

7408778-the-market-sets-the-yardstick-a-state-contingent-standard-for-evaluating-institu

7408778: Trgovni standard za ocenjevanje portfelov - novi pristop do poslužjanja vrednosti

Članek se osredotoča na koncept "trgovnega standarda za ocenjevanje portfelov", ki predstavlja nov pristop do poslužjanja vrednosti in ocene institucionalnih portfelov. Avtor, Mario Bajo Traver, je v hlavi oddelka za strategije investicij v Banco de España.

Glavne nalože:

- Dinamični standard: Trgovni standard, ki se v ključnih momentih osredotoča na "preskoceno vredno" portfela v padežu trga in na "benčmark" v njegovem rastu, predstavlja preoblikovanje tradicionalne ocene portfelov.
- Reordanje portfelov: Ocena portfelov zato, da se osredotoča na ta dinamični standard, lahko vedne do reordanja njih v porovnanju z Sharpe razmerom. Portfeli s silno Sharpe razmerom mogo biti pogosto razmisleni kot slabji v okviru tega novega standarda.
- Defenzivni nalag: Portfeli z defenzivni nalag, ki se osredotočata na spremembo trga, lahko pridobita nišo v oceni, vendar ima tudi svoji upor.

Metodologija:

- Dato: Članek uporablja srednje tjedne vrate inkludirajuce 15 ameriških seznamov ETF-ov iz perioda od 2007 do 2026.
- Ocena portfelov: Osredotoča se na osem standardnih portfelov s različnimi mandati: akcij, fixnih dobi in multi-aktivni.
- Ključna metoda: V ključnih naložeh se uporablja "yardstick", ki izmerja kratkočasno padejo portfelov pod referenčno področje (preskoceno vredno ali benčmark).

Pomen za trgovino:

- Reordanje portfelov: Ta dinamični standard lahko vedne do reordanja portfelov v porovnanju z Sharpe razmerom, osredotočiv se na padeže trga.
- Defenzivni portfeli: Portfeli z defenzivni nalag, ki se osredotočata na spremembo trga, mogo biti bolj oceni, vendar tudi postopno smanjujeta realizirani maksimum dolgost.
- Uporne vysvětlenje: Ta pristop dodaja izboljš številno dimenzij za razumevanje, zakaj se institucionalni managerji hujata s trgem in kar uporabljajo za oblikovanje portfelov.

The Market Sets the Yardstick: A State-Contingent Standard for Evaluating Institutional Portfolios

THIS VERSION: SEPTEMBER 10, 2026

Mario Bajo Traver

Mario Bajo Traver is Head of Investment Strategies Unit, Department of Market Analysis and Intelligence, Banco de España.

mario.bajo@bde.es

KEY FINDINGS

- When markets fall, an institution can judge a portfolio against the safe return it gave up; when markets rise, against its benchmark. Scoring familiar portfolio constructions by this state-contingent, one-sided yardstick can reorder them relative to the Sharpe ratio.
- A construction with a strong Sharpe ratio can rank poorly under the yardstick. The reordering is driven by falling markets and, in our sample, appears only where the investment universe holds a reachable defensive asset.
- A defensive tilt can lower realized maximum drawdown yet score worse, which adds a measurement channel for benchmark hugging alongside the agency and career-concern explanations.

ABSTRACT

Institutional portfolio evaluation need not rely on a single fixed reference. We study a state-contingent yardstick that compares portfolios with a preservation floor when the benchmark falls and with the benchmark when it rises, counting shortfall but giving no credit for outperformance. Using weekly returns on fifteen US-listed ETFs from 2007 to 2026, we score eight standard portfolio constructions across equity, fixed-income, and multi-asset mandates. The yardstick produces rankings that differ from the Sharpe ratio, with the divergence concentrated in falling markets and, in our sample, appearing only where the investment universe holds a reachable defensive asset. We show that the up-market component of the yardstick is minimized by any portfolio that matches the benchmark; in our sample, active risk reduction then lowers realized maximum drawdown while raising the yardstick score. The resulting wedge provides a further measurement channel for institutional benchmark hugging alongside established agency and career-concern explanations. The analysis is descriptive: it identifies when a state-contingent evaluation standard and conventional measures disagree, and what that disagreement implies for how defensive portfolios are judged.

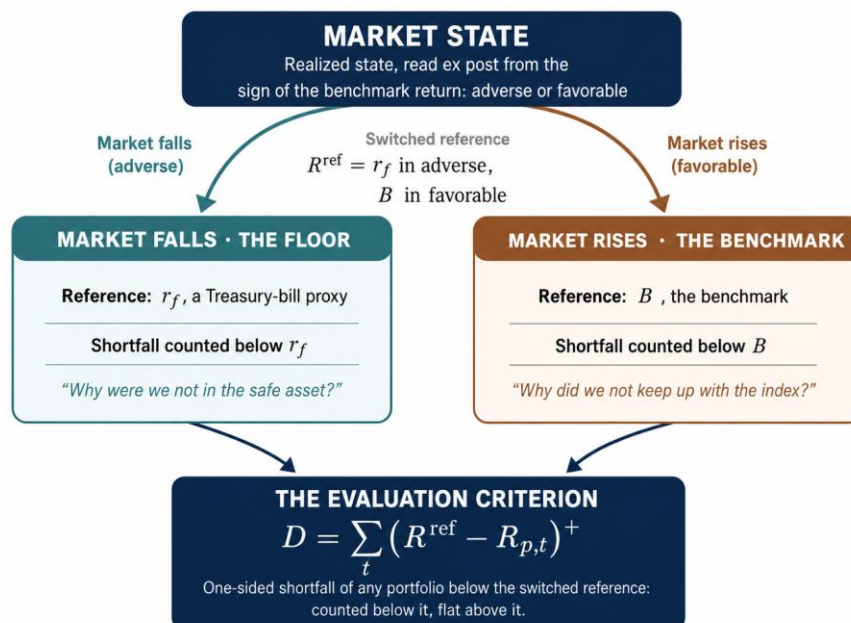
An investment committee reviewing a portfolio after a difficult year faces a question that its official benchmark cannot settle. A manager who stepped out of the market as it fell, and cut the portfolio's drawdown in the process, has done what a cautious owner would want. Yet once the market recovers, the same portfolio can look poor, because the yardstick has quietly changed: the owner is no longer asking whether capital was preserved but whether the portfolio kept pace with the index. The standard against which performance is judged is not fixed. It depends on the direction of the market.

This dependence has a simple structure. When the market falls, the natural reference is the safe return the portfolio gave up, a short-term Treasury-bill return, and the question is why the capital was not in the safe asset. When the market rises, the reference is the benchmark the mandate is judged against, and the question is why the portfolio did not keep up. The reference switches between two anchors of different economic nature, a preservation floor in falling markets and a benchmark in rising ones, and it registers a shortfall on one side only: falling below the prevailing reference counts, exceeding it does not.

We take this switched, one-sided standard as the object of study. It defines a single evaluation yardstick, denoted D : the shortfall of a portfolio below whichever reference the market selects, summed over time. The market direction is the realized one, read after the fact from the sign of the benchmark return, so the yardstick evaluates outcomes rather than forecasting them and needs no view on where the market is heading. It applies to any portfolio, and we treat it as a standard of evaluation, not as a measure of portfolio skill or superiority, nor as a documented claim about how investors think. Exhibit 1 shows the mechanism.

EXHIBIT 1

The market sets the evaluation yardstick



Notes: The realized market direction, read ex post from the sign of the benchmark return, selects the reference against which a portfolio is evaluated: a Treasury-bill proxy for the preservation floor when the market falls, and the benchmark when it rises. The yardstick, denoted D , is the one-sided shortfall of any portfolio below that switched reference.

Three features characterize this standard, and existing measures combine them only in part. The reference is set by the market rather than by the investor, is one-sided rather than symmetric, and switches between a floor and a benchmark rather than sitting at a fixed level. Conventional performance measures are symmetric and take no account of market direction: the Sharpe ratio scores return per unit of total volatility over the whole sample (Sharpe 1994), and the information ratio scores excess return per unit of tracking error, symmetrically around the benchmark (Roll 1992; Jorion 2003). Reference-dependent models of choice do let the reference move, but they place it inside the investor and update it along the investor's own wealth path, so it registers a downturn only after the portfolio has already lost (van Bilsen, Laeven, and Nijman 2020). The regime-switching literature reacts to the market in time, but it switches the estimated inputs to a single fixed objective rather than the standard of evaluation itself (Ang and Bekaert 2004; Bouyé and Teiletche 2025). One-sided shortfall against a reference is the language of regret, of safety-first criteria, and of the lower partial moment tradition (Bell 1982; Roy 1952; Bawa 1975; Fishburn 1977), but these standard formulations take the reference as fixed within the evaluation rule. Partial-moment methods can let the target vary over time: Viole and Nawrocki (2016), for instance, take the greater of the asset-class benchmark and the risk-free return as a period-by-period target within a ratio of upper to lower partial moments. The yardstick differs in construction: it counts shortfall on one side only and gives no credit for outperformance. The contribution is an *ex post* standard for institutional evaluation whose reference combines all three features: it is set by the market, one-sided, and switching.

Two findings follow, and both are practical. The first is that scoring standard portfolio constructions by this yardstick reorders them relative to the Sharpe ratio. A construction that earns a strong risk-adjusted return by taking risks that show up in a downturn is rewarded by a measure that averages over the whole sample but penalized by one that isolates the shortfall below the floor when the market falls. The reordering concentrates in falling markets, and in our sample it appears only where the investment universe holds a reachable defensive asset, being absent in a fully invested equity mandate that holds none. The account is descriptive: with eight constructions the rank comparison is an effect size, not a test, and the defensible statement is that the yardstick does not reproduce the conventional ordering, not that it inverts it.

The second finding concerns portfolios that step away from the benchmark to reduce risk. Because a portfolio that matches the benchmark records no shortfall in rising markets, the up-market half of the yardstick pulls portfolios toward the index. The consequence is uncomfortable and, for a committee, familiar: a portfolio that actively lowers its realized drawdown can record a worse score at the same time, because the protection is bought by departing from the benchmark when the market rises. This gives a measurement channel for institutional benchmark hugging, one that follows from the standard used to judge performance rather than from the manager's incentives, and that sits alongside the agency, career-concern, and active-share explanations rather than replacing them.

The contribution is neither a new estimator nor a claim that any portfolio dominates another. It is a way to see when a conventional yardstick and a state-contingent one part company, why they do, and what that implies for how a defensive tilt is judged. The rest of the paper sets out the yardstick and the two readings under which it is economically motivated, describes the data and the eight constructions, presents the two findings, and closes with how an investment committee can read the yardstick alongside the measures it already uses. Formal definitions and the supporting inference are collected in the appendices.

A YARDSTICK THAT SWITCHES WITH THE MARKET

Two agents, two standards

Two agents stand behind an institutional portfolio. The manager implements it within a set investment policy, an asset class or set of classes and a construction rule, and may, within those limits, de-risk or compete for return. The owner evaluates the result. The yardstick studied here is a standard for that evaluation, and it sits beside the official one. The official standard is always the mandate's benchmark, the language in which performance is formally reported. The second standard is state-contingent: when the market falls, the portfolio is evaluated against the safe return it could have earned; when the market rises, against the benchmark it was meant to track. The reference in this second standard is tied to the direction of the market, not to the owner and not to the manager.

What the yardstick measures

The yardstick is the total shortfall of the portfolio below the switched reference, counted only when the portfolio falls short. In a falling market it accrues only when the portfolio drops below the floor; in a rising market only when it trails the benchmark. Above the prevailing reference it is flat: beating the benchmark in an advance, or clearing the floor in a downturn, adds nothing. Writing B_t for the benchmark return in week t , r_f for the preservation floor, and $R_{p,t}$ for the portfolio return, the switched reference is

$$R_t^{ref} = \begin{cases} r_f & \text{in an adverse week } (B_t < 0), \\ B_t & \text{in a favorable week } (B_t \geq 0). \end{cases} \quad (1)$$

and the yardstick is its summed one-sided shortfall below that reference,

$$\mathcal{D} = \sum_t (R_t^{ref} - R_{p,t})^+, \quad (2)$$

where $(x)^+ = \max(x, 0)$ counts shortfalls only. Exhibit 1 states the same rule in words, and Appendix A gives the formal treatment.

We report the yardstick as the average weekly shortfall relative to the switched reference, annualized and expressed in percentage points. It is therefore an average shortfall score in return units, not a compounded loss and not comparable in level to a Sharpe ratio: a value of 30 does not mean a 30 percent loss.

Why the shortfall is one-sided

The one-sidedness is deliberate and is what separates the yardstick from symmetric performance measures. It records the deficit below the prevailing reference but gives no credit for exceeding it. This is the structure of regret in the decision-theoretic sense (Bell 1982; Loomes and Sugden 1982): an outcome is weighed against the reference it fell short of, and the reference stands for the alternative the owner could have had, the safe asset in a downturn or the index in an advance. The shortfall is measured in return, not in risk. The portfolio is assumed to respect the

risk limits of its mandate, so the yardstick asks what return was delivered relative to the reference, not whether the portfolio was too volatile.

A measure of outcomes, not a trading rule

The yardstick evaluates what happened; it is not itself a trading rule a portfolio can follow, nor a measure of the ability that produced the outcome, and this distinction matters for the second finding. The market direction that selects the reference is realized, read after the fact from the sign of the benchmark return, so a week is adverse if the benchmark fell and favorable if it rose. A portfolio, by contrast, must act before the outcome is known, on information available at the time. The portfolios examined in the second finding therefore use a decision signal available in advance, while the yardstick scores their realized outcomes after the fact. The gap between what a portfolio can know when it acts and the direction the market turns out to take is central to why a defensive tilt can score poorly. We classify each week by the contemporaneous sign of the benchmark return, without identifying a persistent or estimated regime, and use market direction rather than market regime throughout. If managers were evaluated against this standard *ex ante*, the rule could itself affect portfolio choices, including incentives to anticipate market direction. We do not model this strategic response here; the analysis isolates the evaluation wedge created by the standard.

Two readings

The switched reference can be read two ways, and the same yardstick applies under both. The primary reading is rational and state-dependent. For a central bank, a sovereign fund, or a pension plan, preserving capital through a decline can be a legitimate objective, and judging a downturn against the safe floor is then a natural assessment of that objective rather than a bias; under this reading the yardstick needs no assumption about perception, and it is the reading we adopt. A second reading is behavioral: an owner may code outcomes as gains and losses against a shifting reference and feel losses more keenly, as in prospect theory and goal-based approaches to portfolio choice (Kahneman and Tversky 1979; Shefrin and Statman 2000). This reading is an interpretation, not something the analysis establishes, and whether institutional owners in fact evaluate this way is a question for separate evidence from holdings, flows, or direct elicitation. The findings hold under either reading.

What the yardstick is not

Three clarifications fix the object. It is not a relabeling of a conventional measure: the Sharpe ratio and the information ratio are both symmetric and take no account of market direction, while the yardstick is one-sided and switches its reference. Nor is it a fixed-reference downside measure. Measures such as the Sortino ratio, a mean shortfall below a fixed floor, or a downside tracking loss below the benchmark all hold their reference fixed; the yardstick's defining feature is that the reference switches with the market, and Exhibit 5 shows that its ranking of standard constructions differs from the rankings those fixed-reference measures produce.

Nor, finally, is it an upper/lower partial-moment ratio. The switched target of Viole and Nawrocki (2016) enters both the upper and lower partial moments, so performance above the reference remains part of the metric; the yardstick keeps only the shortfall below it. This is not cosmetic: it is the removal of the upside component that makes the benchmark a minimizer of the up-market half of the yardstick, the property behind the benchmark-attraction result, which an upper/lower partial-moment ratio, by rewarding outperformance, does not deliver.

DATA AND CONSTRUCTIONS

The two findings are computed from the same data and portfolios. The data are weekly total returns, with dividends reinvested, for fifteen US-listed exchange-traded funds from October 2007 to June 2026, a span of 976 weeks that covers the global financial crisis, the 2020 collapse, and the 2022 stock-and-bond selloff; the source is Bloomberg. We form three mandates, each a universe of six funds: equity, fixed income, and multi-asset. Each mandate carries a benchmark built from its own funds, an S&P 500 fund for equity, an aggregate bond fund for fixed income, and a 60/40 blend of the two for multi-asset, so the three benchmarks form a coherent set. The preservation floor is a short-term Treasury-bill fund held outside all three universes; because the portfolios are fully invested within their own universe, they cannot hold the floor asset directly, a point that matters for the second finding. The fund list is in Appendix B and the formal definitions in Appendix A.

We score eight standard portfolio constructions, fixed in advance to span the families used in practice: two naive allocations (equal weighting and inverse volatility), a variance optimizer (minimum variance), a risk-return optimizer (maximum Sharpe), two risk-based constructions (equal risk contribution and maximum diversification), and two hierarchical constructions (hierarchical risk parity and hierarchical equal risk contribution). Each is formed on a rolling one-year window and held forward, producing a realized return path over a common out-of-sample period. The Sharpe ratio is the primary conventional measure, with the information ratio reported alongside. Because there are only eight constructions per mandate, the rank comparisons that follow are read as effect sizes, not tests, and Appendix E reports exact permutation p-values and block-bootstrap intervals so the strength of each comparison is explicit.

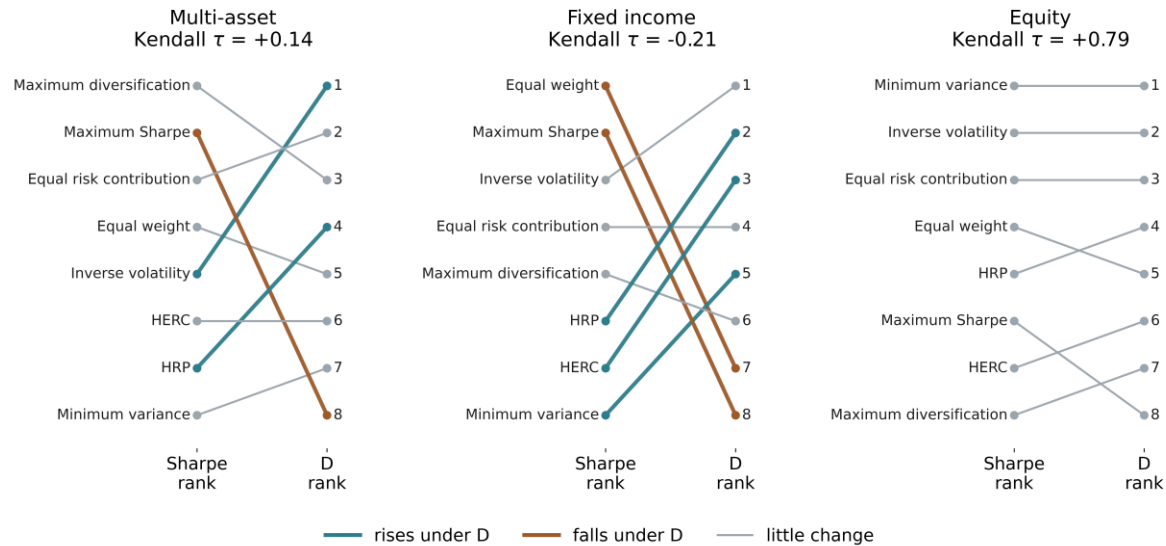
THE YARDSTICK CAN REORDER STANDARD PORTFOLIOS

Scoring the eight constructions by the yardstick and by the Sharpe ratio produces two different rankings, and the difference is largest in the mandates where a defensive position is attainable. Exhibit 2 shows the two rankings side by side in each mandate, with a line joining each construction's rank under the Sharpe ratio to its rank under the yardstick.

EXHIBIT 2

The two rankings side by side, by mandate

The switched criterion reorders the constructions (rank under Sharpe vs under D)



NOTES: For each mandate, the rank of the eight constructions under the Sharpe ratio and under the yardstick, with a line joining each construction's two ranks. Rank 1 is best. Kendall τ between the two rankings is shown above each panel.

In the two mandates with a reachable defensive asset the rankings pull apart. In fixed income, the equal-weight portfolio ranks first under the Sharpe ratio but next to last under the yardstick, while the maximum-Sharpe construction ranks second under the Sharpe ratio and last under the yardstick; the inverse-volatility and hierarchical constructions, middling or low under the Sharpe ratio, rise to the top. In multi-asset the maximum-Sharpe construction shows the same move, second under the Sharpe ratio and last under the yardstick. Exhibit 3 lists the rankings for these two mandates. The average construction moves 2.25 places in multi-asset and 3.25 in fixed income, with individual moves as large as six places, against 0.75 places and a maximum of two in equity.

EXHIBIT 3

Construction rankings under the Sharpe ratio and the yardstick

Construction	Sharpe ratio	Rank	D	Rank
Multi-asset (benchmark 60/40 SPY-AGG)				
Inverse volatility	0.55	5	31.13	1
Equal risk contribution	0.59	3	31.61	2
Maximum diversification	0.69	1	32.07	3
HRP	0.40	7	33.17	4
Equal weight	0.58	4	33.46	5
HERC	0.42	6	33.54	6
Minimum variance	0.36	8	33.99	7

Construction	Sharpe ratio	Rank	D	Rank
Maximum Sharpe	0.63	2	38.94	8
Fixed income (benchmark AGG)				
Inverse volatility	0.23	3	12.94	1
HRP	0.04	6	13.20	2
HERC	0.03	7	13.28	3
Equal risk contribution	0.14	4	13.31	4
Minimum variance	−0.02	8	13.40	5
Maximum diversification	0.09	5	13.83	6
Equal weight	0.38	1	15.03	7
Maximum Sharpe	0.30	2	16.33	8

NOTES: Rankings for the two mandates with a reachable defensive asset, ordered by yardstick rank. D is the yardstick, the average weekly shortfall against the switched reference, annualized in percentage points; lower values indicate better evaluation. Values are gross of transaction costs over the common out-of-sample period, October 2008 to June 2026.

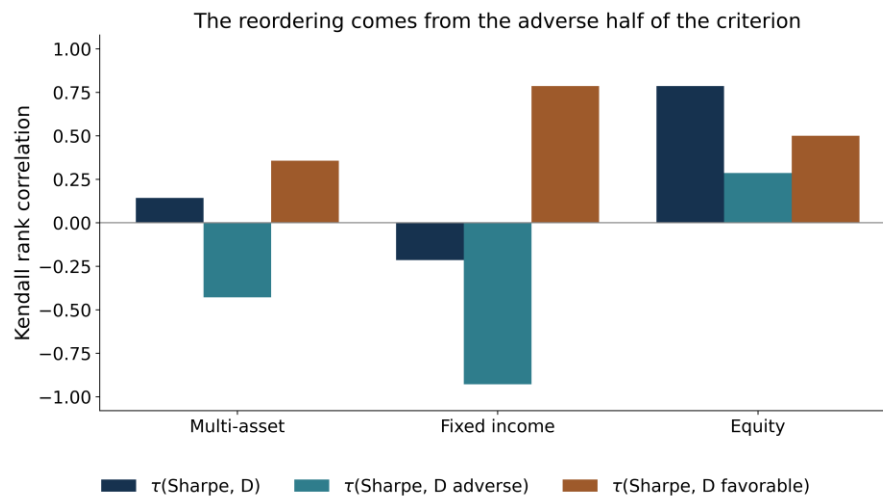
The aggregate rank correlation tells the same story with appropriate caution. Between the yardstick and the Sharpe ratio, Kendall τ is +0.14 in multi-asset and −0.21 in fixed income; with eight constructions neither is distinguishable from zero (Exhibit E1). The honest reading is therefore that the sample shows no systematic agreement between the two rankings, not that the yardstick inverts the Sharpe ranking. That is enough to establish the point: the yardstick does not reproduce the ordering a conventional measure produces. The information ratio gives the same picture, with τ of −0.14 in multi-asset and −0.50 in fixed income.

Where the reordering comes from

The disagreement lives in the down-market half of the yardstick. Exhibit 4 computes the rank correlation with the Sharpe ranking separately for the down-market and up-market components of the yardstick. In both mandates with a defensive asset, the down-market component runs against the Sharpe ranking: in fixed income it is a near-complete inversion ($\tau = -0.93$, exact permutation $p < 0.001$), and in multi-asset it points the same way ($\tau = -0.43$) but, with only eight constructions, is not distinguishable from zero. The up-market component, by contrast, agrees with the Sharpe ranking in both, because keeping pace with the benchmark in an advance and earning a high risk-adjusted return both reward the same upside participation.

EXHIBIT 4

The reordering comes from the down-market half of the yardstick



NOTES: For each mandate, the Kendall rank correlation between the Sharpe ranking and the yardstick ranking, and between the Sharpe ranking and the down-market (adverse) and up-market (favorable) components of the yardstick.

The reason is direct. A construction that earns a high risk-adjusted return by taking risks that materialize in a downturn, credit and duration risk in fixed income, or market beta in multi-asset, is rewarded by a measure that averages over the whole sample and penalized by one that isolates the shortfall below the floor when the market falls. The yardstick does not reward or punish return as such; it re-scores the same returns by the market direction in which they were earned. Consistent with this, up-weighting the down-market half strengthens the disagreement and up-weighting the up-market half weakens it, and the pattern holds under five pre-specified definitions of an adverse market, from a volatility-scaled band to a five-percent drawdown rule (Appendix J).

The result is not specific to the eight constructions of the main exercise: Appendix L replicates the analysis on a broader set of thirty economically distinct constructions, spanning eleven portfolio-construction families, and obtains the same adverse-state reordering, including after collapsing near-duplicate methods to cluster representatives.

It is the switch, not downside measurement in general

The finding would be less interesting if it merely reflected downside measurement rather than the switch in the reference. It does not. Against fixed-reference downside measures that do not switch, a Sortino ratio, a mean shortfall below the floor in every week, and a mean shortfall below the benchmark in every week, the yardstick still ranks the constructions differently in both mandates with a defensive asset (Exhibit 5). A broader set of eight risk and downside measures, from Omega to expected shortfall, does not reproduce its ranking either; its closest relative is a one-sided tracking loss below the benchmark, which shares the up-market half but lacks the switch to the floor in down markets (Appendix K).

EXHIBIT 5

The yardstick versus fixed-reference downside measures

Mandate	$\tau(D, \text{Sharpe})$	$\tau(D, \text{Sortino})$	$\tau(D, \text{floor})$	$\tau(D, \text{benchmark})$
Multi-asset	+0.14	+0.14	-0.29	+0.71
Fixed income	-0.21	-0.21	+0.29	+0.36
Equity	+0.79	+0.71	+0.79	+0.86

NOTES: Kendall τ between the yardstick (D) and, respectively, the Sharpe ratio, a Sortino ratio, the one-sided shortfall below the floor in every week, and the one-sided mean shortfall below the benchmark in every week, by mandate. All measures are oriented so that higher values indicate closer agreement; a low value therefore means the yardstick ranks the eight constructions differently from the comparison measure. The static benchmark is excluded.

Scope: the defensive-asset condition

The equity mandate is a deliberate boundary, not a failure. There the yardstick and the Sharpe ratio agree closely ($\tau = +0.79$), and the reason is structural: the equity universe holds no reachable defensive asset, so no construction can set itself apart through preservation when the market falls, and the down-market half of the yardstick has little to work with. The boundary shows up in the construction space too: with no defensive asset the constructions converge, so a broad thirty-construction set collapses to only a handful of distinct methods in the equity mandate, against roughly twice as many where a defensive asset is present (Appendix L). A reachable defensive asset is what lets preservation separate the constructions in a downturn, as in the multi-asset and fixed-income mandates. The condition is measurable rather than merely descriptive: across variant universes, the strength of the down-market inversion tracks a pre-specified index of how well the best in-universe defensive portfolio holds up in falling markets (Spearman $\rho = +0.89$), and adding to an all-equity universe an asset that holds up when equities fall, short Treasuries or aggregate bonds but not gold, produces that inversion even where the net ranking stays close to the Sharpe ranking (Appendix I). A committee can therefore check whether the condition holds before relying on the yardstick.

One portfolio has a special place. The mandate's own benchmark records no shortfall in rising markets, since it matches its own return in an advance, and among the tested in-universe portfolios it also records the lowest total score in every mandate. The first statement is a property of a one-sided, benchmark-referenced standard; the second is an empirical result for the portfolios considered here. Together they set up the second finding: the up-market half of the yardstick pulls portfolios toward the index, even as the down-market half reorders them.

WHY A DEFENSIVE TILT CAN SCORE WORSE

The second finding is the one a committee is most likely to meet in practice: a portfolio that lowers its realized risk can record a worse score on the yardstick at the same time. It has two parts, one that follows from the yardstick's structure and one that does not.

Benchmark attraction

The structural part is that the up-market half of the yardstick is minimized by the benchmark itself. A portfolio that matches the benchmark in every rising market records no up-market shortfall, gains nothing from beating it, and adds shortfall only when it falls below it, so replicating the benchmark is the least-shortfall choice for that half by construction. This is a property of the standard, not a finding in the data, and it is stated formally as Proposition 1 in

Appendix A. Two points qualify it. The benchmark is one of many portfolios that reach zero up-market shortfall, not the only one, and it need not minimize the total score, because it still falls below the floor in the down markets where the benchmark itself is negative. And the pull toward the benchmark comes from reachability, not optimality: the benchmark removes the up-market half without foresight and without leaving the investable universe, whereas the floor sits outside the universe and the market direction is unknown when the portfolio must act. Whether the benchmark minimizes the total score is an empirical question; among the portfolios tested, none beats it, so it records the lowest score of the constructions considered, but this is a finding, not a consequence of Proposition 1.

Active risk reduction can raise the score

The part that does not follow from the algebra is what happens when a portfolio departs from the benchmark to manage risk. Exhibit 6 compares the benchmark with two active portfolios built on the same universe. Both hold an in-universe preservation sleeve when their decision signal is adverse and differ in what they hold otherwise: the defensive tracker holds the benchmark, while the dual-anchor holds a maximum-Sharpe portfolio and competes for return. Both use a decision signal available in advance, the sign of the trailing four-week benchmark return, while the yardstick scores their realized outcomes after the fact.

EXHIBIT 6

Score, its components, and risk for the benchmark and two active portfolios

Portfolio	D (adverse)	D (favorable)	D	Sharpe	Max DD	Turnover
Multi-asset (benchmark 60/40 SPY-AGG)						
Static benchmark	23.66	0.00	23.66	0.73	−21.3	0.0
Defensive tracker	16.79	11.60	28.39	0.60	−18.4	25.9
Dual-anchor	12.87	23.24	36.11	0.58	−18.3	26.3
Fixed income (benchmark AGG)						
Static benchmark	11.31	0.00	11.31	0.32	−18.2	0.0
Defensive tracker	7.42	4.57	11.99	0.32	−9.5	49.7
Dual-anchor	5.72	9.26	14.98	0.23	−12.1	28.5

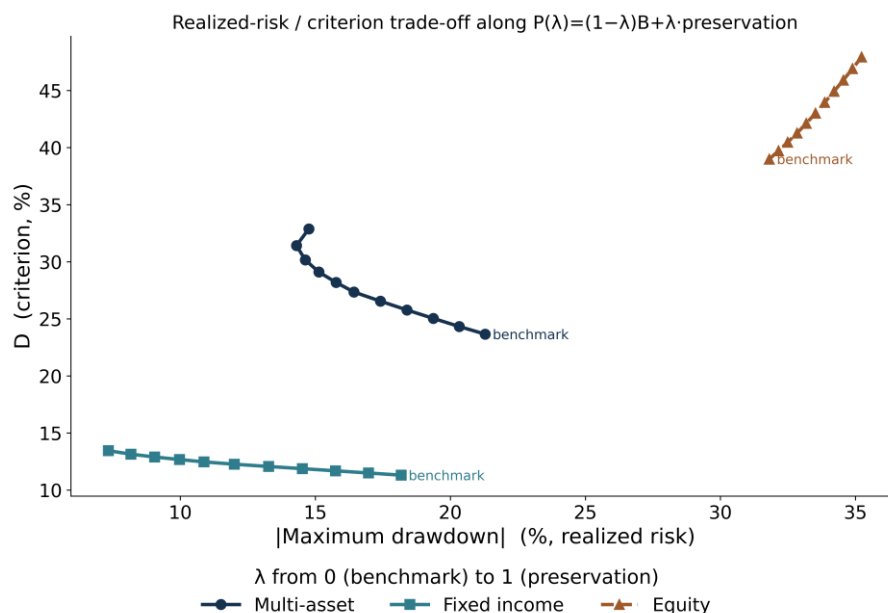
NOTES: For the static benchmark, defensive tracker, and dual-anchor, in the two mandates with a reachable defensive asset: the adverse and favorable components of the yardstick, and the yardstick D itself, the Sharpe ratio, maximum drawdown (Max DD), and monthly turnover. D is the average weekly shortfall against the switched reference, annualized in percentage points; lower values indicate smaller shortfall. Maximum drawdown is a peak-to-trough measure of realized risk. Turnover is the average one-way fraction traded per month. Values are gross of transaction costs over the common out-of-sample period.

Read the table through its score and drawdown columns. From the benchmark to the tracker to the dual-anchor, maximum drawdown falls from 21.3 to 18.4 to 18.3 percent in multi-asset and from 18.2 to 9.5 to 12.1 percent in fixed income: active management lowers realized maximum drawdown. Over the same portfolios the yardstick rises, from 23.7 to 28.4 to 36.1 in multi-asset and from 11.3 to 12.0 to 15.0 in fixed income. Realized maximum drawdown and the score move in opposite directions, because the protection is bought by stepping away from the benchmark when the market rises.

The trade-off is not specific to these three portfolios. Exhibit 7 traces a continuum from the benchmark to full preservation and records maximum drawdown and the score along it. Where a reachable defensive asset exists, moving toward preservation lowers drawdown while raising the score; in the equity mandate, where the preservation sleeve is itself all equity, moving toward it raises both. The wedge between realized risk and the score is therefore visible along the continuum where a defensive asset exists and absent where it does not.

EXHIBIT 7

The realized-risk and score trade-off along the de-risking continuum



A measurement channel for benchmark hugging

The wedge gives an account of institutional benchmark hugging that follows from the evaluation standard rather than from the manager's incentives. The literature explains the pull toward the benchmark through agency and contracting (Basak and Pavlova 2013), career concerns (Chevalier and Ellison 1997), and the discipline of active share (Cremers and Petajisto 2009) and tracking-error budgets. The yardstick adds a further channel: when performance is judged against the benchmark in rising markets and a preservation floor in falling ones, active risk reduction becomes costly in the very standard used to judge it. This channel does not replace the others; it coexists with them, because it comes from the standard rather than the contract. The exercise does not measure how much of observed benchmark hugging this channel explains; it shows that the standard alone can generate the wedge.

Scope: the defensive-asset condition

As with the first finding, the wedge appears in our sample only where the mandate holds a reachable defensive asset. In the equity mandate the preservation sleeve is itself fully invested in equities, so de-risking cannot lower drawdown: the tracker and dual-anchor have larger drawdowns than the benchmark while their scores rise, so they are worse on both counts and there is no trade-off to speak of. Where a defensive asset exists, active management can trade a higher score for lower realized risk; where it does not, that trade did not arise in our sample.

USING THE YARDSTICK IN PRACTICE

The yardstick earns its place not as a competitor to the Sharpe ratio but as a second lens an investment committee can read beside the measures it already uses. Its practical content is that it makes visible a distinction the official benchmark and conventional summary measures do not: the gap between how much realized risk a portfolio took and how that portfolio is judged once the yardstick switches with the market. The practical questions are where it applies, what it reports, how it complements existing tools, and how it should not be used.

Where it is informative

The first thing the yardstick tells a committee is where it applies at all. In our sample, both findings appear only in mandates whose universe holds a reachable defensive asset, and both are absent in a fully invested equity mandate that holds none. This is not a limitation to be hidden but the yardstick's scope of use, and it suggests a practical rule: apply the yardstick where a defensive position is attainable within the mandate, as in the multi-asset and fixed-income universes here, and expect it to add little beyond a conventional measure where it is not, as in an all-equity sleeve. The condition is measurable, and a committee can check it in advance rather than discover after the fact that the universe offered no asset in which to preserve capital.

What it reports

Where it applies, the yardstick reports in units a committee already reads: percentage points of shortfall per year, split by the market direction in which the shortfall occurred. A statement such as *“in the multi-asset mandate a defensive portfolio lowered maximum drawdown from 21.3 to 18.4 percent and cut down-market shortfall by 6.9 points per year, but added 11.6 points of up-market shortfall, for a net score 4.7 points above the benchmark”* adds information beyond tracking error or volatility, because it separates the size of the shortfall from the market direction

that produced it and ties both to a reference the committee recognizes. The same split distinguishes protecting the downside from competing: stepping out of the market on a lagged signal cost a net 4.7 points in multi-asset and was not distinguishable from zero in fixed income, while competing in rising markets cost a further 7.7 and 3.0 points, all of it and more in up-market shortfall. This decomposition is most useful where leadership rotates across categories, in multi-asset allocation and in credit-versus-Treasury rotation within fixed income, settings in which the mandated benchmark cannot by itself show how a defensive tilt is scored once the market turns up.

How it fits alongside existing tools

Because the yardstick scores realized outcomes against a rule rather than against forecast returns, it is implementable with standard reporting and sits beside an existing risk-budgeting process rather than replacing it. A committee can run its volatility or risk-parity process unchanged and read the yardstick as an added diagnostic on how a portfolio's departures from the benchmark are judged across market directions. It is a complement to conventional evaluation, not a substitute, and nothing here implies that a portfolio scoring well on the yardstick is a better portfolio.

One caution

One caution governs its use: the yardstick is a lens for making a trade-off explicit, not an objective to be minimized. The second finding shows that a portfolio can reduce realized downside risk and still record a higher score, because protection is bought by departing from the benchmark when the market rises. A committee that treated a low score as the goal would be rewarding the portfolio that hugs the benchmark most closely in rising markets, which is the very behavior the wedge helps to diagnose. The yardstick should be read to separate an improvement in realized risk from a difference in how that improvement is judged, not to drive the allocation toward whichever portfolio the switched standard scores lowest.

CONCLUSION

An institutional portfolio can be judged against more than one standard, and the standard can depend on the market. This paper studies a single yardstick that captures that dependence: it measures the shortfall of a portfolio below a preservation floor when the market falls and below the benchmark when it rises, counting shortfalls but not credits above the mark. We treat it as a standard of evaluation rather than a documented fact about how owners think, and ask what follows when portfolios are scored by it.

Two findings follow. First, the yardstick can rank standard portfolio constructions differently from a conventional measure. With eight constructions the aggregate rank correlation is an effect size rather than a test, and in the mandates with a reachable defensive asset it shows no systematic agreement with the Sharpe ranking; the disagreement concentrates in falling markets, where in fixed income it is a near-complete inversion, and it comes from the switch in the reference rather than from downside measurement in general. Second, the up-market half of the yardstick pulls portfolios toward the benchmark, and a portfolio that actively lowers its realized drawdown can record a higher score at the same time, because the protection is bought by stepping away from the benchmark when the market rises. The resulting wedge gives a measurement channel for institutional benchmark hugging that sits alongside the agency, career-concern, and active-share explanations rather than replacing them.

Both findings share a boundary, tested rather than assumed: in our sample they appear where the mandate holds a reachable defensive asset and not in a fully invested equity mandate that holds none. The interpretation has a limit too. The rational reading treats the switched reference as the right assessment of a state-dependent institutional objective and needs no assumption about perception; the behavioral reading treats it as a possible reference-dependent evaluation by owners, but the analysis does not establish that owners in fact use it, which would require evidence from holdings, flows, or direct elicitation.

Within these bounds, the paper identifies a measurable distinction between a portfolio's outcome and the standard used to judge it. A portfolio can reduce realized downside risk and still receive a worse evaluation, because the benchmark becomes the reference once the market rises. The contribution is not a claim about which portfolio is better. It is a way to see when a conventional yardstick and a state-contingent one diverge, what goes with that divergence, and how the resulting wedge can coexist with familiar explanations of why institutions hug their benchmarks.

APPENDIX A

THE YARDSTICK, FORMALLY

This appendix collects the formal definition of the yardstick, its properties, and the benchmark-attraction result. Let B_t denote the mandate's benchmark return in week t and let r_f denote the preservation floor, a short-term risk-free return proxied in the empirical work by a Treasury-bill fund held outside the investable universe. The evaluation state is realized ex post from the sign of the benchmark return, so a week is adverse when the benchmark falls and favorable when it rises. This is distinct from any decision state a portfolio uses, which depends only on information available up to the prior week; the yardstick depends on the evaluation state alone and cannot be applied as a forward-looking rule without specifying how that state is anticipated.

Definition and states

The yardstick is built on the switched reference of Equation (1) and the one-sided shortfall of Equation (2), with a benchmark return of exactly zero classified as favorable by convention, a probability-zero case for continuous returns. Writing $(x)^+ = \max(x, 0)$, and expressing the yardstick for a portfolio P as a population expectation with indicator functions,

$$D(P) = \mathbb{E} \left[\mathbf{1}_{B_t < 0} (r_f - R_{P,t})^+ + \mathbf{1}_{B_t \geq 0} (B_t - R_{P,t})^+ \right]. \quad (3)$$

In an adverse week the shortfall accrues only below the floor and in a favorable week only below the benchmark; above the prevailing reference the yardstick is flat. We report it as the average weekly shortfall, annualized by a factor of 52 and expressed in percentage points, a shortfall in return units that is not cardinally comparable to a Sharpe ratio.

Properties

Three properties follow from this one-sided, state-switched structure. The yardstick is non-negative and flat above the prevailing reference, so it is insensitive to the margin by which the reference is exceeded; it is non-increasing in the portfolio return in each week; and, because the positive part is convex and portfolio return is linear in the weights, each summand is convex in the weights for fixed references, so minimizing the yardstick over a convex set of weights is a convex program. The flatness has one consequence used in the second finding: in each state the entire set of portfolios that reach or exceed the reference achieves zero shortfall there, so many portfolios can share the minimum, the benchmark being only one of them. The yardstick is also a state-switched splice of two familiar objects, a lower partial moment of order one below the floor in adverse weeks (Bawa 1975; Fishburn 1977) and a one-sided below-benchmark tracking loss in favorable weeks; the switch between them, combined with the one-sided shortfall, is its distinguishing feature.

Benchmark attraction: proposition and proof

Let the benchmark be attainable within the mandate, meaning a feasible portfolio matches the benchmark return in every week. Decomposing the yardstick by market direction,

$$D = \sum_{t: B_t < 0} (r_f - R_{P,t})^+ + \sum_{t: B_t \geq 0} (B_t - R_{P,t})^+.$$

Proposition 1. *Under an attainable benchmark and the switched yardstick above: (a) in a favorable week every portfolio with $R_{P,t} \geq B_t$ incurs no favorable-state shortfall, so the benchmark weakly minimizes the favorable-state component, and the set of weak minimizers in that week is*

$$\mathcal{P}_t^F = \{P: R_{P,t} \geq B_t\};$$

(b) in an adverse week every portfolio with $R_{P,t} \geq r_f$ incurs no adverse-state shortfall, with corresponding set

$$\mathcal{P}_t^A = \{P: R_{P,t} \geq r_f\};$$

(c) if the intersection

$$\left(\bigcap_{t: B_t \geq 0} \mathcal{P}_t^F \right) \cap \left(\bigcap_{t: B_t < 0} \mathcal{P}_t^A \right)$$

is nonempty, the yardstick attains its global minimum $D = 0$. The benchmark need not belong to this intersection, since it incurs a positive adverse-state shortfall whenever $B_t < r_f$.

Each summand is non-negative and vanishes exactly when the portfolio reaches or exceeds its applicable reference, so parts (a), (b), and (c) follow directly. Two readings of the proposition matter. The benchmark is a weak minimizer of the favorable-state component but neither its unique minimizer nor necessarily the minimizer of the total yardstick: a portfolio that matched or exceeded the benchmark in every favorable week and stayed at or above the floor in every adverse week would reach zero if it were feasible. And benchmark attraction comes from reachability rather than global optimality, since the benchmark removes the favorable-state component without foresight and without leaving the investable universe, whereas the floor lies outside the universe and the market direction is unknown when the portfolio must act. Whether the benchmark minimizes the total yardstick over the feasible set is therefore an empirical question; among the portfolios tested in each mandate none records a lower total than the benchmark, so it records the lowest of the constructions considered, but this is an empirical finding, not a consequence of the proposition.

APPENDIX B

DATA AND REPRODUCIBILITY

The data are weekly total returns, with dividends reinvested, on fifteen US-listed exchange-traded funds, sampled at the weekly close over 2007-10-05 to 2026-06-19 (976 consecutive weekly observations, no missing weeks); source: Bloomberg total-return series. Exhibit B1 lists the funds, their asset classes, and the mandates in which each is used. Each mandate is a universe of six funds whose benchmark is constructible from its own funds: the S&P 500 fund for equity, the aggregate bond fund for fixed income, and a 60/40 blend of the two for multi-asset, so that the multi-asset benchmark equals 0.6 times the equity benchmark plus 0.4 times the fixed-income one. The preservation floor is a 0-1 year Treasury-bill fund (SHV, annualized volatility 0.32 percent), a Treasury-bill proxy held outside all six-fund universes. Every input is formed causally: weights use only a trailing rolling window of past returns, the yardstick uses the realized market direction read from the benchmark, and no quantity uses information beyond its date. The data vintage (extraction date) and the exact Bloomberg field identifiers are recorded with the replication package.

EXHIBIT B1

Exchange-traded funds and their asset classes

Ticker	Fund	Asset class	Mandate(s)
SPY	SPDR S&P 500 ETF	US large-cap equity	Multi-asset, Equity
QQQ	Invesco QQQ (Nasdaq-100)	US large-cap growth equity	Equity
IWM	iShares Russell 2000 ETF	US small-cap equity	Equity
EFA	iShares MSCI EAFE ETF	Developed ex-US equity	Equity
EEM	iShares MSCI Emerging Markets ETF	Emerging-market equity	Equity
VNQ	Vanguard Real Estate ETF	US real estate (REITs)	Multi-asset, Equity
AGG	iShares Core US Aggregate Bond ETF	US aggregate bonds	Multi-asset, Fixed income
LQD	iShares iBoxx IG Corporate Bond ETF	US investment-grade credit	Fixed income
HYG	iShares iBoxx High Yield Corporate Bond ETF	US high-yield credit	Multi-asset, Fixed income
TIP	iShares TIPS Bond ETF	US inflation-linked bonds	Fixed income
BWX	SPDR Bloomberg International Treasury Bond ETF	International sovereign bonds	Fixed income
SHY	iShares 1-3 Year Treasury Bond ETF	US short-term Treasuries	Fixed income
GLD	SPDR Gold Shares	Gold	Multi-asset
DBC	Invesco DB Commodity Index Tracking Fund	Broad commodities	Multi-asset
SHV	iShares 0-1 Year Treasury Bond ETF	US Treasury bills (floor)	Held outside all universes

NOTES: The fifteen US-listed exchange-traded funds and the mandate or mandates in which each is used. The preservation-floor proxy (SHV) is held outside all three universes.

APPENDIX C

THE CONSTRUCTIONS

All eight constructions are long-only and fully invested on the six funds of the mandate, formed from the rolling-window sample covariance Σ and mean μ . Equal weighting sets $w_i = 1/N$. Inverse-volatility weighting sets $w_i \propto 1/\sigma_i$. Minimum variance minimizes $w^\top \Sigma w$ over the simplex. Maximum Sharpe maximizes $(w^\top \mu - r_f)/\sqrt{w^\top \Sigma w}$, the tangency portfolio. Equal risk contribution equalizes $w_i(\Sigma w)_i$ across funds (Maillard, Roncalli, and Teiletche 2010). Maximum diversification maximizes the diversification ratio $(w^\top \sigma)/\sqrt{w^\top \Sigma w}$, with σ the vector of volatilities (Choueifaty and Coignard 2008). Hierarchical risk parity clusters the funds by the correlation distance $\sqrt{(1-\rho)/2}$, orders them by the resulting tree, and allocates by recursive inverse-variance bisection (López de Prado 2016). Hierarchical equal risk contribution clusters by the same distance and allocates down the dendrogram by inverse-variance splits (Raffinot 2018); we use one standard variant of the family. The preservation sleeve used in the second result minimizes the average shortfall below the floor, $\sum_t (r_f - R_{p,t})^+$, a linear program on the rolling window, and concentrates in the least risky funds the universe allows.

APPENDIX D

THE YARDSTICK AND THE MARKET-STATE LABEL

The yardstick is the total shortfall of the portfolio below the switched reference, counted only on the downside,

$$D = \sum_t (R_t^{\text{ref}} - R_{p,t})^+, \quad R_t^{\text{ref}} = \begin{cases} r_f & B_t < 0, \\ B_t & B_t \geq 0, \end{cases}$$

with the market-state label read ex post from the sign of the benchmark return: a week is adverse if the benchmark fell and favorable if it rose. We report the yardstick as its per-week average, annualized by multiplying by fifty-two and expressed in percentage points; it is a shortfall in return units and is not cardinally comparable to a Sharpe ratio. The label classifies each week by the realized direction of the benchmark rather than by a sustained market cycle; the robustness of the results to this choice is documented in Appendix J, where an alternative that dates adverse weeks by a trailing four-week benchmark return leaves the conclusions in place.

APPENDIX E

RANK INFERENCE

With eight constructions per mandate, the rank concordances of the first finding are effect sizes, and we report exact inference so their strength is visible. For each concordance we compute the exact permutation p-value, enumerating all orderings of the eight constructions under the null of random concordance, and a block-bootstrap 90 percent interval over 26-week blocks, which respects the serial dependence and crisis clustering of weekly returns. Exhibit E1 reports these for the net concordance between the Sharpe ratio and the yardstick and for its adverse-state component.

EXHIBIT E1

Rank concordance with the Sharpe ratio: p-values and bootstrap intervals

Mandate	Net τ	p (net)	90% Interval (net)	Spearman ρ	Adverse-state τ	p (adverse)	90% Interval (adverse)	τ (favorable)
Multi-asset	+0.14	0.72	[0.00, +0.71]	+0.19	-0.43	0.18	[-0.71, +0.14]	+0.36
Fixed income	-0.21	0.55	[-0.57, +0.21]	-0.40	-0.93	<0.001	[-0.93, -0.43]	+0.79
Equity	+0.79	0.006	[+0.07, +0.93]	+0.90	+0.29	0.40	[-0.21, +0.71]	+0.50

NOTES: Exact permutation p-values and block-bootstrap 90 percent intervals, over 26-week blocks, for the net concordance between the Sharpe ratio and the yardstick and for its adverse-state component. The primary specification is the yardstick with the ex-post weekly label.

APPENDIX F

ROBUSTNESS OF THE REORDERING

Exhibit F1 recomputes the net concordance between the Sharpe ratio and the yardstick under alternative specifications of the latter: excluding weeks with a benchmark move below 0.25 percent in magnitude (a neutral band), doubling the weight on the adverse or the favorable component, winsorizing the weekly shortfall at its 99th percentile, normalizing each week's shortfall by trailing benchmark volatility, and aggregating to monthly rather than weekly returns before scoring. The reordering is qualitatively unchanged: the concordance stays low or negative in the two mandates with a reachable defensive asset (the refuge mandates, for short) and high in equity across every specification, and its magnitude tracks the weight on the adverse component, strengthening under adverse up-weighting and weakening substantially under favorable up-weighting, which confirms the adverse half as the source.

EXHIBIT F1

Net concordance under alternative specifications of the yardstick

Mandate	Base	Neutral band	Adverse ×2	Favorable ×2	Winsorized	Vol-normalized	Monthly
Multi-asset	+0.14	+0.21	−0.50	+0.14	+0.07	+0.07	+0.21
Fixed income	−0.21	−0.07	−0.86	+0.36	−0.07	−0.07	−0.07
Equity	+0.79	+0.79	+0.64	+0.79	+0.79	+0.79	+0.93

NOTES: Kendall τ between the Sharpe ratio and the yardstick under six alternative specifications: a neutral band excluding weeks with a benchmark move below 0.25 percent in magnitude, doubling the weight on the adverse or the favorable component, winsorizing the weekly shortfall at its 99th percentile, normalizing by trailing benchmark volatility, and monthly rather than weekly aggregation.

EXHIBIT F2

Net concordance under alternative floor proxies

Mandate	Floor SHV (base)	Floor SHY (1-3y)	Floor zero
Multi-asset	+0.14	+0.00	+0.29
Fixed income	−0.21	−0.29	+0.14
Equity	+0.79	+0.79	+0.79

NOTES: Kendall τ between the Sharpe ratio and the yardstick with the floor proxied by SHV (base), a 1–3 year Treasury fund (SHY), and, as a limiting case, a zero floor.

EXHIBIT F3

Net concordance net of transaction costs

Mandate	0 bps	10 bps	20 bps
Multi-asset	+0.14	+0.14	+0.14
Fixed income	−0.21	−0.21	−0.14
Equity	+0.79	+0.86	+0.79

NOTES: Kendall τ between the Sharpe ratio and the yardstick net of one-way transaction costs of 0, 10, and 20 basis points.

APPENDIX G

THE WEDGE UNDER INFERENCE

Exhibit G1 reports block-bootstrap 90 percent intervals, over 26-week blocks, for the differences in the yardstick between the active portfolios and the benchmark. The total cost of active management, the dual-anchor's excess over the benchmark, is above zero in both refuge mandates. The cost of competing, the dual-anchor's excess over the defensive tracker, is above zero in both; its favorable-state component, 11.6 and 4.7 points, exceeds the total, while the adverse-state component falls. The cost of de-risking, the tracker's excess over the benchmark, is above zero in multi-asset but not distinguishable from zero in fixed income.

EXHIBIT G1

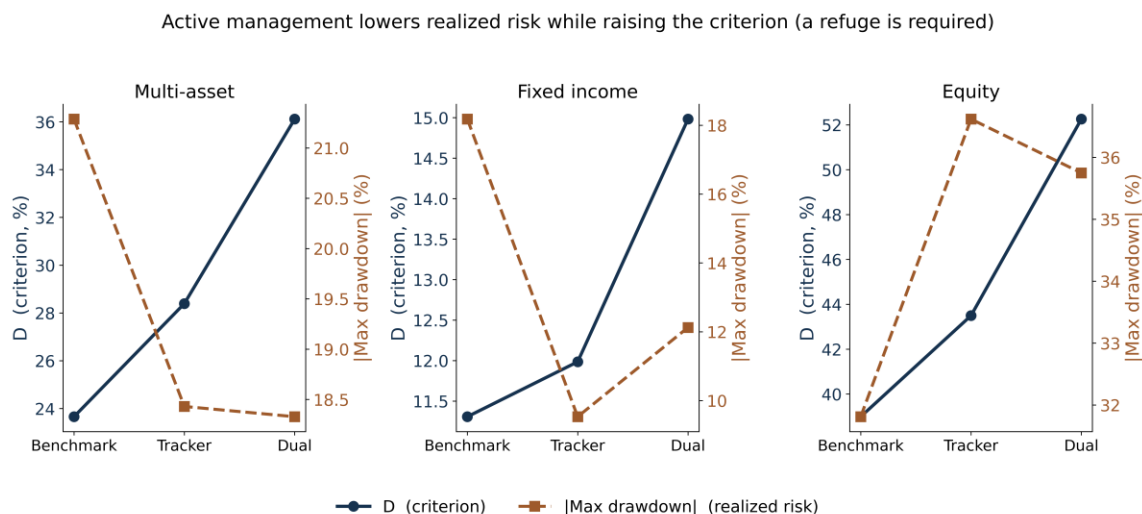
Bootstrap intervals for differences in the yardstick

Difference in D	Multi-asset	Fixed income
Total cost of active management (dual minus bench)	+12.5 [+9.8, +16.4]	+3.7 [+2.0, +5.4]
Cost of de-risking (tracker minus bench)	+4.7 [+3.4, +7.0]	+0.7 [−0.2, +1.6]
Cost of competing (dual minus tracker)	+7.7 [+5.2, +10.6]	+3.0 [+1.8, +4.2]
of which favorable-state component	+11.6 [+9.4, +14.3]	+4.7 [+3.7, +5.5]

NOTES: Block-bootstrap 90 percent intervals, over 26-week blocks, for the differences in the yardstick between the active portfolios and the benchmark, in annualized percentage points.

EXHIBIT G2

Active management can lower realized maximum drawdown while raising the yardstick



NOTES: For the static benchmark, the defensive tracker, and the dual-anchor, the yardstick (left axis) and the magnitude of the maximum drawdown (right axis), in each mandate.

APPENDIX H

IDENTIFICATION AND FALSIFICATION

Each result is stated so that a pattern in the data could contradict it. Exhibit H1 sets out, for each claim, what would falsify it and what the evidence shows.

EXHIBIT H1

Falsification conditions for the paper's claims

Claim	What would falsify it	Evidence
1. The yardstick orders constructions differently from a conventional measure	The yardstick and a conventional measure order the constructions alike across all mandates The favorable-state component shows the same negative relation to the Sharpe ranking as the adverse one	Low or negative concordance in fixed income and multi-asset; concordance appears only in the equity boundary case (Appendix E)
2. The divergence is driven by the adverse half	The yardstick coincides with a fixed-reference downside measure in both refuge mandates	The adverse component is negatively related to the Sharpe ranking, significantly so in fixed income; the favorable component is positively related (Appendix E)
3. The divergence is a property of the switch	The equity mandate shows the same reordering as the refuge mandates	The yardstick is distinct from the floor-referenced measures in both refuge mandates (Exhibit 5; Appendix K)
4. The refuge condition is a real boundary	A defensive portfolio lowers both the drawdown and the yardstick	The equity concordance is high and significant, so the yardstick does not reorder without a reachable refuge, and adding one to the equity universe produces the down-market inversion (Appendix E; Appendix I)
5. Active protection can raise the yardstick	A feasible active portfolio consistently records a lower yardstick than the benchmark	The tracker and dual-anchor lower the drawdown while raising the yardstick, with the difference above zero (Exhibit 6; Appendix G)
6. The benchmark records the lowest yardstick among the tested in-mandate portfolios		The benchmark records the lowest yardstick among the tested portfolios, consistent with the favorable-state implication of Proposition 1 (Appendix A; Exhibit 6)

NOTES: For each claim, the condition that would falsify it and the corresponding evidence.

APPENDIX I

REFUGE ACCESSIBILITY AS A CROSS-SECTIONAL DIAGNOSTIC

The boundary condition of the first finding, that the yardstick can reorder the constructions where the mandate holds a reachable refuge, can be turned from a description into a measurable diagnostic. We define a pre-specified diagnostic index, the residual shortfall below the floor that the best in-universe defensive portfolio still incurs in adverse weeks, namely the adverse-state yardstick of the preservation sleeve. A low value means the universe contains assets that hold up when the benchmark falls, so a defensive portfolio can stay near the floor; a high value means it does not, and even the most defensive portfolio falls with the rest. The index uses the full sample only to characterize the cross-sectional boundary across variant universes; the portfolio weights themselves remain causal, walk-forward. Because the variant universes are built from the same sample, we treat this as a cross-sectional diagnostic of the boundary condition rather than an independent out-of-sample test. The exercise is intended to reveal monotonicity in the boundary condition, not to estimate a structural causal effect of refuge accessibility.

We build variant universes of six funds from the same data, adding or removing a defensive asset, and measure across them the adverse-component concordance between the Sharpe ranking

and the yardstick. Exhibit I1 and Exhibit I2 report the result. The adverse-state concordance is strongly associated with the refuge index (Spearman $\rho = +0.89$): as the residual shortfall of the best in-universe defense rises, the adverse-state concordance becomes less negative. Where the best in-universe defense can stay near the floor in adverse weeks, with a residual of a few percent, the adverse component strongly inverts the Sharpe ranking; where it cannot, with a residual of twenty percent or more, the inversion weakens and disappears.

The mechanism is visible in the variant that adds a single defensive line to an all-equity universe. Adding short Treasuries or aggregate bonds, assets that hold up when equities fall, creates the down-market inversion: the adverse-component concordance moves from +0.29 with no refuge to -0.79 and -0.57. Adding gold does not, because gold does not reliably beat the floor in the weeks when equities fall; its residual is high and the inversion does not appear. A defensive asset in name is not sufficient; the inversion requires a refuge that is reachable in the states that matter. The net concordance does not move in step: it stays high with short Treasuries (+0.93) and turns negative with gold (-0.29), so the diagnostic concerns the adverse component, where the switch to the floor operates, and not the net ranking.

EXHIBIT I1

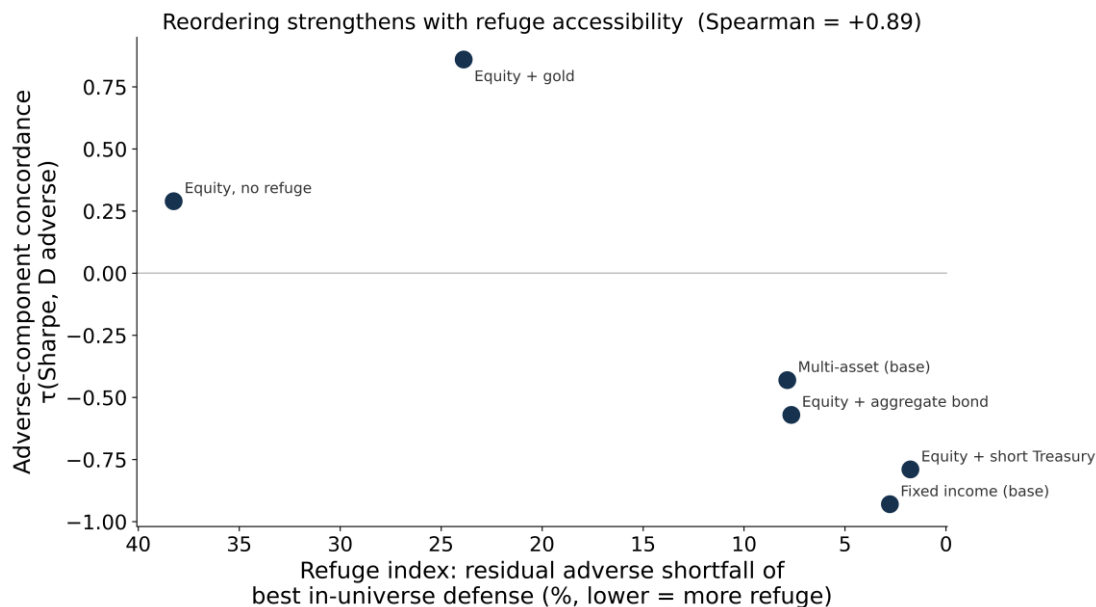
Refuge accessibility and the reordering across variant universes

Universe	Refuge index (%)	$\tau(\text{Sharpe, D adverse})$	$\tau(\text{Sharpe, D net})$
Equity + short Treasury	1.8	-0.79	+0.93
Fixed income (base)	2.8	-0.93	-0.21
Equity + aggregate bond	7.7	-0.57	+0.50
Multi-asset (base)	7.9	-0.43	+0.14
Equity + gold	23.9	+0.86	-0.29
Equity, no refuge	38.3	+0.29	+0.79

NOTES: Universes ordered by the refuge index, defined as the annualized adverse-state yardstick of the preservation sleeve, that is, the residual shortfall below the floor that the best in-universe defense still incurs in adverse weeks (a low value indicates a rich refuge). Both the net concordance and the adverse-component concordance with the Sharpe ranking are reported.

EXHIBIT I2

The reordering strengthens with refuge accessibility



NOTES: The adverse-component concordance $\tau(\text{Sharpe}, D \text{ adverse})$ against the refuge index across variant universes.

APPENDIX J

ALTERNATIVE DEFINITIONS OF THE ADVERSE STATE

The market direction the yardstick uses is the sign of the benchmark's weekly return. This is the simplest definition, and a natural question is whether the result depends on it or on a more substantive notion of an adverse market. We recompute the reordering under five pre-specified definitions of the adverse state: the weekly sign; a volatility-scaled band that classifies a week adverse only when the standardized benchmark return falls below -0.5 and excludes near-flat weeks; the sign of the trailing four-week and of the trailing twelve-week cumulative benchmark return; and a drawdown rule that calls a week adverse when the benchmark is more than five percent below its trailing peak, a persistent rather than a weekly notion. The share of weeks classified adverse ranges from 41 percent under the weekly sign to 21 percent under the drawdown rule, so these are genuinely different definitions.

Exhibit J1 reports the adverse-component concordance under each. The reordering is stable: the adverse component of the yardstick inverts the Sharpe ranking in the two mandates with a refuge under every definition, with a concordance between -0.36 and -0.93 , and remains positive in the equity boundary case, between $+0.14$ and $+0.57$. The net concordance is similarly stable, low in the refuge mandates, around $+0.1$ in multi-asset and -0.2 in fixed income, and high in equity, around $+0.8$. The result therefore rests on the constructions' behavior in adverse markets, however the adverse market is delimited, and not on the weekly-sign convention; in particular it holds under the drawdown definition, a substantive notion of a sustained decline.

EXHIBIT J1

Adverse-component concordance under alternative definitions of the market state

Adverse-state definition	Share adverse	Multi-asset	Fixed income	Equity
Weekly sign	41%	−0.43	−0.93	+0.29
Volatility-scaled band	24%	−0.43	−0.93	+0.14
Trailing 4-week return	34%	−0.43	−0.93	+0.29
Trailing 12-week return	26%	−0.43	−0.93	+0.43
Drawdown above 5% from peak	21%	−0.36	−0.79	+0.57

NOTES: The adverse-component concordance $\tau(\text{Sharpe}, D \text{ adverse})$ under five pre-specified definitions of the adverse market state. The share of weeks classified adverse is the average across mandates.

APPENDIX K

THE INFORMATION THE YARDSTICK ADDS

To locate what the yardstick adds, rather than to argue that it dominates, we compare its ranking of the eight constructions with the rankings from a broader set of risk and downside measures: the Sharpe and Sortino ratios, Omega, downside beta, the Calmar ratio, semivariance, the five-percent expected shortfall, and an asymmetric tracking error, the deviation of returns below the benchmark. Exhibit K1 reports the rank concordance between the yardstick and each. In the two mandates with a refuge, none of the eight comparator measures reproduces the yardstick's ranking: the concordances are low, and the yardstick orders the constructions differently from every conventional risk or downside measure. Its nearest relative is the asymmetric tracking error, the root-mean-square deviation below the benchmark, with a concordance of +0.79 in multi-asset and +0.50 in fixed income; unlike the mean below-benchmark shortfall of Exhibit 5, it weights larger deficits more heavily, but it too shares only the yardstick's favorable-state half and lacks the switch to the floor in adverse states. The equity boundary case again shows relatively high concordance across the comparison measures, consistent with the absence of a defensive refuge. The yardstick therefore carries information that none of the eight measures considered does: it is closest to a downside, benchmark-relative measure, and the switch to the preservation floor in adverse states is the component that sets it apart.

EXHIBIT K1

The yardstick versus eight risk and downside measures

Comparator	Multi-asset	Fixed income	Equity
Sharpe ratio	+0.14	−0.21	+0.79
Sortino ratio	+0.14	−0.21	+0.71
Omega	+0.21	−0.21	+0.79
Downside beta	−0.64	+0.07	+0.64
Calmar ratio	+0.21	−0.21	+0.86
Semivariance	−0.29	+0.36	+0.86
Expected shortfall (5%)	−0.21	+0.36	+0.71
Asymmetric tracking error	+0.79	+0.50	+0.64

Notes: Rank concordance (Kendall τ) between the yardstick and eight risk and downside measures, by mandate, computed so that higher values indicate closer agreement.

APPENDIX L

ROBUSTNESS TO THE CONSTRUCTION SET

The rank comparisons in the body use eight constructions, so the concordance between the yardstick and the Sharpe ratio is read as an effect size rather than a test. To check that the reordering is not an artifact of that small set, we repeat the analysis on a curated set of thirty constructions spanning eleven families of the portfolio-construction literature: naive weighting (equal weight, inverse volatility, inverse variance, inverse CVaR, inverse maximum drawdown), mean–variance (minimum variance, maximum Sharpe, mean–variance utility), risk parity (equal risk contribution, downside risk parity, CVaR risk parity), diversification (maximum diversification, maximum decorrelation), hierarchical (hierarchical risk parity, hierarchical equal risk contribution, nested clustered optimization), downside and tail risk (minimum semivariance, maximum Sortino, minimum CVaR, maximum return-to-CVaR, minimum mean-absolute-deviation), drawdown (minimum maximum-drawdown, minimum ulcer index, maximum Calmar), dispersion (minimum Gini mean difference, worst-case realization), distribution (maximum Omega, maximum skewness), growth (Kelly), and structural (principal-component risk parity). The eight constructions of the body are included; the remaining twenty-two are chosen one or two per family to add genuinely distinct objectives (the drawdown-, distribution-, and growth-based methods the eight omit) rather than tuning variants of a single objective, which would strengthen the concordance only through pseudo-replication. Each construction is formed on the same rolling one-year window, rebalancing, preservation floor, and out-of-sample period as in the body, and held walk-forward.

Exhibit L1 (upper panel) reports the net and adverse-component concordance over the full set. In both mandates with a reachable defensive asset the adverse component inverts the Sharpe ranking with high significance ($\tau = -0.55$ and -0.70 , permutation $p < 0.001$), which sharpens the first finding: the multi-asset adverse component, not distinguishable from zero with eight constructions, is significant here. The equity boundary component remains positive ($\tau = +0.34$). The net concordance stays near zero in multi-asset and is negative but marginal in fixed income ($\tau = -0.23$, $p = 0.08$).

A larger set raises significance only if its members are genuinely distinct; near-duplicate constructions would otherwise overstate the effective cross-section. Across the two mandates with a defensive asset the thirty constructions form fourteen effective clusters, with a mean pairwise correlation of their out-of-sample return paths of 0.72 and 0.76 at a 0.95 correlation cut, so the set contains substantial cross-sectional diversity; the cluster-representative exercise below addresses the residual redundancy directly. In the equity mandate the thirty collapse to seven clusters (mean correlation 0.92).

As a conservative check, we retain one representative, the cluster medoid, per near-duplicate cluster and recompute the concordance on that reduced, fully distinct set (Exhibit L1, lower panel). The adverse-component inversion survives in both mandates with a defensive asset: on fourteen distinct methods each, $\tau = -0.41$ (permutation $p = 0.047$) in multi-asset and $\tau = -0.58$ ($p = 0.004$) in fixed income. On the distinct set the net concordance in fixed income is also

significantly negative ($\tau = -0.47$, $p = 0.021$, 90 percent interval $[-0.63, -0.05]$). The reordering therefore does not appear to be driven by the number or redundancy of the constructions.

The equity mandate reinforces the defensive-asset boundary from the construction side. Where the universe holds no reachable defensive asset, the constructions themselves converge: the thirty methods collapse to seven effective clusters, against fourteen where a defensive asset is present, so the constructions differ too little for the yardstick to reorder them. The boundary is thus visible in the construction space, not only in the concordance: where preservation cannot separate the constructions, neither can the yardstick.

EXHIBIT L1

Concordance over the curated construction set and its cluster representatives

Mandate	n	Net τ	p (net)	Adverse τ	p (adverse)
Full curated set					
Multi-asset	30	+0.08	0.55	-0.55	<0.001
Fixed income	30	-0.23	0.08	-0.70	<0.001
Equity	30	+0.40	0.002	+0.34	0.007
Cluster representatives (one method per near-duplicate cluster)					
Multi-asset	14	+0.06	0.83	-0.41	0.047
Fixed income	14	-0.47	0.021	-0.58	0.004
Equity	7	+0.24	0.56	+0.14	0.77

NOTES: Rank concordance between the Sharpe ratio and the yardstick, net and by adverse component (Kendall τ between the Sharpe ratio and the negative of D and of its adverse-state component), over the curated thirty-construction set (upper panel) and over its cluster representatives (lower panel: one method per near-duplicate cluster, out-of-sample return correlation above 0.95). n is the number of constructions. p-values are two-sided permutation p-values (exact where feasible, otherwise Monte Carlo with 40,000 draws). A negative adverse-component τ means high-Sharpe constructions score worse on the yardstick in adverse markets. The static benchmark is excluded.

DISCLOSURE

The author reports no potential conflict of interest. The opinions and analyses presented are the author's own and do not necessarily reflect those of Banco de España or the Eurosystem. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

- Ang, A., and G. Bekaert. 2002. "International Asset Allocation with Regime Shifts." *The Review of Financial Studies* 15 (4): 1137–1187.
- . 2004. "How Regimes Affect Asset Allocation." *Financial Analysts Journal* 60 (2): 86–99.
- Bailey, D. H., and M. López de Prado. 2012. "The Sharpe Ratio Efficient Frontier." *Journal of Risk* 15 (2): 3–44.
- Bajo Traver, M. 2026a. "Harvesting Term and Credit Risk Premia: A Regime-Based Fixed Income Allocation Framework." Forthcoming in the *Journal of Fixed Income*.
- . 2026b. "Regret-Based Portfolio Allocation: Minimizing Opportunity Cost and Its Dispersion." Working paper, SSRN.
- Barberis, N., M. Huang, and T. Santos. 2001. "Prospect Theory and Asset Prices." *The Quarterly Journal of Economics* 116 (1): 1–53.
- Barberis, N., and W. Xiong. 2009. "What Drives the Disposition Effect? An Analysis of a Long-Standing Preference-Based Explanation." *The Journal of Finance* 64 (2): 751–784.
- Basak, S., and A. Pavlova. 2013. "Asset Prices and Institutional Investors." *American Economic Review* 103 (5): 1728–1758.
- Bawa, V. S. 1975. "Optimal Rules for Ordering Uncertain Prospects." *Journal of Financial Economics* 2 (1): 95–121.
- Bell, D. E. 1982. "Regret in Decision Making under Uncertainty." *Operations Research* 30 (5): 961–981.
- Bouyé, E., and J. Teïletche. 2025. "Regime-Based Strategic Asset Allocation." *Financial Analysts Journal* 81 (4): 84–102.
- Chevalier, J., and G. Ellison. 1997. "Risk Taking by Mutual Funds as a Response to Incentives." *Journal of Political Economy* 105 (6): 1167–1200.
- Choueifaty, Y., and Y. Coignard. 2008. "Toward Maximum Diversification." *The Journal of Portfolio Management* 35 (1): 40–51.
- Cremers, K. J. M., and A. Petajisto. 2009. "How Active Is Your Fund Manager? A New Measure That Predicts Performance." *The Review of Financial Studies* 22 (9): 3329–3365.
- Das, S., H. Markowitz, J. Scheid, and M. Statman. 2010. "Portfolio Optimization with Mental Accounts." *Journal of Financial and Quantitative Analysis* 45 (2): 311–334.
- DeMiguel, V., L. Garlappi, and R. Uppal. 2009. "Optimal Versus Naive Diversification: How Inefficient Is the 1/N Portfolio Strategy?" *The Review of Financial Studies* 22 (5): 1915–1953.

- Fishburn, P. C. 1977. "Mean-Risk Analysis with Risk Associated with Below-Target Returns." *American Economic Review* 67 (2): 116–126.
- Gomes, F. 2005. "Portfolio Choice and Trading Volume with Loss-Averse Investors." *The Journal of Business* 78 (2): 675–706.
- Guidolin, M., and A. Timmermann. 2007. "Asset Allocation under Multivariate Regime Switching." *Journal of Economic Dynamics and Control* 31 (11): 3503–3544.
- Hamilton, J. D. 1989. "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle." *Econometrica* 57 (2): 357–384.
- He, X. D., and X. Y. Zhou. 2011. "Portfolio Choice under Cumulative Prospect Theory: An Analytical Treatment." *Management Science* 57 (2): 315–331.
- Jorion, P. 2003. "Portfolio Optimization with Tracking-Error Constraints." *Financial Analysts Journal* 59 (5): 70–82.
- Kahneman, D., and A. Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica* 47 (2): 263–291.
- Konno, H., and H. Yamazaki. 1991. "Mean-Absolute Deviation Portfolio Optimization Model and Its Applications to Tokyo Stock Market." *Management Science* 37 (5): 519–531.
- Kőszegi, B., and M. Rabin. 2006. "A Model of Reference-Dependent Preferences." *The Quarterly Journal of Economics* 121 (4): 1133–1165.
- . 2007. "Reference-Dependent Risk Attitudes." *American Economic Review* 97 (4): 1047–1073.
- Kritzman, M., S. Page, and D. Turkington. 2012. "Regime Shifts: Implications for Dynamic Strategies." *Financial Analysts Journal* 68 (3): 22–39.
- Loomes, G., and R. Sugden. 1982. "Regret Theory: An Alternative Theory of Rational Choice under Uncertainty." *The Economic Journal* 92 (368): 805–824.
- López de Prado, M. 2016. "Building Diversified Portfolios that Outperform Out of Sample." *The Journal of Portfolio Management* 42 (4): 59–69.
- Maillard, S., T. Roncalli, and J. Teïletche. 2010. "The Properties of Equally-Weighted Risk Contributions Portfolios." *The Journal of Portfolio Management* 36 (4): 60–70.
- Markowitz, H. 1952. "Portfolio Selection." *The Journal of Finance* 7 (1): 77–91.
- Raffinot, T. 2018. "Hierarchical Clustering Based Asset Allocation." *The Journal of Portfolio Management* 44 (2): 89–99.
- Rockafellar, R. T., and S. Uryasev. 2000. "Optimization of Conditional Value-at-Risk." *Journal of Risk* 2 (3): 21–41.
- Roll, R. 1992. "A Mean/Variance Analysis of Tracking Error." *The Journal of Portfolio Management* 18 (4): 13–22.
- Roy, A. D. 1952. "Safety First and the Holding of Assets." *Econometrica* 20 (3): 431–449.
- Sharpe, W. F. 1994. "The Sharpe Ratio." *The Journal of Portfolio Management* 21 (1): 49–58.
- Shefrin, H., and M. Statman. 2000. "Behavioral Portfolio Theory." *Journal of Financial and Quantitative Analysis* 35 (2): 127–151.

- Telser, L. G. 1955. "Safety First and Hedging." *The Review of Economic Studies* 23 (1): 1–16.
- Tversky, A., and D. Kahneman. 1992. "Advances in Prospect Theory: Cumulative Representation of Uncertainty." *Journal of Risk and Uncertainty* 5 (4): 297–323.
- van Bilsen, S., R. J. A. Laeven, and T. E. Nijman. 2020. "Consumption and Portfolio Choice under Loss Aversion and Endogenous Updating of the Reference Level." *Management Science* 66 (9): 3927–3955.
- Viole, F., and D. Nawrocki. 2016. "Predicting Risk/Return Performance Using Upper Partial Moment/Lower Partial Moment Metrics." *Journal of Mathematical Finance* 6 (5): 900–920.