribution to YFYS test result from side-pocket benchmarked against CPI+3%





The analysis suggests that the side-pocket itself is unlikely to lead to test failure purely due to benchmarking against CPI+3%, especially if it is restricted to a 5%-10% weight. Even at a 20% weight, we observe a 10-year TE of only 0.69% (5-year TE of 1.15%) and a 10-year test failure rate over the analysis period of just 1.15% (4.6% over 5-years).

These estimates will understate the TE implications of a side-pocket with a CPI+ benchmark for four reasons:

- ***Analysis period does not fully capture the high inflation episode of the 1970s*** – The first 10-year rolling period covers December 1972 to December 1982. As a consequence, the estimated failure rate is likely to significantly understate the risks associated with exposure to a sustained rise in inflation going forward. If the data series were commenced 5- or 10-years earlier, more rolling periods would have been captured covering the inflationary episode of the 1970s, probably resulting in an increase in the estimated failure rate. The relative performance estimates to December 1982 are more revealing. The side-pocket would have generated underperformance in excess of -0.5% over the 10-years to December 1982 at a weighting of 10% or above (and 5% over 5-years), even though the side-pocket return is assumed the same as the rest of the portfolio. This is an indication that the shift in benchmark could itself lead to test failure under an episode of poor real returns in markets, which in turn cannot be ruled out (see our April 2026 report *Market exposure as a major source of systemic risk for super – Part A: Likelihood of extended weakness*).
- ***Super funds will look forward*** – The 5-year rolling series might provide a better guide to how super funds could behave in real time. A fund suffering underperformance over 5-years will begin to worry about the implications of underperformance continuing, while being aware that it takes time to reconfigure a side-pocket containing private assets. The 5-years series may hence be a better guide to the risk that super funds are placed in a position where they may wish to limit exposure to the side-pocket in order to manage TE risk, potentially leading to reduced willingness to fund emerging and

alternative assets. The 5-year estimates suggest both higher probability and greater magnitude of potential underperformance from a CPI+ benchmarked side-pocket.

- ***Investments included in the side-pocket might have TE implications*** – The analysis assumes that the side-pocket delivers the same return as the baseline 70/30 portfolio. In practice, the side-pocket would likely include investments that have differing characteristics that may add or subtract from TE. For example, investment with J-curve effects could add further to TE, while investments with inflation hedging characteristics could help reduce TE under a CPI+ benchmark.
- ***Side-pocket will be one of multiple sources of TE*** – Our analysis focuses only on the contribution to TE arising from the side-pocket in isolation. Managing YFYS test TE requires considering all sources of potential deviation from the test benchmark, with various investments competing for a slice of the 'TE budget'. From this perspective, the TE arising from the side-pocket may have a more influential impact on the willingness to invest in emerging and alternative assets than appears on first sight. Previous research[3] by the Conexus Institute suggests that 1% is an appropriate level of TE, while we have picked up anecdotes that funds with buffer run at around 2% TE.

Overall, we feel that a 5% side-pocket with an option to extend to 10% should not give rise to significant concerns that an allocation may lead to test failure. Nevertheless, TE considerations may act as a natural constraint on how much super funds could be willing to allocate above the 5%-10% level.

4.1.3. Implementation considerations

Below are some preliminary thoughts on the key considerations in implementing option 1.1, bearing in mind that our view is that most of these matters should be addressed with the assistance of a TAG and further consultation (see Section 8):

- **Conceptual basis for the CPI+x% target** – A threshold issue is whether the +x% target is based around the concept of a 'fair' real expected return or set at a concessional rate in order to encourage investments that policymakers wish to support. The fair real return is more consistent with best financial interest duties and would avoid political criticism around the Government directing super funds how to invest or 'picking winners'.
- **Identifying a 'fair' +x target** – While a 'fair' x% target would ideally be set for individual investments, this is impractical. At the other extreme, applying a common +x% target to the entire side-pocket would introduce a range of distortions including favouring higher risk over lower risk investment. Something in-between may be more appropriate, e.g. different targets for growth and defensive assets. However, this approach may be open to manipulation if funds are able to nominate how investments are categorised. There is also the matter of whether the +x% target should (or could) be varied over time with market conditions, and what party would be responsible for adjusting the target. These are complex issues that would be suitable for consideration by a TAG.
- **Allocation cap** – The main question is whether a hard allocation cap should be set under policy or the allocation can be chosen at trustee discretion. Issues include the scope provided to funds that want to make significant investments in emerging and alternative assets under the rules versus whether accommodating trustee discretion may create opportunities for manipulation. Our sense is that TE considerations will serve as a natural constraint on allocations to any side-pocket, which should limit the risk associated with allowing trustee discretion. In any event, a 5% cap seems far too low. Intuitively we feel that 10% would be more reasonable.
- **Eligible investments** – The main question is whether funds may decide at their discretion what investments are included in the side-pocket versus declaring eligible investments under policy or regulation. The latter may be problematic to manage under policy or regulation as asset categorisation is not always clear and hence requires exercise of discretion on a case-by-case basis. For example, some climate transition and property-related investments might equally sit under unlisted infrastructure and property as well as within a side-pocket. Either approach may be open to manipulation in differing ways.

[3] *Your Future, Your Super Performance Test - Exploring the Impact on Super Fund Investment Strategies*

- **Transparency** – We suggest ensuring transparency around investments included in any emerging and alternative asset class. The case for transparency would be heightened where these investments are afforded different treatment such as inclusion based on 'nation-building' credentials and then being benchmarked in a different manner to other asset classes. This may require specifying some distinct asset sub-categories within the asset class.
- **Transition considerations** – There are many complexities around transition, with issues including whether only new or existing investments may be included in the side-pocket and how side-pocket returns are taken into account over time especially if investments are early in their return J-curve. Whether funds are permitted to transfer existing investment into the side-pocket is an important issue. While there is some merit in allowing them to do so to smooth out transition issues around performance measurement by immediately creating a history for the side-pocket (rather than allowing it to build over time), it could create opportunities for manipulation.

4.1.4. Alternative: cash+ benchmarking

Primary issues with a CPI+ benchmark include that it is an un-investible benchmark that is insensitive to the return opportunity set available in the markets, and the fact that it exposes funds to inflation as a source of TE. A way of partially addressing these issues might be to base the benchmark return on available yields on low-risk cash proxies such as bank bill yields, which will tend to reflect and vary with the baseline (real) return that is readily available in the markets. The '+' then plays the role of a required risk premium for taking additional risk. A cash+ return benchmark could be estimated on a per quarter basis and accumulated over time, similar to what would occur for a CPI+ benchmark.

Cash+ is closer to a 'fair' benchmark than CPI+. It represents an investible alternative for super funds that readjusts for the general level of baseline expected returns available in the market, and assesses fund performance based on the extent to which investments generate a return premium over cash. Nevertheless, it still leaves some of the problems with CPI+ unresolved. Realised performance would remain largely a function of available return premiums over cash on riskier assets (in particular growth assets), over which funds have no direct control. That is, realised performance would remain subject to the whims of the markets to a significant degree. Another issue would be determining the appropriate level for '+' versus cash. If the decision is made to pursue the side-pocket, we recommend that cash+ benchmarking be considered alongside CPI+ as a possible alternative.

4.1.5. Alternative: market-based benchmarks

Another possibility is assessing performance of investments within any emerging and alternative asset class against market-based benchmarks. For instance, investments could be assigned to a selection of benchmarks comprising combinations of equity and fixed income indices similar to what currently occurs within the existing Alternatives asset class, or perhaps a tailored combination of market benchmarks determined under certain criteria. One consideration is that the emerging and alternative investments that policymakers are aiming to support could be based in either Australia or overseas.

Advantages of a market-based benchmarking approach is that it would maintain consistency with the existing test structure and would assess performance against investible benchmarks that account for the general level of returns available in the market and hence have less impact on TE. The challenge would be to devise a benchmarking scheme that is both fair and difficult for funds to manipulate.

4.2. Option 1.2: Improve the existing Alternatives covered asset class

Option 1.2 could provide an alternative way of accommodating investment in certain emerging and alternative assets classes within the existing test structure. Consolidating selected investments that do not fit well within the asset classes with well-defined benchmarks into a single 'other' category seems, at first pass, much 'tidier' than having two asset class buckets that are performing a similar function. If option 1.2 is pursued, some of the issues discussed for option 1.1 would carry across including benchmarking approach, setting of any allocation cap, determining eligible investments, transparency and certain transition considerations. However, option 1.2 also gives rise to a range of other issues.

One challenge is to resolve the role of asset sub-categories and associated benchmarks. Currently the Alternatives asset class uses three benchmarks (formed from blends of growth and defensive assets) to

accommodate different asset sub-categories distinguished by level of risk, which in turn is self-assessed by funds. The issue arises of whether these three categories would adequately cover emerging and alternative asset classes, or if some other form of categorisation and associated benchmarks is required. For instance, affordable housing is not a single asset class and hence would need nuanced benchmarking, with public disclosures indicating two funds (HESTA and AustralianSuper) approaching the sector as equity investors while CBUS has expressed preparedness to be a debt provider.

Applying a CPI+ (or cash+) benchmarking approach within the Alternatives asset class might be another option. However, this would seem an uncomfortable fit with the prevailing benchmarking method. Accordingly, any CPI+ benchmark is probably better implemented through creating a separate asset class as per Option 1.1.

Another consideration is whether and how an exposure cap is imposed. Currently there is no cap on allocations to the Alternatives asset class. If a cap was to be imposed upon introduction of emerging and alternative investments into the Alternatives asset class, an issue is how this cap might interact with the existing role for the Alternatives asset class (which accommodates assets such as hedge funds, commodities and insurance-based investments). A low cap could unreasonably constrain investments across a range of areas, while a high cap might provide too much latitude and may open up avenues to manipulate the test.

4.3. Overall assessment

We are not against either option 1.1 (the 'side-pocket') or option 1.2, which we view as ways to provide a bit more flexibility to trustees to invest in emerging and alternative assets classes. However, we doubt either option would spur any significant uplift in investment due to the presence of two key constraints that would persist under option 1:

- **TE risk considerations** – These would persist within the context of 'TE budgeting 'and indeed could be sufficient under a CPI+ benchmark to significantly limit the strength of any investment response from funds.
- **Limited availability of investment opportunities offering clearly attractive returns** – We suspect that this is the main reason that there is not more investment in emerging and alternative asset classes is. We discuss this issue in the context of climate investing in our *Climate-related investment activity by superannuation funds* report of October 2025. We suspect the issues raised in that report extend to equity investment in housing as an asset class that seems priced for low returns given current valuations. Meanwhile, venture capital is inhibited by the difficulty of accessing attractive private equity investments at scale and J-curve effects. Super funds can always find money to buy assets which are considered clearly attractive (witness data centres). We are suggesting that the underlying issue is limited availability of such attractive assets, notwithstanding the large requirement for capital in certain sectors. A large opportunity to invest does not ensure high risk-adjusted returns. If the Government wants to encourage investment in certain areas, it might be better examining what could be done to improve the attractiveness of these investments from a risk-return perspective.

We are ambivalent between option 1.1 and option 1.2. Option 1.1 would afford more latitude for funds and would make investment in targeted 'nation-building' asset classes more transparent. An alternative worth considering under option 1.1 would be to apply cash+ benchmarking (see Section 4.1.4) or market-based benchmarks (see Section 4.1.5). Option 1.2 would support a tidier test in terms of a single 'other' asset bucket that is assessed in a manner consistent with the existing test, but it would require a review of the benchmarks which could require a reasonable degree of change if the three alternatives' buckets were not deemed suitable. Further, option 1.2 may restrict overall activity and hence portfolio diversification if a tight exposure cap is imposed on an asset category that embeds a wide range of investments. In any event, if a place is to be carved out for 'nation building' emerging and alternative assets classes, there would be call for a comprehensive review of the benchmarks including allowing for the possibility that the investments that policymakers are aiming to support may be based in either Australia or overseas. Such implementation matters are quite complex and technical and should be considered under a multi-phase process such as that described in Section 8.

5. Option 2: Introduce an assessment of risk-adjusted returns

While we support introducing assessment of risk-adjusted returns within any bright line test, we recommend that this is done under a multi-metric test regime (see Section 3.3.2). We are not supportive of shifting to a single reference portfolio metric, or indeed any single test metric. The unintended consequences stemming from the benchmarking approach and methods applied to 'manage' the test will only change in nature versus the existing test rather than disappear. Further, if a reference portfolio approach is to be used, the design of the reference portfolio and the choice of risk proxy are critical issues and would need deep consideration. Our approach in this section is to outline a number of options for implementing a reference portfolio test that go beyond the adoption of benchmarking against a SRP frontier using standard deviation (SD) as a risk proxy.

We make the observation that, in its proposed form, option 2 would render option 1 redundant.

5.1. Key considerations with a reference portfolio test metric

We see considerable merit in reference portfolios as a governance tool, i.e. as a mechanism for the trustee board to express risk tolerance and create a benchmark against which an investment management team is expected to add value and as one tool (of hopefully a number of tools) for internal performance evaluation. A reference portfolio approach is arguably a better fit with a total portfolio approach (TPA), which super funds are increasingly pursuing. We also see advantages in assessing total portfolio returns and bringing risk considerations into play, the lack of which are major shortcomings of the existing test.

However, we see some significant challenges with creating an effective reference portfolio test for the purposes of performance assessment of individual funds in an YFYS testing environment, especially where it was to be used as a single metric within a bright line test with existential consequences for failure. Key issues include construction of a reference portfolio and the choice of risk proxy being used in forming any reference portfolio frontier, with these design choices being fundamental to determining the benchmark return against which fund performance is assessed and hence the effectiveness of the test. Other issues include the incentive that the test creates and hence the potential for unintended consequences, and the degree to which the test can be 'managed' (i.e. manipulated) by super funds. We discuss these considerations in the sub-sections below.

5.1.1. Effectiveness at identifying value added by the super fund

The fundamental principle is that any performance test should effectively and reliably identify value added or detracted by super funds for their members with minimal probability of false signals (i.e. Type 1 and Type 2 errors). Avoiding false failures is particularly desirable in the context of the YFYS test where the consequences of failure are existential for super funds. However, it is also desirable that the test is able to distinguish poor fund performance (i.e. false positives) and hence not be too easy to pass. This can be a difficult balance to strike. We have doubts over the ability to form a reference portfolio test that is sufficiently effective and reliable to stand as a single metric, given the issues discussed below and some very challenging design choices that can have a significant impact on test results.

5.1.2. Construction of the reference portfolio

The construction of the reference portfolio is a critical consideration for both determining the benchmark return and the type of incentives created, in particular with regard to how funds may define and then manage TE. Key issues include what asset classes are included in the reference portfolio and their weights. Another key issue is treatment of currency hedging within the reference portfolio.

Assets included in the reference portfolio

We identify two broad options for the type of reference portfolio, and within that a further choice of exactly what asset classes are incorporated:

(a) *Simple reference portfolio (SRP)* – SRPs comprise a narrow set of passive investments that are readily available to members at low cost. The Treasury discussion paper refers to including equity and fixed income assets, which are used by APRA in its SRP test metric within the *Comprehensive Product*

Performance Package (CPPP). The PC review of 2018 applied a listed benchmark portfolio (BP1) incorporating other asset classes including listed proxies for infrastructure, property and private equity. A key feature of SRPs is that they assess funds against a subset of their investment opportunity set. Test results are thus subject to the relative performance of off-benchmark assets versus assets within the SRP, in particular the performance of private and other alternative assets versus public market assets. Two main consequences can arise from this design feature:

- It could give rise to a meaningful probability of a fund failing the test through underperformance of off-benchmark assets, even though holding these assets may have been in member best interests on an ex-ante basis.
- An incentive may be created for funds to converge on the reference portfolio under certain conditions, most notably where test buffer is thin and there is high concern over future test failure. We say more about this in Section 5.1.4.

(b) *Broad reference portfolio (BRP)* – This includes return indices for other asset classes that are not passively available such as private infrastructure and property. The PC also examined a BRP in its 2018 review denoted as a blended benchmark portfolio (BP2). While a BRP gets closer to capturing the opportunity set faced by super funds, it does not fully address this issue as the benchmarks are an incomplete representation of the unlisted asset classes. However, a BRP may limit the consequences outlined for a SRP related to probability of failure stemming from performance cycles in non-benchmark asset classes and incentive to converge on a narrow set of assets. A BRP approach is further discussed in Appendix A.

The two approaches differ in their application to funds with different liquidity profiles. All funds can in theory replicate a SRP, making it an active decision to construct a different portfolio. Under a BRP, funds that resolve to run a liquid strategy may not be able to replicate a BRP containing significant illiquid assets. We maintain a strong view that a fund's potential allocation to illiquid assets should be informed by its liquidity profile. For example, the operating structures of some for-profit funds requires them to emphasise listed assets. It may be unfair to benchmark these funds against a BRP containing private market assets at higher weights than are prudently achievable under their operating structure.

We can identify strengths, weaknesses and challenges under both SRP and BRP approaches and do not take a strong view either way[4]. In fact, our strongest reflection is that working through the issues of SRP versus BRP benchmarking further heightens our concerns about complexity and the impact of test design details and the associated risks of switching to an alternative single metric.

Asset weights

A number of options present themselves for weighting assets to form a reference portfolio. The central issue is the extent to which the reference portfolio weights are specified independently or attempts are made to reflect fund SAA weights. Independently specified weights would impose a benchmark that is consistently applied across the industry, but in doing so, may influence how the industry invests to the extent that funds might converge towards the reference portfolio in order to limit TE. On the other hand, reflecting fund SAA weights in any reference portfolio would reduce herding incentives with respect to the reference portfolio SAA. However, it would limit the effectiveness of the test in assessing the impact of SAA decisions and reduce the test towards an assessment of the implementation of declared SAA (as occurs under the existing test). A trade-off thus exists, to which there is no clearly superior choice.

The table over describes potential weighting schemes along the spectrum from imposing entirely independent weights under scheme #1 to fully tailored weights under scheme #4. Note that this table does not consider the asset class indices that might be considered for the reference portfolio, which were discussed in Section 5.1.2. Weighting scheme #1 involves a single reference portfolio and is included largely for completeness rather than as a likely option for use in the test. Nevertheless, it does reflect the approach where a trustee board establishes a reference portfolio as baseline for the management team (e.g. NZ Super). Weighting scheme #2 entails establishing growth and defensive reference sub-portfolios with independently specified weights that are combined to trace out a frontier. This approach equates to what appears proposed in the discussion paper and is used by APRA in its

[4] In our view, SRPs are reasonable way of assessing the value that the super industry at large has added relative to what members could have readily accessed for themselves at low cost, as per the PC's 2018 review.

CPPP SRP metric. Weighting scheme #3 tailors the weights within the growth and defensive components of the reference portfolio towards fund weights, which are then used to create a reference portfolio frontier. Weighting scheme #4 matches the reference portfolio weights to a fund's SAA, which reflects the situation under the existing test. All schemes accommodate accounting for risk differences between the fund and reference portfolios arising from variance between asset classes and investments held, but do so differently. While scheme #1 and #4 entail a single reference portfolio, a risk/return performance frontier can still be formed by combining the reference portfolio with a risk-free proxy (likely cash). Schemes #1 and #2 support using SD and G/D weight as risk proxies; while schemes #3 and #4 preclude using G/D weight as the fund and reference portfolios would have the same G/D weight by construction.

Potential weighting schemes in forming a reference portfolio

Reference portfolio weighting scheme	Reference portfolio frontier formulation	Test metric and risk proxy
1. Generic single reference portfolio with independently specified asset weights	Generic single reference portfolio is combined with risk-free proxy (e.g. cash) to trace out a 'performance frontier' to enable risk adjustment	Fund return benchmarked against return on frontier portfolio of same risk, with two feasible risk proxies: • Risk most likely based on *SD* • *G/D weight* could be applied
2. Generic growth and defensive portfolios, each with independently specified assets weights, are combined to form reference portfolios	Generic growth and defensive portfolios are combined to trace out a reference portfolio frontier	Fund return benchmarked against return on frontier portfolio of same risk, with two feasible risk proxies: • *SD* - Approach that appears to be proposed by discussion paper • *G/D weight* - Used in APRA's CPPP
3. Tailored growth and defensive portfolio, each reflecting fund's asset weights within growth and defensive components, are combined to form reference portfolios. For example: - Fund with 65/35 overseas/local equity mix within its growth component benchmarked against reference portfolio with same 65/35 mix within its growth portfolio - Currency hedging also matched	Tailored growth and defensive reference portfolios are combined to trace out reference portfolio frontier	Fund return benchmarked against return on frontier portfolio of same risk, likely using SD as a risk proxy *Note:* Use of *G/D weight* as risk proxy redundant as fund and reference portfolio G/D weights should be equivalent by construction
4. Tailored reference portfolio weights to reflect fund SAA	Tailored single reference portfolio is combined with risk-free proxy (e.g. cash) to trace out a performance frontier to enable risk adjustment	Fund return is benchmarked against return for performance frontier portfolio of same SD *Note:* Use of *G/D weight* as risk proxy redundant as fund and reference portfolio G/D weights should be equivalent by construction

We have no strong preference over the above options. However, we trust our discussion may highlight the potential importance of the construction of any reference portfolio.

5.1.3. Risk proxy

Whatever risk proxy is used should be relevant to members, i.e. capture risks that matter to members. The two main risk proxies that present themselves as practical options include SD as referred to in the Treasury discussion paper, and G/D weight as is currently used by APRA for its SRP test metric within the CPPP. We hence focus on these two risk proxies with the main theme being that each raises specific issues and neither is clearly preferred, such that they are better viewed as complementary rather than competing measures. Advantages and disadvantages of each risk proxy are outlined in the table below.

SD versus G/D weight as a risk proxy

Risk proxy	SD	G/D weight
Advantages	• Captures risk reduction benefits from diversification or strategies that reduce return volatility • Easy to estimate • Relatively familiar to professionals	• Can be revealing of underlying economic risk exposures that may result in shortfall versus objectives over the long run • Familiar, including to members
Disadvantages	• Questionable measure of risk of delivering poor member outcomes over long periods • Prone to distortion from smoothed returns arising from valuation methods in alternative assets[5] • Can obscure underlying risk such as embedded economic exposures	• Diversification benefits not captured effectively: diversifying may provide more risk reduction than evident in G/D weight change • Blunt risk proxy that may not properly distinguish differing risk levels for assets within the growth and defensive categories • Difficulty of standardising G/D weightings across assets and across funds • May be open to manipulation by funds

A key point of distinction is that SD accounts for diversification benefits and their impact on short-term return volatility while G/D weight may be more effective at capturing underlying economic risk exposures that can be a key driver of poor long-term performance. A good example is how both risk proxies treat private equity and private debt. These assets can generate equity-like returns due to presence of risk premia related to embedded economic exposures and leverage. Further, risk in private debt can manifest as occasional episodes of very poor returns that may not occur over any particular measurement period. However, these embedded risks are obscured by SD as a consequence of valuation processes. In contrast, using G/D as a risk proxy can better account for the underlying risks through treating private equity as a 100% growth asset and treating private credit as a mix of growth and defensive exposures. On the other hand, only SD and not G/D can give credit for additional diversification benefits arising from return sources associated with these assets that are not highly correlated with equities and fixed income. Both risk proxies are lacking in key dimensions.

Our main point is that no comprehensive and reliable risk measure exists. This is another reason why we favour a multi-metric approach as a way of incorporating multiple risk proxies into the assessment process. Both risk measures should be considered complementary rather than mutually exclusive.

5.1.4. Incentives created

Page 16 of the Treasury consultation paper states:

> "*By reducing reliance of asset class benchmarks, this approach could more directly address concerns about benchmark hugging and unintended investment constraints.*"

We disagree with Treasury's assessment as an unconditional statement, noting that under the reference portfolio approach the benchmark merely changes in nature. The incentives created are contingent on test design, in particular the construction of the benchmark and how circumstances evolve particularly regarding developments in markets and fund performance. TE to any test benchmark is fundamental to the incentive for benchmark hugging and constraining investment in off-benchmark assets, especially when combined with the existential nature of test failure. Funds are incentivised to manage test TE to ensure that the probability of failure is very low. Under certain situations, an incentive will arise to reduce TE, especially where limited test 'buffer' exists as a consequence of developments. The impact could be the opposite to that suggested in the above quote.

[5] Potential for smoothing is not just a function of the frequency of revaluations. Return smoothing also stems from how appraisal values are estimated whereby they may remain largely unchanged until fundamental news arrives. They also involve considerable subjectivity, which can leave valuations open to manipulation. SD may also inappropriately benefit funds where poor valuation governance practices have contributed to smoothed returns.

TE under the existing test is largely determined by deviations from the indices through investments held within each asset class[6]. Under a reference portfolio test, TE would reflect two components:

(a) *Deviations in the asset class weights versus the reference portfolio (reflecting SAA)* – TE can arise from investing in any asset classes outside of those reflected in the reference portfolio, most notably alternative assets to equities and fixed income in the context of a SRP.

(b) *Deviations from the benchmark indices within the reference portfolio asset classes* – TE can also arise from holding investments within asset class portfolios that deviate from the benchmark indices for those asset classes, e.g. differing equity and fixed income securities to those in the indices that comprise a SRP. This effect is similar to the existing test but the scope of the effects could be narrower to the extent that some asset class indices currently being used disappear from the mix.

Deviations associated with (a) are likely to swamp those from (b) in determining TE under a reference portfolio metric. Funds will thus be incentivised to pay closer attention to SAA weights relative to the reference portfolio than the make-up of their individual asset class portfolios, although the latter would still remain relevant. This would be a significant change in how TE and performance versus the test is currently determined under the existing test.

The effect should be that the incentive shifts from managing TE relative to the asset class indices to managing both SAA deviations from the reference portfolio weights and security selection within the reference portfolio assets, but with much greater emphasis on SAA. Funds may focus on *how much* deviation they are willing to take from the reference portfolio SAA in setting their SAA. The following aspects would be relevant under the reference portfolio frontier approach:

(i) *Impact on expected return* – The expected returns offered by off-benchmark asset classes will be a key consideration. However, from an ex-ante perspective, returns are the most uncertain component of test outcomes as they are never guaranteed. As a consequence, the expectation of higher returns is in most cases unlikely to be assigned sufficient confidence to justify large deviations from the reference portfolio[7].

(ii) *Impact on expected risk* – Extent to which diversifying into off-benchmark assets reduces portfolio risk is the next consideration. Under SD as a risk proxy, the impact of off-benchmark assets on risk will to some extent depend on how SD is measured especially how private assets are treated, i.e. how much risk reduction they yield due to return smoothing. Under G/D weight as a risk proxy, the impact will depend on the G/D categorisation of the assets involved. The impact may also depend on the slope of the frontier, with changes in SD or G/D weight making more of a difference as the slope steepens[8]. As a general rule, the impact of diversifying into off-benchmark investments on risk measures is more predictable[9] than the potential impact on future returns.

(iii) *Test failure threshold* – The test failure threshold arising from off-benchmark asset classes also matters. The failure threshold can interact with existing buffer to influence how much TE that a super fund may be willing to take to limit the probability of failure.

(iv) *Extent of 'test buffer'* – The disincentive to deviate from the reference portfolio weights would be greater as the distance narrows between fund returns and the failure threshold such that a fund has limited buffer hence raising the probability of test failure. Test buffer will evolve dynamically, potentially leading to changing behaviour over time. Here funds may consider what returns will remain or drop out of the 10-year history looking forward. This may lead to emphasis being placed on relative returns versus the reference over latter years of the look-back period on the basis that this establishes a foundation for future test results as earlier returns run off. A forward-looking

[6] TE also stems from deviations from the stated SAA through impacts from market fluctuation and rebalancing activity, as well as deviation in currency hedging relative to that embedded in the YFYS benchmarks.
[7] This may depend on the investment in question. For instance, super funds have been willing to increase credit exposure within fixed income portfolios on the basis of high likelihood of outperforming the benchmark; whereas much less confidence can be attached to equity investments outperforming.
[8] Perversely if growth (e.g. equities) under-perform defensive assets (e.g. fixed income) so the reference portfolio frontier slopes downwards, diversification that lowers SD or G/D weight will then raise the benchmark return.
[9] Under SD as a risk proxy, the volatility of individual assets tends to be relatively predictable, although correlations are less stable and may be regime dependent. Under G/D as a risk proxy, the impact on the risk measure should be largely known.

perspective would require allowing sufficient time to adjust the portfolio, especially with regard to long lead times involved in initially investing and then adjusting exposures in private market assets.

To provide some sense for the trade-offs referred to above, Appendix B presents some modelling of the impact on portfolio risk and return from holding mid-risk assets such as infrastructure and property where SD is used as the risk measure. Section 5.3 also reflects on existing buffer across the industry under a version of the proposed SRP metric with SD and G/D as risk proxies.

If super funds decide to manage the risk of failure as the first priority, they could become more wary about investing in asset classes not included in the reference portfolio given they would give rise to considerable TE. This could induce convergence towards the asset classes within the reference portfolio and therefore restrict or even reduce allocations to emerging and alternative asset classes. If this were to occur it would be contrary to the review's aim to limit benchmark hugging and reduce constraints around investment in certain assets. There may also be unintended systemic impacts arising from policy influencing how super funds set their SAA and inducing increased commonality in asset exposures across the industry.

Whether funds would move in the direction of limiting TE risk relative to any reference portfolio is a moot point. It may depend in part on the design of the reference portfolio test regarding aspects such as composition, failure threshold and risk proxy, as well as prevailing buffers that could shift over time. Nevertheless, the potential exists for such perverse outcomes under certain conditions, which should be given consideration in deciding whether to proceed with a single reference portfolio test.

Comment on currency hedging

The implications for currency hedging practices deserve special consideration. Currency hedging decisions can be quite influential for fund performance, and only likely to increase in importance as the industry grows in size and exposure increases to international assets. Any deviations in the currency hedging position adopted by a fund versus that implicitly embedded within the reference portfolio will give rise to TE. To gauge the potential magnitude, currency hedges that deviate from that within the reference by 10% of the total portfolio would impact on yearly performance versus the SRP by 1.0% for a 10% currency movement, although this would be diluted to approximately 0.1% for a 10% currency move when spread over 10 years. While this may seem modest, recall that the A$ has historically undergone large swings over extended periods, having ranged between roughly 50 cents and $1.10 versus the US$ over the post-float period. This is a -55% to +220% range, depending on the numeraire.

Imposing a particular hedge ratio on the reference portfolio could inadvertently drive super fund hedging practices to the extent that they aim to set hedging positions to mimic those within the reference portfolio to limit TE risk. Funds may be reluctant to assign much of their test TE budget to currency hedging to the extent that currency movements are hard to predict. Meanwhile, funds may have good reasons for adopting a different currency hedge ratio to that with the reference portfolio. One reason would be to constrain hedge ratios in order to manage liquidity risk. Another is, as large funds grow further in size and increase international exposure, they may be constrained in hedging if counterparties are unwilling to increase their foreign exchange books and counterparty exposure limits. These considerations suggest that substantial thought and care should be put into the hedging assumption embedded in any reference portfolio.

5.1.5. Scope for funds to manage the test

Fund attempts to 'manage' a reference portfolio test would involve two elements. The first is managing TE to limit probability of failure, which was discussed above. The second could be attempting to manipulate the risk measure in order to be assessed against a reference portfolio that delivers a lower benchmark return. Assuming an upward sloping frontier, this would involve aiming to have the fund portfolio categorised as lower risk under the particular risk proxy being used.

- *SD-based test metric* – Funds would be incentivised to reduce measured SD if possible. This might involve attempts to use discretion over the valuation process where available, particularly with regard to private assets, to ensure that valuation changes occur gradually if possible. It may also involve favouring assets that may carry underlying risk that do not show up in return volatility, e.g. private assets, with private credit being a notable case as discussed above. We make the observation that a SD

risk proxy may inadvertently and inappropriately benefit those funds where poor valuation governance practices are contributing to smoothed returns.

- *G/D-based metric* – Funds would be encouraged to use any discretion to categorise growth assets as defensive assets. They might also be incentivised to seek out assets with higher expected return within the growth and defensive components, at least where there is a high enough probability that extra return would be delivered. Such behaviour is evident under the existing test through the use of credit within fixed income asset classes. Similar behaviour could extend across the palette of investments within the growth and defensive components, e.g. favouring higher risk growth and defensive assets that offer relatively high expected returns versus similar types of assets with a reasonable probability of delivering a premium.

5.2. Further implementation considerations

The Treasury consultation paper asks for inputs on the following implementation issues:

- asset classes to be included in the SRP and related market indices
- whether SD is an appropriate risk measure, and if there are any alternatives
- implementation and transition considerations

Reference portfolio construction was addressed in Section 5.1.2 and risk measures in Section 5.1.3. Below we discuss selected other implementation issues. Our intent is not to make recommendations, but rather to highlight the implementation challenges that would be involved.

- **Indices used** – Indices will need to be chosen to represent each asset class. There will be multiple index options available, which we will not address in any detail at this stage. One choice is whether indices versus passive funds or ETFs are used for estimating returns. Investible alternatives like passive funds or ETFs offer the advantage that they embed any implementation costs but may not provide adequate history covering all asset classes. *Note:* We have received numerous complaints from super funds about the cost of index subscriptions, especially for those indices which are part of the performance test. We have previously recommended that if these claims have any foundation, then the concerns should be passed on to the ACCC.

- **Rebalancing assumptions** – A decision needs to be made on whether the reference portfolio is assumed to be rebalanced yearly, quarterly, monthly or even daily for the purpose of estimating the return and potentially SD over the course of the assessment period. Another issue is whether to allow for notional rebalancing costs.

- **Failure threshold** – Failure threshold will determine which and how many funds would fail. It will also influence the extent to which funds are incentivised to manage TE and hence the extent to which the test may induce benchmark-hugging and constraints off-benchmark investment including in emerging and alternative assets. The 50bps threshold under the existing test may not be appropriate under a reference portfolio test. A larger failure threshold may be appropriate to recognise greater uncertainty around a reference portfolio relative to the existing test due to the test metric being subject to volatility in asset classes as well as individual investments within asset classes and a larger set of design choices and assumptions. However, it may be a challenge for policymakers to introduce a larger threshold than 0.50% to the extent that it could be viewed as weakening the test. The 'right' failure threshold will be tricky to calibrate and should ideally involve analysis of confidence intervals.

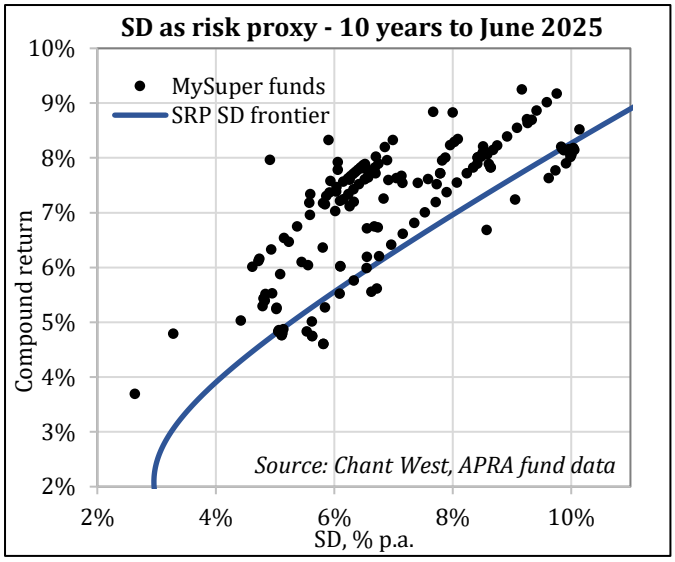
- **Estimation of SD** – It is assumed that SD would be estimated with respect to the assessment period (i.e. 10 years). A key estimation choice is the data interval, i.e. whether monthly, quarterly or yearly return observations are used. There is a trade-off between having more observations and diluting the distortion from assets with smoothed returns, e.g. private assets. Another option might be to apply a model to account for the impact of any serial correlation in the data.

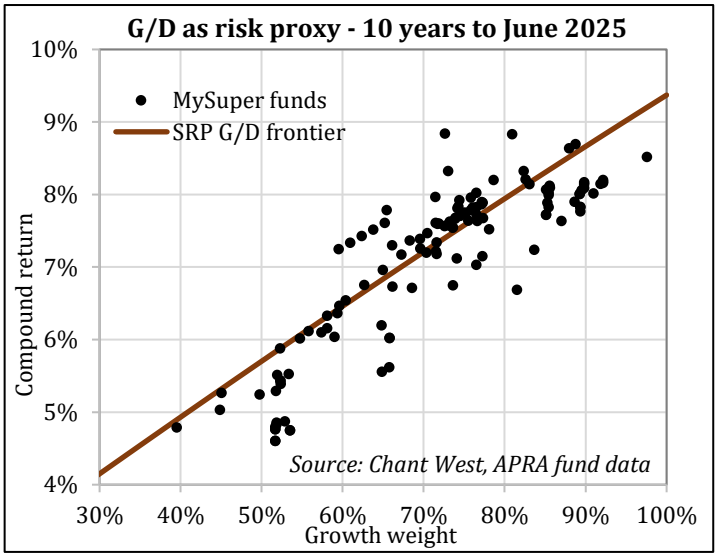
- **Transition to a reference portfolio test** – A reference portfolio test might be introduced on an immediate, deferred or transition period basis, with the latter potentially involving running the reference portfolio and existing test in parallel for a period. A key issue is the extent to which any new test is applied retrospectively or introduced progressively over time.

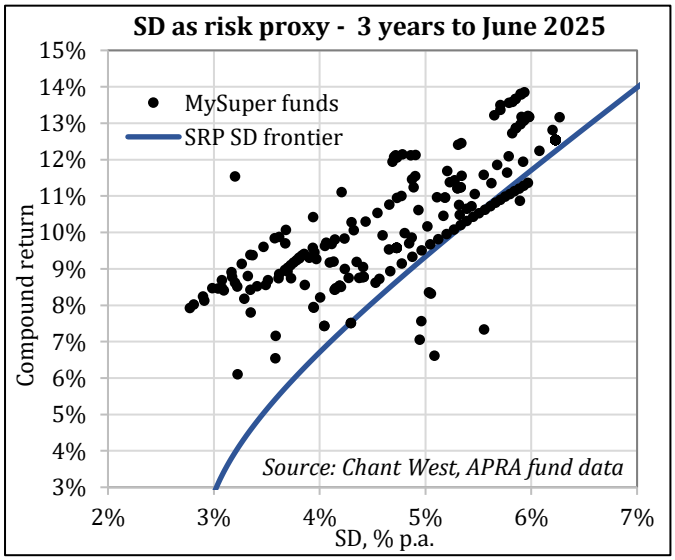
5.3. Industry performance record and buffer under two reference portfolio tests

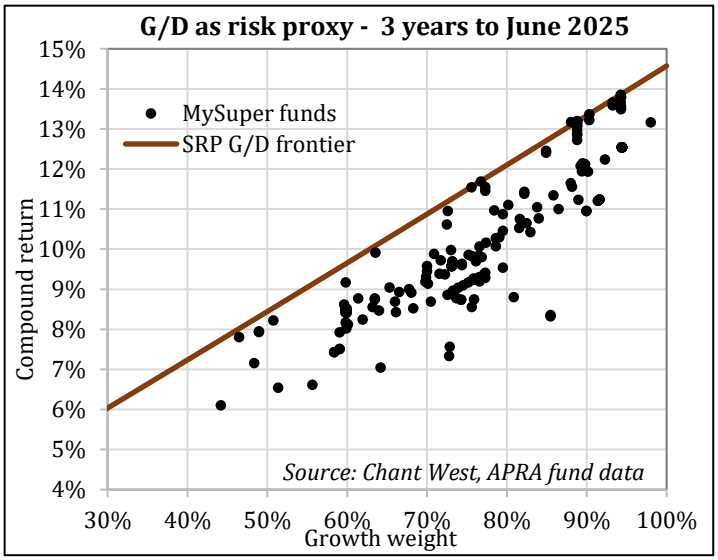
This section presents analysis of performance using MySuper fund data through until June 2025 as undertaken and provided to the Conexus Institute by Chant West. Our aim is to illustrate the potential for significant impacts to arise from test design choices. We specifically focus on the impact of risk proxy choice through presenting and discussing results based on SD and G/D weight. The panel of charts below plots the performance of MySuper funds against SRP frontiers formed using both risk proxies over 10-year and 3-year look-back periods. The 3-year period is shown for an indication of how funds may respond under either metric going forward, noting that the 3-years' period will remain in the history under FY2032. The results are presented over the page as charts (with estimation notes shown underneath). The analysis presents a stark example of how design choices can have a large impact.

MySuper fund returns vs. SRP frontier using SD and G/D over 10- and 3-year look-back periods









Estimation notes

- Growth and defensive portfolios formed and combined to form SRP and G/D frontiers under two risk measures
- *Weights* (same as used in APRA CPPP):
  Defensive portfolio: 20% cash, 40% Australian bonds, 20% international bonds
  Growth portfolio: 50% Australian equities; 25% international equities, unhedged; 25% international equities, unhedged
- *Returns:* Compound over the look-back period
- *Risk proxies*:
  - SD of quarterly returns over the look-back period
  - Average G/D weight over the look-back period
- Indices used to construct reference portfolio frontiers:
  Cash: Bloomberg AusBond Bank Bill
  Australian bonds: Bloomberg AusBond Government 0+ Yr
  International bonds: Bloomberg Global Treasury, $A Hedged
  Australian equities: S&P/ASX 300
  International equities, unhedged: MSCI ACWI ex Australia with Special Tax, Unhedged
  International equities, hedged: MSCI ACWI ex Australia with Special Tax, Hedged
- ***Source:*** Charts sourced from Chant West, which used APRA-reported data on super fund returns and growth weights

The main takeaway is how much the results vary both between the two risk proxies and over the various look-back periods. While most funds outperform the SD-based SRP benchmark, a majority underperform the G/D-based SRP benchmark. In the latter case, the underperformance is largely sourced from the three years to June 2025 (based on an analysis of APRA CPPP data, which uses the G/D method). We surmise that this spate of underperformance reflects a combination of poor performance from certain private assets and security selection (i.e. active management) relative to public market indices. A key feature of the G/D-based SRP frontier is that it does not capture the additional reduction in measured SD-based risk arising from a combination of diversification and return smoothing stemming from private assets, which appears to largely account for the difference between the two metrics. It is clear that the chosen risk proxy would be pivotal to the test results.

This motivates the question of which risk proxy is more suitable as a measure of the risks that matter most to members. As discussed previously, G/D is a coarse measure but might better reveal underlying exposure to economic risk. Meanwhile, SD can better recognise diversification benefits but may hide underlying risk in part due to return smoothing. The real risk exposure probably resides somewhere between the two measures. It is plausible that the G/D-based metric is too harsh and could indicate failure for funds that are taking sensible steps to diversify their portfolios, while the SD-based measure is too lax on funds and may not be effective at identifying genuine underperformers.

The above analysis further supports our conclusion that a multi-metric test incorporating various proxies for risk would be more appropriate.

Implications

The analysis above highlights the potential for design choices to have a significant impact under a reference portfolio test. Two potential implications arise, which are explored below.

Result reliability and failure threshold

It seems evident from our analysis that a reference portfolio test will be more sensitive to design choices than the existing test. The contrast between the SD-based and G/D-based SRP results highlight the sensitivity to design choices around the selection of risk measure used to identify a fund's matching reference portfolio. Other design choices also have potential to impact the test results. The construction of the reference portfolio itself as discussed in Section 5.1.2 will be important, including the choice of assets included (especially the choice of SRP versus BRP), the weighting scheme used to construct the reference portfolio performance frontier and treatment of currency hedging. Other design choices such as assumptions around indices, fees and rebalancing will also make a difference at the margin.

Such sensitivity to design choices cast doubt over the suitability of a single reference portfolio metric for use as a bright line test with existential consequences. As suggested above, the test failure threshold may need to be varied from the 50bps as applied under the existing test to account for the additional variation arising from 'model risk' under a reference portfolio test where model design choices are somewhat arbitrary and debatable.

Test buffer and incentives

Both design choices and performance patterns over time will impact on the test buffer that funds will be facing into under a reference portfolio test, which will in turn influence the incentive to converge towards the reference portfolio to limit TE and hence the probability of test failure going forward. Adoption of a SD-based SRP frontier could initially provide most of the industry with considerable buffer, thus leading to weak incentives to converge towards the SRP at the current time. The existence of a meaningful buffer carries over into the 3-year look-back period for most funds, suggesting that absence of an immediate incentive to converge towards such a benchmark to limit TE. However, to the

extent that an SRP test is easy to pass due to design features that favour funds, it may also be ineffective at correctly identifying poor performers[10].

Alternatively, adoption of a G/D-based approach would suggest that the industry could enter into a reference portfolio test regime with considerable concerns over available test buffer. The very poor relative performance over the three years to June 2025 is particularly noteworthy as it would remain within the 10-year history for some years while better performance numbers drop out. Further, the underperformance of active management appears to have continued beyond June 2025 while listed markets (led by international equities) look to have continued to perform well overall, raising the prospect of the industry having four years of underperformance under its belt. This could provide a much stronger incentive to converge towards an SRP. The question would also arise over whether G/D may be too harsh on funds by not giving enough credit for diversification benefits.

Importantly, this is the situation as it currently stands based on the two methods examined. The results might not only shift depending on design choices but also with the evolution of fund performance over time especially the relative returns of assets held that are outside of the reference portfolio benchmark. There is no guarantee that the significant buffer evident under the SD-based SRP test or the thin buffer under the G/D-based test will continue into the future, depending on relative performance versus equities and fixed income of other assets held in fund portfolios. Further, the relation between the reference portfolio return and risk proxies inverts under a (unlikely but plausible) scenario where growth assets underperform defensive assets, which has potential to up-end the normal relationship between reduced risk and the benchmark return hurdle. Establishing a sustainable test requires considering the implications of any particular test design over the passage of time and different market scenarios, and not just how the test result looks under current circumstances.

5.4. Assessment

Introducing a single reference portfolio metric under a bright line test seems challenging and risky. Design choices, especially around construction of the reference portfolio and the risk metric, will be critical to the effectiveness of the test and the incentives it creates. Further, a reference portfolio test would establish deviations from the reference portfolio's SAA as *the* major source of test TE in a context where failure has existential consequences. Under certain circumstances this could lead to herding towards the reference portfolio asset classes, which could potentially reduce availability of funds for emerging and alternative asset classes to the contrary of one of the aims of the review. It could also lead to widespread test failures across the funds if the industry at large suffers underperformance from assets such as private markets if they are not included in any reference portfolio. Depending on the circumstances, the unintended consequences could be worse than those occurring under the existing test! In any event, the only certainty is that any unintended consequences would be of a different nature.

The range of important design choices are worth emphasising. The table below summarises the critical choices and selected options, focusing on options that may be more able to be readily implemented. The overarching point is that implementation of a reference portfolio test requires addressing a range of complex and technical issues. As a consequence, if a reference portfolio test is pursued either in isolation or as part of a multi-metric test, we strongly recommend design proceeds with input from a TAG and then further consultation as discussed in Section 8.

Testing against any single metric is deeply flawed. Changing to an alternative single metric would introduce its own unintended consequences and set of risks. While our preference for a bright line test would be a multi-metric test (see Section 3), if there is to be a single test metric, we would rather the decision be made to retain the existing test rather than taking on the design risks associated with a shift to a single reference portfolio test.

[10] The possibility that SD inappropriately benefits super funds that have been lagging with regards to uplifting their unlisted asset valuation processes is relevant in this regard. APRA's governance review (December 2024) found that 12 of the 23 in-scope RSE licensees reviewed require material improvements in either or both their valuation governance or liquidity risk management frameworks to meet the requirements of SPS 530.

Critical design choices for a reference portfolio test

Design choice	Selected options
Reference portfolio type and asset classes included	a) SRP – Equities and bonds only b) SRP – All passively-available listed assets c) BRP – All assets with acceptable asset indices
Reference portfolio weighting scheme and formation of reference portfolio performance frontier	a) Generic single reference portfolio with independently specified asset weights is combined with risk-free asset to form a frontier b) Generic growth and defensive portfolios, each with independently specified asset weights, are combined to form a frontier c) Tailored growth and defensive portfolio, each reflecting fund's asset weights within growth and defensive components, are combined to form a frontier d) Tailored reference portfolio weights to reflect fund SAA is combined with risk-free asset to form a frontier
Risk proxy	a) Standard deviation (SD) b) Growth/defensive weight (G/D weight) c) Another measure of observed risk such as factor exposures, equivalent equity exposure[11] or drawdowns. d) No explicit proxy (risk captured via tailored asset weights)
Failure threshold	Should be contingent on test design, which will require analysis

6. Option 3: Introduce a routine review of the benchmarks

We support option 3 proposing introduction of a routine review of the test benchmarks. The benefit would be greatest if the review of indices was conducted under a formalised process by a body that is structurally independent from the political process and the super industry. Formal index reviews could address any evident shortcomings and help to ensure that the design of the test remains suitable as industry practice evolves. It can consider whether more appropriate indices may be available that (say) better capture the effects of rebalancing, fees and taxes; or impose lower costs on funds and the overall system. Finally, formalising the review process and assigning responsibility to an independent body would help extract the test from what has proven a charged political environment.

Candidate parties to review benchmarks could be Treasury, APRA, or an independent body, noting that Treasury and APRA could utilise a TAG. Each model has strengths and weaknesses which we do not explore. In any event, our sense is that implementation of the index review process should be left up to the body that is responsible for undertaking the review, rather than undertaken as a policy process. However, policymakers may consider whether they may want to impose some parameters or directives around the review, e.g. how to respond to requests to broaden the indices to cover ESG, new asset classes, etc; extent to which the benchmark indices should track industry practice; and so on.

What sort of review process is needed will depend on how the test moves forward, i.e. existing test is retained, a shift is made to a reference portfolio or multi-metric test, etc. In any event, the index review process could be addressed as part of the multi-phase process we suggest in Section 8.

An additional reflection is that making changes to the index listing carries implementation issues. Judgement may be required rather than a standard rule. For instance, whether an index change is applied on a forward-looking or retrospective basis may depend on the rationale. For instance, a small adjustment to particular details of the test may support applying the change on a forward-looking basis while leaving the history unchanged, whereas a change such as introducing ESG indices might need to be accompanied by a retrospective application. Again complexity arises in nearly every consideration.

[11] Used by the Future Fund.

7. Option 4: Expanding the performance test

Performance testing has delivered benefits, making it relevant to consider whether it can be reasonably applied to other sectors beyond MySuper and trustee-directed products (TDPs). We consider four candidate sectors for expansion: externally directed multi-asset products (EDPs), separately managed accounts (SMAs), single sector products and account-based pensions (ABPs). Our approach entailed initially exploring whether the pre-conditions that led to the PC's recommendation to establish a bright line performance test (discussed in Section 3.2) apply to these areas. We conclude that they do to a meaningful degree. After exploring each area through this lens, a range of additional issues came to light that generally reduce our enthusiasm for extending performance testing to these four areas. These include: (1) whether alternative avenues exist for uplifting governance to improve member outcomes; (2) whether members who invest through these channels are exercising choice or are disengaged and placing trust in their fund; (3) presence of 'flex' in provider operating models that could be used to circumvent performance testing; and (4) other challenges. These additional considerations led us to conclude that there would be minimal benefit but considerable cost in extending the test to many of these areas. Our broad conclusions are:

- While a case can be made for applying performance testing to *EDPs*, we do not feel that the case is sufficiently compelling and hence remain equivocal about extending the test into this area.
- We see a case for limited testing of *ABPs* where the super fund is running the same strategy as an existing tested accumulation option and it is not part of a broader, integrated retirement solution.
- We do not support expansion of the test to *SMAs* or *single sector options*.

7.1. Specifying and applying the criteria for expansion of the test

The breakout box below lists the criteria we consider in evaluating whether an extension of a bright line performance test to a candidate area is likely to be beneficial. Criteria (i) to (iv) reflect the initial pre-conditions for introducing a bright line test stemming from the PC's 2018 review. Criteria (v) to (viii) are the additional considerations identified after exploring each candidate expansion area, and reflect on the potential effectiveness of introducing the test to the candidate area. We apply and summarise these criteria in the table over followed by discussion of each candidate area in subsequent sub-sections.

Criteria for evaluating the case for introducing a bright line performance test

(A) Initial pre-conditions

(i) *Performance dispersion* – identifying and addressing the poorly performing funds

(ii) *Fee dispersion* – weeding out over-charging of consumers

(iii) *Comparative data* – establishing a base of comparative product data where none exists

(iv) *Regulatory capabilities* – introducing assessment where no regulatory capability exists for assessment

(B) Additional considerations

(v) *Governance uplift* – A performance test may not be required where alternative avenues exist for uplifting governance to improve member protections. In particular, an uplift in some candidate expansion areas is likely following Treasury's *Enhancing member protections in the superannuation system* consultation.

(vi) *Nature of users* – Performance testing is most relevant for consumer protection where members are relatively disengaged and placing trust in trustees as a fiduciary, rather than operating towards the active end of the choice spectrum or under professional financial advice. For example, SMA's, single sector options and (to a varying degrees) EDPs reflect a greater degree of choice and/or adviser guidance. Measures that focus on governance and transparency may provide more benefits than performance testing in these instances.

(vii) *Operating model flex* – Performance testing may prove of limited effectiveness where providers or users can adjust access to the product through an alternative implementation channel that is not performance tested. This consideration applies most strongly to the SMA sector.

(viii) *Other challenges* – Includes relevant considerations not covered under other criteria, including the likely effectiveness of testing to identify poor performance given the nature of the area and the products offered.

Applying the criteria to candidate areas for expansion of the performance test

Candidate area	Externally-directed multi-asset products (EDPs)	Separately managed accounts (SMAs)	Single sector products	Account-based pensions (ABPs)
(A) Pre-conditions				
(i) Performance dispersion	Likely sizable in some product areas, particularly multi-asset products with total return focus	Performance visibility across products currently very limited, but highly likely to exhibit reasonable dispersion	Performance likely to be highly dis-perse, however there is often visibility around relative performance given wide availability of asset class indices	Appears to be more dispersion than in accumulation, only some of which seems due to differing investment strategies being applied for the retirement phase
(ii) Fee dispersion	Dispersion exists, but element of demand-based fee setting is suspected where 'in-demand' fund managers charge higher fees	There are many potential fees involved, which can be difficult to understand and limit overall visibility	Visibility around fees across products currently limited but likely to be highly disperse as likely to be a degree of 'in-demand' pricing power	Appears to be more dispersion than accumulation products, which could reflect either higher costs and/or lack of scrutiny of retirement options
(iii) Comparative data	No broadly available comparison of EDP offerings across all platforms; although data and comparisons may be made available via research houses	Largely non-existent; although data and comparisons may be made by research houses, platform trustees as part of their governance duties, and the financial advice industry	Performance data across products can generally be accessed on platforms, while financial advisers can access research houses	Performance data not easily accessible to members; although data and comparisons may be available via research houses while APRA CPPP covers ABPs offered as TDPs
(iv) Regulatory capabilities	Neither APRA nor ASIC appear positioned to actively monitor the performance of EDPs, i.e. not viewed as part of remit, or armed with appropriate data and resources	No formal regulatory assessment, although ASIC is currently examining the sector	No formal regulatory assessment	APRA monitors ABPs offered as TDPs through its CPPP framework

(Table continued over)

Candidate area	Externally-directed multi-asset products (EDPs)	Separately managed accounts (SMAs)	Single sector products	Account-based pensions (ABPs)
(B) Additional considerations				
(v) Governance uplift	Effective governance uplift may be delivered through measures being considered in Treasury's *Enhancing member protections in the superannuation system* consultation, namely governance requirements for platform super funds	Effective governance uplift may be delivered through measures being considered in Treasury's *Enhancing member protections in the superannuation system* consultation, namely governance requirements for platform super funds	Effective governance uplift may be delivered through measures being considered in Treasury's *Enhancing member protections in the superannuation system* consultation, namely governance requirements for platform super funds	A range of mechanisms already exist for ensuring good governance over ABPs, including regulatory oversight and trustee obligations
(vi) Nature of users	Users of EDPs operate towards the active end of the choice spectrum, and potentially under financial advice suggesting they should take greater responsibility for their investment decisions. Some challenges relate to users relying on financial adviser agency (although protected by duties) and capabilities (less protected)	Users of SMAs have made an explicit choice and are often operating under financial advice, suggesting they should take greater responsibility for their investment decisions. Some challenges relate to users relying on financial adviser agency (although protected by duties) and capabilities (less protected). SMA performance can be difficult to assess	Users of single sector options operate towards the active end of the choice spectrum, and potentially under financial advice, suggesting they should take greater responsibility for their investment decisions. Some challenges relate to users relying on financial adviser agency (although protected by duties) and capabilities (less protected)	Invested members are often relying on the super fund trustee, which buffers the case for performance testing
(vii) Operating model flex	Nothing to note here.	May push industry towards using combinations of single-sector SMAs or MDAs, which are difficult to test as more personalised; or investing via SMSFs through MISs or IDPS, which circumvent performance testing	Members may be moved by advisers to alternative implementation channels such as investing via SMSFs through MIS or IDPS, which circumvent performance testing	Super fund trustees cannot circumvent the application of the test if imposed
(viii) Other challenges	• SAA test does not work well for products based around particular objectives, e.g. real returns • Attention may lead to trustees removing products, which disrupts consumer portfolios and can impose transaction or tax related costs • Test requirement of 5-years track record limits ability to test all EDP products and identify frauds early. Testing would not have caught Shield and First Guardian	• Performance testing may need to be undertaken on the basis of an average of client account outcomes • Attention may lead to trustees removing products, which disrupts consumer portfolios or imposes costs from transaction or tax • Test requirement of 5-years track record limits effectiveness in a fast-growing sector • Many SMAs are single sector offerings, so may not be tested	• Testing would be applied to component parts rather than the overall portfolio ("the ingredients rather than the cake") • Increases difficulty of offering style-based products (e.g. low volatility equities)	• Retirement products are one component part of fund retirement solutions and retirement income strategies, which should be the focus of assessment - Failure of an ABP need not imply the overall solution is poor, and will be difficult to isolate out without broader consequences • Retirement products may be managed towards retiree needs (e.g. smoothing return path, inflation, tax), and may not be treated appropriately by YFYS test - This issue might be addressed under any test re-design

7.2. Externally directed multi-asset products (EDPs)

There appears to be a prima facie case for extending the test to EDPs based on the four initial pre-conditions (refer to Part A of prior table). However, while there are potential benefits we can also see many challenges in extending the performance test to EDPs. We view the following considerations as important:

- While formal performance testing may improve accountability in this area, this already exists to a significant extent through the layers of research ratings and product research by platform trustees, licensees and financial advisers. These areas of scrutiny provide mechanisms to identify and weed out poorly performing products.
- Other potential changes in governance requirements (for platform trustees) being considered in Treasury's *Enhancing member protections in the superannuation system* consultation may limit the call for performance testing of EDPs.
- EDPs entail a move along the choice spectrum where we believe it is reasonable that consumers should take more responsibility for their investment decisions.
- Where members are relying on financial advisers, challenges can exist related to agency, i.e. reliance on advisers operating in member best interest (although protected by duties), and their capabilities to assess a product and its suitability (less protected and not guaranteed).
- The test could treat some product types unfairly relative to their mandate or promise to members, especially under the current design. For example, multi-asset portfolios managed to a real return objective with a risk management overlay are a very poor fit with the existing test, and also may not be fairly addressed by a test focused on risk-adjusted returns. (Note: This depends on test design, suggesting that the case for expanding to EDPs may interact with decisions around adjusting the structure of the test.)
- Flux in available choice options arising from retiring of existing products and introduction of new products would be problematic for creating a comprehensive testing regime for EDPs. Further, testing may lead to trustees removing some products, which could disrupt consumer portfolios or impose costs from transaction or tax. Meanwhile, testing of new products will be hampered by the difficulty of accessing a 5-year performance history.
- Inclusion of EDPs in a performance testing framework would not prevent fraud. For example, testing of EDPs may not capture early-stage products that have been formed to enable fraud due to limited history (as occurred in the case of Shield or First Guardian). However, performance testing may still act as a deterrent.

Our position: *We express caution with respect to extending the performance test to EDPs. While we see some case for doing so, we do not feel that it is sufficiently strong to overtly advocate for the extension of performance testing into the EDP area.*

7.3. Separately managed accounts (SMAs)

Fast growth in the SMA sector underlines the importance of considering the need for performance testing as a consumer protection mechanism in this area. The four pre-conditions suggest a strong case for assessing SMAs (refer to Part A of prior table), with significant lack of visibility around performance and fees and an absence of close regulatory scrutiny. Further, this is a somewhat opaque part of the industry where the sector has not fully embraced setting their own standards, and many consumers will have low capacity to identify problems or compare the offering of different providers. However, the challenges of establishing a formal testing regime for SMAs are considerable. We view the following considerations as important:

- Assessment would supply a big improvement in transparency, and likely lead to institutionalisation of governance and processes – although we are already seeing signs of institutionalisation motivated by industry consolidation.

- Some level of governance and accountability should already exist through the oversight of SMAs by platform trustees. This may be accelerated through the governance requirements being considered in Treasury's *Enhancing member protections in the superannuation system* consultation especially to the extent obligations around the suitability of products placed on platforms are heightened.
- We consider it reasonable that as consumers move further along the choice spectrum they need to take more responsibility for their investment decisions. Further, many investments in SMAs occur under a financial advice arrangement, where some accountability exists via the layers of research ratings and product research by licensees and financial advisers. However, it can be challenging for users and indeed financial advisers to assess performance in the SMA area, which operate under a range of nuanced investment processes.
- Where members are relying on financial advisers, challenges can exist related to agency, i.e. reliance on advisers operating in member best interest (although protected by duties), and their capabilities to assess a product and its suitability (less protected and not guaranteed).
- We are not sure of the additional degree of consumer protections that performance testing of SMAs would provide. Testing SMA's may accelerate a switch to other channels that circumvent the test such as greater use of single sector SMAs, the use of managed discretionary accounts (MDAs) or SMSFs invested through MIS or an investor directed portfolio service (IDPS).
- In our view it is difficult for extreme frauds to occur in SMAs due to beneficial ownership arrangements, independent custodian arrangements, model portfolio transparency and potentially platform trustee oversight. Nevertheless, there remain elements of risk. Similarly to EDPs, inclusion of SMAs in a performance testing framework may deter, but not prevent, frauds.

Our position: *We do not support extension of the test to SMAs.*

7.4. Single sector products

While the initial pre-conditions may seem to be in place for extending the test to single sector products (refer to Part A of prior table), we see few benefits in extending performance testing to this area. The central issue is that single sector products are essentially choice options involving an explicit decision to allocate capital by the member and thus sit much further along the choice spectrum than EDPs and SMAs. We have two main concerns with testing single sector products:

- **Intentional choice** – The option may be intentionally selected based on personal preferences. For instance, investment could be made based on a desire to invest sustainably, a positive view on a particular asset class or wanting to access a particular investment style (e.g. value). A member who is prepared to choose a single asset class should be under an onus to undertake their own research (e.g. review the PDS or realised performance), or perhaps rely on research and direction provided through a financial adviser, in judging the suitability of the product for their needs.
- **Only one component** – Single sector products are often likely to be operating as one component of a broader portfolio. Performance testing of single sector products that could be operating as one portfolio component is tenuous, as it is the aggregate-level outcome of diversified portfolio created through blending various options together that really matters.

Another consideration is that a degree of governance and accountability should already exist through the oversight of single sector options by super fund trustees, or by financial advisers where the member is operating under financial advice. However, where members are relying on financial advisers, challenges can exist related to agency, i.e. reliance on advisers operating in client (member) best interest (although protected by duties), and their capabilities to assess a product and its suitability (less protected and not guaranteed)

Our position: *We do not support extension of the test to single sector products. We suggest that reliance be placed on other mechanisms to protect members who invest in single sector products, in particular relying on them to look out for their own interests.*

7.5. Account-based pensions (ABPs)

We hold the view that retirees deserve the protection of some form of assessment in retirement as much as members in accumulation. Further, the pre-conditions for a performance test notionally appear to be in place (refer to Part A of the prior table). Nevertheless, we are hesitant to recommend introducing performance testing of investment products in the retirement phase for the following reasons:

- **One part of a bigger picture** – Testing ABPs would only be assessing one component of what determines member outcomes in retirement. We have consistently been vocal that there are many aspects to delivering a good retirement income strategy, including delivering a range of integrated retirement solutions and assistance to members in identifying a solution that is suitable for their needs. A related consideration is that the consequences of failure under the existing test do not gel with a situation where the investment option forms part of an integrated retirement solution. For instance, failure of the investment component does not imply that the overall solution is poor and that members should not continue to participate. In short, testing one component part in terms of investment performance without considering the overall picture is tenuous and sub-optimal.
- **Retirement is different** – A test that is valid for products in the accumulation phase may not translate directly to the income-focused retirement phase where other investment objectives may come into play such as limiting drawdowns to address sequencing risk and placing emphasis on inflation protection and tax efficiency. We acknowledge that these elements might be partly addressed under a performance test based around risk-adjusted returns or under a multi-metric test. Hence the force of our second misgiving may be subject to the nature of any test re-design.

Our preference would be that performance testing of ABPs is introduced at a later date as one component within a broader assessment framework for retirement income strategies. However, we offer one caveat on this general position that may open the door to introduce some level of performance testing of retirement products in the interim. This would involve testing of ABPs where two conditions are satisfied:

a) The investment strategy is a clone of (or highly similar to) an accumulation product that is subject to the test and has not been evolved for the retirement phase in ways that stretch the abilities of the prevailing performance test. For instance, it would be inappropriate to test an ABP that is managed to limit exposure to drawdown and inflation risk under the existing test structure.

b) The ABP constitutes the centrepiece of a fund's retirement solution, i.e. the ABP is not part of an integrated solution involving multiple building blocks. That is, the ABP is largely intended to deliver investment outcomes through generating returns rather than acting as a component within a retirement solution designed to deliver an income stream constructed from a range of sources[12].

If the ABP satisfies both of the above conditions it would be tested, otherwise it would be excused from the test. This approach accords with the central message from research conducted by Super Consumers Australia, which finds evidence of ABP clones of accumulation products failing the test but not being captured under the current test regime.

Our position: *Our primary preference is for testing of ABPs to be introduced at a later date as part of a broad retirement income strategy assessment framework. Nevertheless, in the interim we are open to the idea of testing ABPs that are effectively clones of existing tested accumulation options while not being embedded within integrated retirement solutions.*

[12] Integrated retirement solutions are currently not offered by any fund outside of those arising from personal financial advice process, although it is likely they will be in due course.

7.6. Assessment

While we identify that many of the same pre-conditions that led to the performance test being applied to MySuper default products also appear to exist in most of the candidate areas, there are many reasons for caution over further expanding the test. Among the four candidate areas considered, we see some case for expanding the test to EDPs but remain equivocal as the challenges of doing so could outweigh the benefits. We also see a case for expanding the test to ABPs on a limited basis where they are a clone of an existing tested accumulation product and do not form part of an integrated retirement solution. We recommend against expansion of the test to SMAs, single sector products or ABPs that have been explicitly designed to meet the needs of retirees.

In any event, we expect that extending the test to these areas will deliver low benefit relative to cost as compared to the initial application of the performance test to MySuper products. This is in part due to industry practices having improved since the initial introduction of the test. However, it is largely in recognition of the difficulty of accounting for the product and operational complexities in these areas through a generic test. In the case of SMAs and (particularly) single sector products, the greater likelihood that investing is occurring as a matter of informed choice coupled with operational flexibility to offer products in ways that could circumvent the test are important considerations.

Finally, one idea worth considering is whether the test could be put in place for a fixed period of time, say five years (i.e. a 'sunset clause'). This should be sufficient time for the benefits with respect to governance and reporting to be realised, after which the test could be withdrawn. In retrospect this would have been effective with regard to the initial introduction of the test. We hence wonder whether a sunset clause could be introduced for any areas of expansion if not the test itself.

8. Recommended multi-phase process for test adjustment

We recommend the following multi-phase process for making adjustments to the test structure:

(a) *Adjustments to test design or management are mandated following the current review*. We envisage that the mandated adjustments would be quite high level, e.g. a decision to (say) implement a side-pocket, shift to an alternative test structure such as introduce a multi-metric test, introduce a formalised process for review on the indices, extend testing to particular areas or perhaps even shift course and examine establishing an independent authority to oversee and manage the test. Target implementation dates may also be specified, if appropriate.

(b) *Implementation details are scoped out with assistance of a technical advisory group (TAG), or groups.* The TAG should comprise experts and ideally be formed from parties with limited direct conflicts as far as practicable. This process could be run by either Treasury or APRA, depending on the nature of the proposed adjustments.

(c) *Implementation details are released for further consultation.* This would give the industry at large a chance to have input and weed out any major problems that may have been overlooked.

(d) *Adjustments to the test are implemented.* Any adjustments are specified, changes of legislation made if required, and the adjustments then implemented while allowing a reasonable time frame for the industry and regulator to prepare.

A process similar to the above is currently evident with respect to the Retirement Reporting Framework, with APRA acting as technical experts by putting forward proposals for the indicators and metrics and the data fields to be provided in accordance with directions provided by Treasury.

A multi-phase process such as that above seems the most efficient path forward, noting that any adjustments will require consideration of challenging complex and often technical matters and careful design. The comment around complexity applies not only to all four options put forward in the consultation paper but also applies where a different course of action is pursued (e.g. implementing a multi-metric test).

APPENDIX A

Broad reference portfolio as an alternative

A test metric based around a broad reference portfolio (BRP) might be given consideration as an alternative to the SRP approach. The idea of a BRP aligns with one of two 'benchmark portfolios' (BP2) examined by the Productivity Commission in its 2018 report into the competitiveness and efficiency of the superannuation industry:

> "*Two types of benchmark portfolios (BPs) were constructed:*
>
> - *a listed benchmark portfolio* ***(BP1****) that captures investment performance (net of fees and taxes) of a set of investment strategies across a range of listed asset classes*
> - *a blended benchmark portfolio (****BP2****) that captures investment performance (net of fees and taxes) of a set of investment strategies across a range of listed and unlisted asset classes that more closely represents how funds implement their asset allocation.*" (PC report, p111)

Rationale to consider a BRP

Below are outlined reasons to consider a BRP in redesigning the YFYS test, which are discussed by focusing on comparison with a SRP.

1. *Conceptual basis* – It makes good sense to use a SRP in assessing the value added by super funds over what a member could readily access for themselves, i.e. a notional low-cost, passive portfolio of listed investments. However, the concept of benchmarking against an investment strategy that members might implement for themselves does not so readily translate to performance evaluation, which should ideally be assessing whether a fund has added value through the decisions made with regard to the available set of investment opportunities. A BRP might better align with this perspective.
2. *A BRP captures more of the investment opportunity set* – The SRP leaves out significant parts of the investment opportunity set, specifically alternatives including private assets which may constitute ~20% or more of the portfolios of many super funds. A BRP would help address this issue.
3. *BRP metric would be less exposed to performance of off-benchmark assets* – There is a risk that a large swath of funds could fail (or pass) a SRP test simply because of significant ex post differences in the relative risk-adjusted performance of alternatives versus the core listed asset classes, notwithstanding that a meaningful allocation to alternatives may be sensible ex ante. The potential for such a scenario to occur would be reduced under a BRP test.
4. *BRP creates clearer incentives to diversify* – As discussed in Section 5, the fact that certain alternative assets would be a significant source of TE under a SRP test may encourage some funds to limit their exposure to reduce their TE and hence risk of failure and consequently deliver less efficient portfolios to members. We note this depends on the circumstances and is thus not a foregone conclusion, as alternatives could be viewed as improving the risk-adjusted performance of portfolios and hence the probability of passing the test under certain assumptions. How a fund reacts might depend on their prevailing test buffer. A BRP test would more reliably encourage funds to diversify, at least in accordance with the asset mix contained within the BRP portfolio.
5. *A BRP test metric could be positioned as improving the structure of the existing test* – The BRP could be formed from the set of benchmark indices used for the existing performance test, and as such could amount to a form of restructuring the existing test. This would involve basing the BRP test metric on total portfolio returns versus the returns along the BRP frontier (in place of a SRP frontier), rather than estimating asset class performance versus the benchmark indices and then aggregating through weighting by declared SAA as occurs under the existing test.
6. *Shifting focus to total portfolio returns offers a number of advantages* – A BRP test would enable restructuring the existing performance test around total portfolio returns and facilitate bringing asset allocation and risk-adjustment into consideration. (Although these advantages are also available under the SRP as well as a peer-relative test based around total returns.)

Potential issues with a BRP test

Any test metric is imperfect and will give rise to issues. Below are the issues we see for a BRP test. Points 1, 2 and 3 also apply in general to a SRP test albeit based around a narrower set of asset indices.

1. *Risk of policy dictating portfolio construction at the asset class level* – The structure of the BRP could act as an anchor for the SAA of super funds, given that deviations from asset allocation embedded in the test would give rise to significant TE. To the extent that the BRP test influences super fund behaviour, it could have the effect of policy dictating industry practice and potentially acting as a barrier to adopting new assets classes and other forms of innovation at the asset class level.
2. *Trading one form of benchmark herding for another that could be more impactful* – Replacing the motivation to herd around the asset class benchmarks as arises under the existing test with an incentive to herd around the asset allocation embedded in the BRP would trade one form of herding for another. Further, herding around asset allocation may be more impactful, to the extent that asset allocation decisions can have a larger effect on portfolio performance and potentially financial market conditions.
3. *Currency hedging assumption is problematic* – The treatment of currency hedging is a key element of point 2 immediately above. The currency hedging assumption embedded in any BRP will be an important design decision. Large mismatches between a super fund's hedge ratio and that within the BRP have the potential to create significant TE that could alone result in a test pass or fail. How currency is treated could thus have significant implications for super fund hedging practices, with possibly systemic implications given the importance of super funds in currency markets.
4. *Liquidity considerations* - Under a BRP, funds that resolve to run a liquid strategy may not be able to replicate a BRP containing significant illiquid assets. The operating structures of some funds requires them to emphasise listed assets, most notably in the for-profit sector. It may be unfair to benchmark these funds against a BRP containing private market assets at higher weights than are prudently achievable under their operating structure.
5. *BRP performance is harder to track* – Super funds may have difficulty in tracking their performance in real time relative to a BRP test. While relative returns may be easier to track within a reasonable margin, risk will be more difficult as it is likely to be measured through analysis of returns realised that are unknown until after the assessment period. This tracking is more problematic with regard to private asset classes where indices are not available in real time.
6. *Implementation challenges in creating a BRP frontier* – While a performance frontier can be relatively straightforward to generate for a SRP through mixing a growth and defensive portfolios, it is less clear how it would work for a BRP where multiple asset classes with differing attributes are involved. (Note: We believe there are workable solutions).
7. *Risk of 'advocacy' for asset class inclusion in the BRP* – Given the impact of having an asset class included in the BRP, its composition is likely to be heavily contested. Two obvious sources of such contest include: (a) funds with sizable allocations to alternative assets versus funds that do not (i.e. jockeying over which of these groups need to deal with greater performance test TE); and (b) funds that represent a particular asset class or are pursuing a particular social cause.
8. *Alternative asset indices are not well suited to the purpose* – It is well-known that the benchmark indices used for private asset classes like unlisted infrastructure and property are not well-suited to performance assessment as they are un-investable and reflect any country or sector skews of assets currently held by funds contributing to the indices. Alternative assets classes such as private equity and hedge funds are also difficult to benchmark. Alternative asset indices may also be subject to return smoothing stemming from valuation processes. However, as this could also apply to super fund portfolios, the impact could wash out for a BRP to some degree.

APPENDIX B

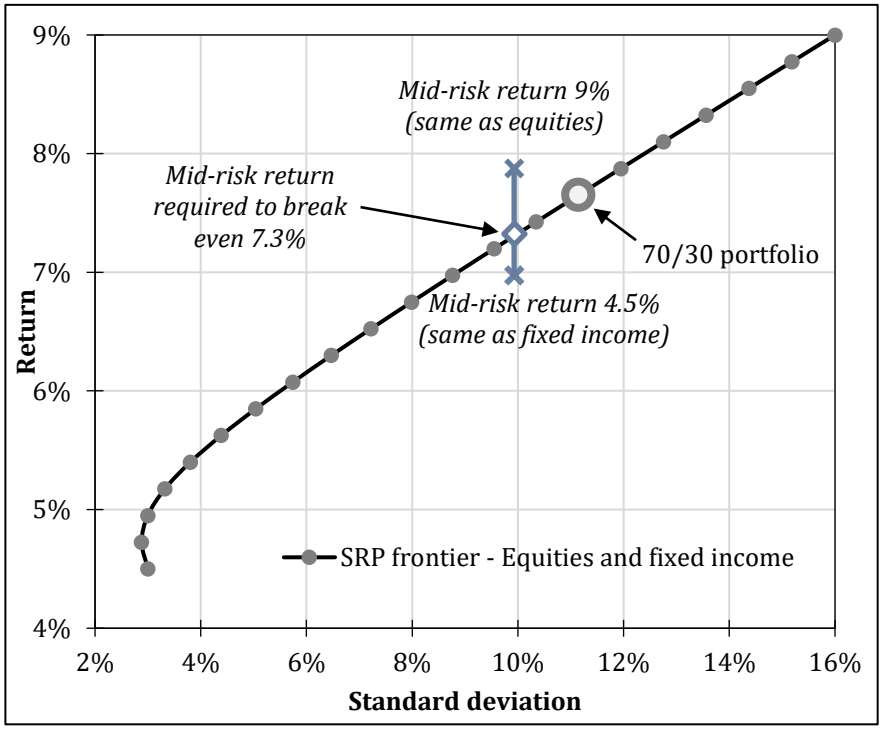
Impact of diversifying mid-risk assets

We present some notional modelling to gauge the incentive to diversify into mid-risk assets such as unlisted infrastructure and property. We assume that mid-risk assets will reduce the SD of the overall portfolio but also offer lower expected returns. Our focus is the 'breakeven' reduction in expected return versus equities where the fund would expect no detriment from adding a 20% weighting to mid-risk assets under the SRP test relative to a 70/30 equity/fixed income portfolio. We assume the allocation to mid-risk assets is funded 15% out of equities and 5% out of fixed income. Key asset return assumptions appear below in the table on the left with results in the chart on the right.

Modelling assumptions

	Fixed income	Mid-risk	Equities
Return	4.5%	*Imputed*	9.0%
SD	3.0%	10.0%	16.0%
Correlations			
Fixed income	1.00		
Mid-risk	0.10	1.00	
Equities	-0.10	0.50	1.00

Impact of mid-risk assets under the SRP test metric



Adding mid-risk assets reduces the SD of the portfolio and accordingly lowers the benchmark return by shifting the portfolio down the frontier relative to a 70/30 equity/fixed income mix. The breakeven expected return on mid-risk assets to match the benchmark return is imputed at 7.3%, or -1.7% below the assumed equity return of 9.0%. A fund might thus be encouraged to allocate to mid-risk assets if they are sufficiently confident that they beat the 7.3% hurdle. If not, they might be encouraged to hug the SRP benchmark. The chart also shows the range where the mid-risk assets offer a return equal to equities and equal to fixed income.

Bear in mind that the breakeven expected return estimate depends on the modelling assumptions. Our analysis does not consider confidence in achieving the expected return relative to equities (and fixed income). If the fund is not confident in the relative expected returns, it may be more inclined to limit exposure to mid-risk assets to limit the risk of test failure. Our assumptions of 10% SD for mid-risk assets may also exceed realisations of measured SD as a consequence of return smoothing stemming from valuation practices. In this case, the return hurdle to invest in mid-risk assets would be lower than 7.3%.