

Assessment of Investment Policy Statement Quality Across U.S. Public Pension Funds

A Comparative Analysis of IPS Documentation Standards

Analysis Date: October 2026

Rubric Version: 1.3

1. Introduction

An Investment Policy Statement (IPS) is the governing document of an institutional investor. It defines the entity and the assets it covers, states the objectives the portfolio is meant to serve, sets the boundaries of risk, prescribes how the portfolio is built and kept in line, allocates decision authority, and fixes the cadence of oversight. For a public pension plan, which manages retirement assets under statutory fiduciary duties and answers to a legislature, a board, members and taxpayers at once, the IPS is the principal artefact by which governance is documented and can be defended.

This study examines the quality of IPS documentation across 121 U.S. public pension plans: statewide systems, teachers' and school-employee systems, county and municipal plans, and police and fire funds. Each document is scored on 14 domains of policy practice, grouped into six categories, and within every domain on four quality lenses. The scores measure the document, not the fund: they say how specific, enforceable and internally coherent the written policy is, and nothing about investment skill or returns.

Research Objectives

1. Measure the overall quality of IPS documentation across a broad sample of public pension plans
2. Locate the systematic strengths and weaknesses, by category and by quality lens
3. Show what separates the strongest documents from the weakest
4. Give plans an ordered set of improvements that the evidence supports
5. Describe how the policies set and organise the strategic asset allocation
6. Show what changes when a plan rewrites its policy, and what the newest and strongest documents suggest best practice should add

2. Methodology

2.1 Assessment Framework

The rubric (version 1.3) evaluates each IPS on 14 domains, grouped into six categories for comparison:

1. Governance & Oversight (2 domains)

- Governance & Decision-Making
- Monitoring, Reporting & Review

2. Objectives Clarity (3 domains)

- Scope, Purpose & Coverage
- Investment Objectives
- Investment Philosophy (*optional*)

3. Risk Framework (3 domains)

- Risk Tolerance & Risk Philosophy
- Time Horizon & Constraints
- Liquidity & Cash Flow Management

4. Portfolio Design (3 domains)

- Strategic Asset Allocation
- Rebalancing Policy
- Spending / Distribution Policy

5. Implementation and Oversight (2 domains)

- Implementation & Manager Oversight
- Performance Measurement & Benchmarks

6. Responsible Investing/ESG (1 domain - optional; scored, but outside the headline number)

- Responsible Investment / ESG

2.2 Scoring System

Each domain is scored on a five-point scale against written anchors:

- **5 (Best practice)**: fully specified, enforceable, aligned with peer practice
- **4 (Strong)**: clear and operational, with minor gaps
- **3 (Adequate)**: the basic structure is there, with notable gaps
- **2 (Weak)**: present but vague or non-operational
- **1 (Deficient)**: missing, or purely aspirational

Within each domain the judge scores four **quality lenses** - presence and clarity, quantification and enforceability, consistency and structure, and governance and process - and the domain score is the mean of the lenses that apply. A category score in this report is the mean of its domain scores.

The headline **overall score** of a plan is the figure the judge itself reports: the mean of its **core** domains. A domain is core when the rubric marks it mandatory, or marks it conditional and the condition applies. For a public pension plan that is 12 of the rubric's 15 domains. Investment Philosophy and Responsible Investment / ESG are optional: both are scored and shown throughout this report but excluded from the headline, and tax considerations is conditional on an institution type that does not apply here, so it is not scored at all. A **critical gap** is a core domain scored at 2 or below; an optional domain never raises one.

2.3 Scoring Procedure: The LLM Judge

Scores are produced by a large language model (Anthropic's Claude) acting as a structured **judge** against the rubric, not by manual review. The procedure is two-stage so that it is auditable and reproducible:

1. **Extraction (retrieval)**. For each domain, a Claude model reads the whole IPS and extracts the relevant facts with verbatim evidence quotes and page references. It does not score. A faster, cheaper model (Claude Haiku) is used because the task is retrieval. The output is a frozen extraction that records the model, the rubric hash and a hash of the source text.
2. **Judging (evaluation)**. A stronger model (Claude Sonnet) scores each domain against its anchors, lens by lens, from the frozen extraction and the source document. Scoring can therefore be re-run without re-extracting. The lens design is what lets the assessment separate "the policy is present" from "the policy is enforceable", the distinction at the centre of this report.

Several conventions keep the scores stable and traceable:

- **Deterministic**: the model runs at temperature zero, so the same input yields the same score.
- **Structured output**: the judge returns its verdict through a forced tool-use schema with enum-constrained reason codes, so every result is well-formed and tied to a stated reason.
- **Evidence-checked**: every quote the judge cites is verified against the source text. The result is reported as a diagnostic; it does not gate the score.

- **Provenance:** each scorecard records the judge model, rubric version and prompt version, so the paper's claims are frozen to an exact configuration and can be regenerated.

Validation against an expert (designed, not yet run). Because the scores are model-produced, their credibility rests on measured agreement with a human expert, and that measurement has not yet been made: the findings below are the judge's alone. The validation is built and ready. A stratified sample of plans has been drawn, to be rated blind, per domain, against the full IPS text, with the judge's score hidden until each human rating is submitted. Agreement will be reported as a quadratic-weighted Cohen's kappa alongside per-domain error, bias and a test-retest ceiling. The sample and ratings are pinned to a judge configuration, so they will double as a regression benchmark whenever the prompt or model changes. The scope is measure and diagnose; the judge is not tuned to the ratings.

2.4 Sample Composition

The corpus holds the current published IPS of 122 plans, drawn from the two hundred U.S. public pension plans with the most assets and covering 49 states plus one federal district. 121 are scored here. The remaining one is tracked but withheld from scoring because the document its sponsor publishes in the place of an IPS is a different instrument - such as a statement of investment beliefs or a set of investment guidelines - and a score derived from it would describe the wrong document. Each of the scored plans is counted once: where one policy governs several plans, the corpus keeps one document. By plan type:

Plan type	Plans
Statewide systems	49
Municipal systems	20
Police and fire systems	16
Teachers' systems	15
County systems	14
School-employee systems	5
University systems	2

2.5 Reading the Numbers

Every statistic is given to two decimals in the tables. In the running text the same figures are rounded to one decimal, so a sentence reads as a summary and the table beside it as the record. Scores describe the written policy. A low score means the document does not commit the plan to something specific and testable; it does not mean the plan behaves badly, and a high score is no evidence of investment skill.

3. Summary of Results

3.1 Overall Performance

The mean overall score across the 121 plans is 3.7 on the five-point scale and the median 3.7: most documents sit between the Adequate and Strong anchors. The spread is wide. About 2.5 points separate the strongest document from the weakest, more than the distance between an Adequate and a Best-practice rating (Table 1).

Table 1: Overall score, all plans

Statistic	Value
Plans scored	121
Mean	3.66
Median	3.71
Minimum	1.99
Maximum	4.50
Standard deviation	0.42

One plan scores below the Weak anchor and is kept in every statistic.

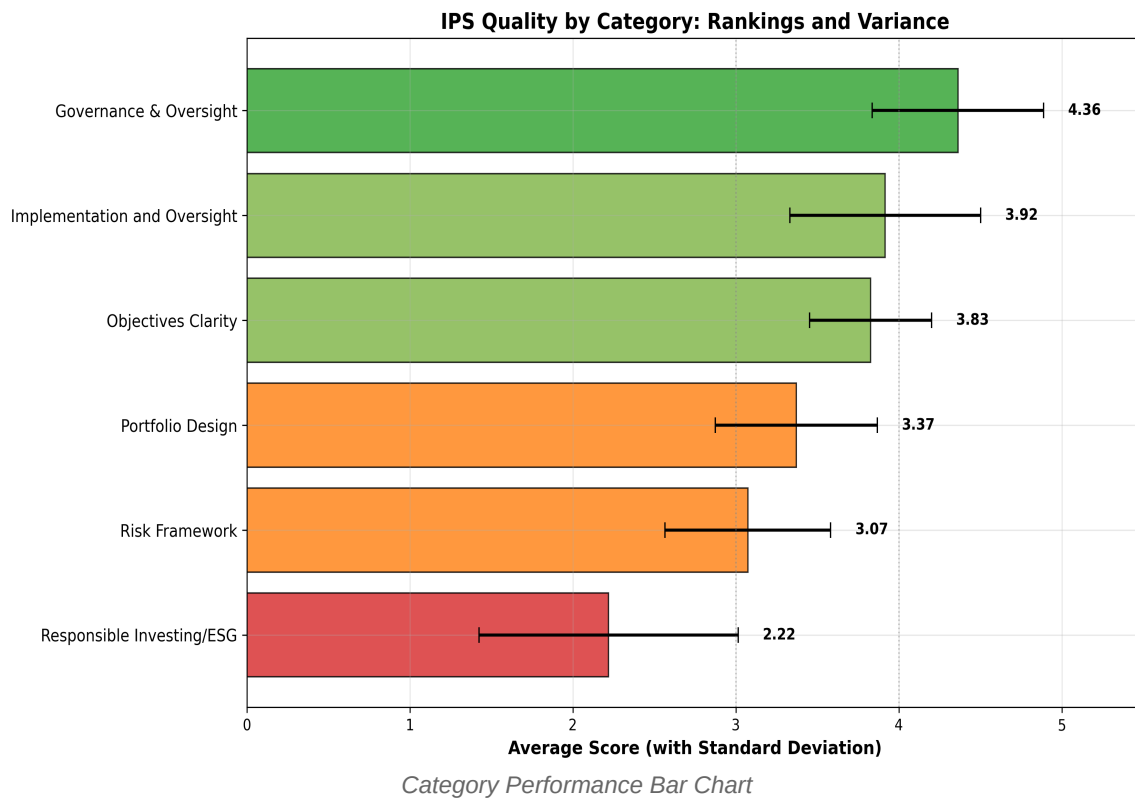
3.2 Category Performance Overview

The six categories fall into a clear order (Table 2). Governance & Oversight leads at 4.4; Implementation and Oversight and Objectives Clarity follow in the high threes; Portfolio Design and Risk Framework sit near the Adequate anchor; Responsible Investing/ESG trails at 2.2, the only category whose median sits at the Weak anchor.

Table 2: Category scores, all plans

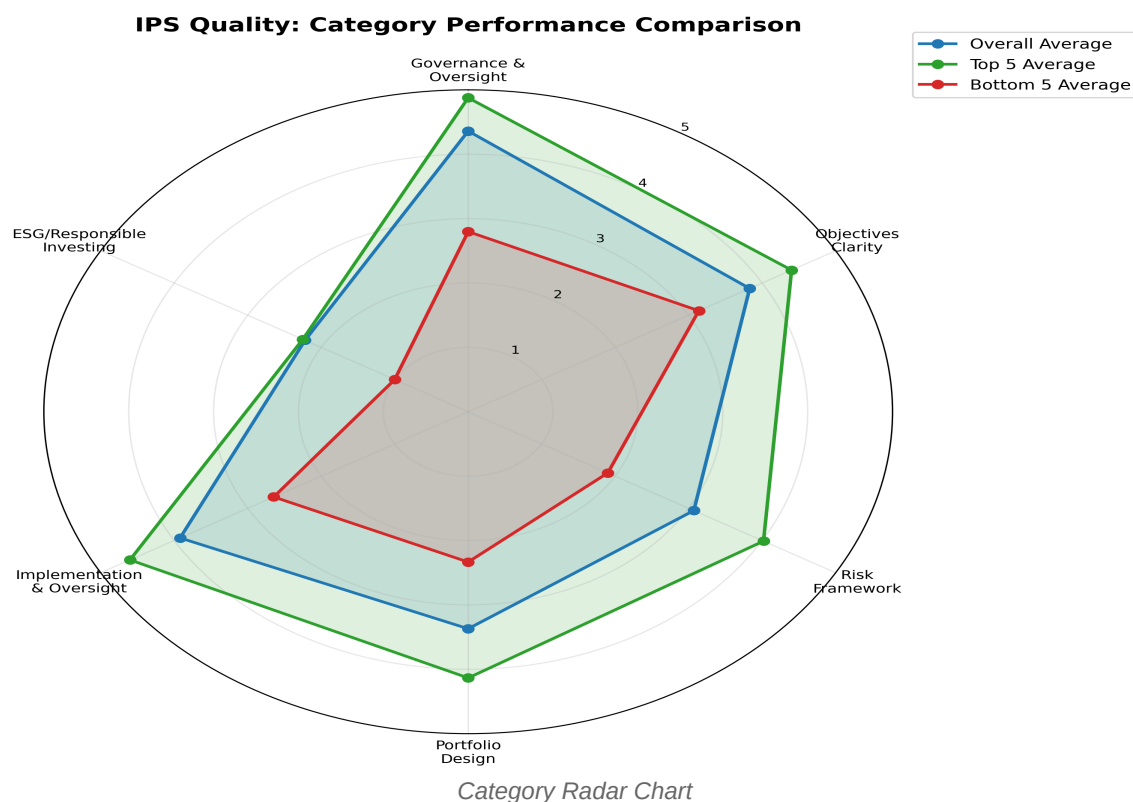
Category	Mean	Median	Min	Max	Std dev	Plans at 4.0 or above	Plans at 2.0 or below
Governance & Oversight	4.36	4.50	2.50	5.00	0.53	100 (83%)	0 (0%)
Implementation and Oversight	3.92	4.00	1.75	4.88	0.58	64 (53%)	1 (1%)
Objectives Clarity	3.83	3.83	2.69	4.67	0.37	43 (36%)	0 (0%)
Portfolio Design	3.37	3.42	1.25	4.42	0.50	11 (9%)	2 (2%)
Risk Framework	3.07	3.08	1.50	4.50	0.51	4 (3%)	4 (3%)
Responsible Investing/ESG	2.22	2.00	1.00	4.75	0.80	4 (3%)	61 (51%)

Figure 1: Category Performance Rankings



The bar chart ranks the categories with standard-deviation error bars. The strengths (Governance & Oversight, Implementation and Oversight) and the weaknesses (Risk Framework, Responsible Investing/ESG) are the same in every cut of the data that follows.

Figure 2: Category Performance Radar Chart



The radar chart overlays the sample mean, the five strongest documents and the five weakest across the six categories. The two outlines have the same shape: the weak documents are weak everywhere, not in one place.

3.3 The Four Quality Lenses

The same scores, decomposed by the lens the judge applied within every domain, locate the weakness precisely (Table 3). Consistency / Structure is the strongest lens at 4.0: documents are, on average, internally consistent and organised. Quantification / Enforceability is the weakest at 3.0, a full point lower. Plans describe their policies; far fewer state them in terms specific enough to bind a decision - a numeric limit, a named body that monitors it, a consequence when it is breached. This is the central finding of the assessment, and every weakness reported below is a form of it.

Table 3: Mean score by quality lens, all applicable domain-lenses

Quality lens	Mean	Domain-lenses scored
Consistency / Structure	4.04	1679
Presence / Clarity	3.66	1693
Governance / Process	3.51	1682
Quantification / Enforceability	2.99	1615

3.4 Results by Domain

A category score averages domains that behave very differently, so Table 3b gives all 14, weakest first. Governance & Decision-Making and Scope, Purpose & Coverage lead, at 4.5 and 4.4, and in 6 of the 14 domains most plans reach the Strong anchor. At the other end, Spending / Distribution Policy (2.2) and Liquidity & Cash Flow Management (2.7) are the weakest core domains by a wide margin. 57 plans score 2 or below in the first and 22 in the second, and these two account for most of the critical gaps in Section 8. Each is averaged in with stronger domains in its category (Portfolio Design and Risk Framework), which is why the category tables understate them.

Table 3b: Domain scores, all plans, weakest first

Domain	Core	Mean	Median	Plans at 4 or above	Plans at 2 or below
Responsible Investment / ESG	optional	2.22	2.00	4 (3%)	61 (51%)
Spending / Distribution Policy	yes	2.23	2.25	1 (1%)	57 (47%)
Liquidity & Cash Flow Management	yes	2.71	2.75	7 (6%)	22 (18%)
Time Horizon & Constraints	yes	3.10	3.25	14 (12%)	5 (4%)
Risk Tolerance & Risk Philosophy	yes	3.41	3.50	24 (20%)	1 (1%)
Investment Philosophy	optional	3.47	3.50	35 (29%)	1 (1%)
Rebalancing Policy	yes	3.56	3.75	46 (38%)	6 (5%)
Investment Objectives	yes	3.58	3.50	37 (31%)	1 (1%)
Performance Measurement & Benchmarks	yes	3.85	4.00	69 (57%)	3 (2%)
Implementation & Manager Oversight	yes	3.99	4.00	68 (56%)	1 (1%)
Monitoring, Reporting & Review	yes	4.24	4.50	93 (77%)	3 (2%)
Strategic Asset Allocation	yes	4.32	4.50	97 (80%)	1 (1%)
Scope, Purpose & Coverage	yes	4.42	4.50	108 (89%)	0 (0%)
Governance & Decision-Making	yes	4.48	4.75	103 (85%)	0 (0%)

4. Distribution Analysis

4.1 Score Distribution by Category

Table 4 counts the plans at each score level, category scores rounded to the nearest anchor.

Table 4: Plans by rounded category score

Category	Mean	Score 5	Score 4	Score 3	Score 2	Score 1
Governance & Oversight	4.36	55	58	7	1	0
Implementation and Oversight	3.92	14	85	19	3	0
Objectives Clarity	3.83	2	99	20	0	0
Portfolio Design	3.37	0	59	56	5	1
Risk Framework	3.07	0	26	80	15	0
Responsible Investing/ESG	2.22	1	10	17	77	15

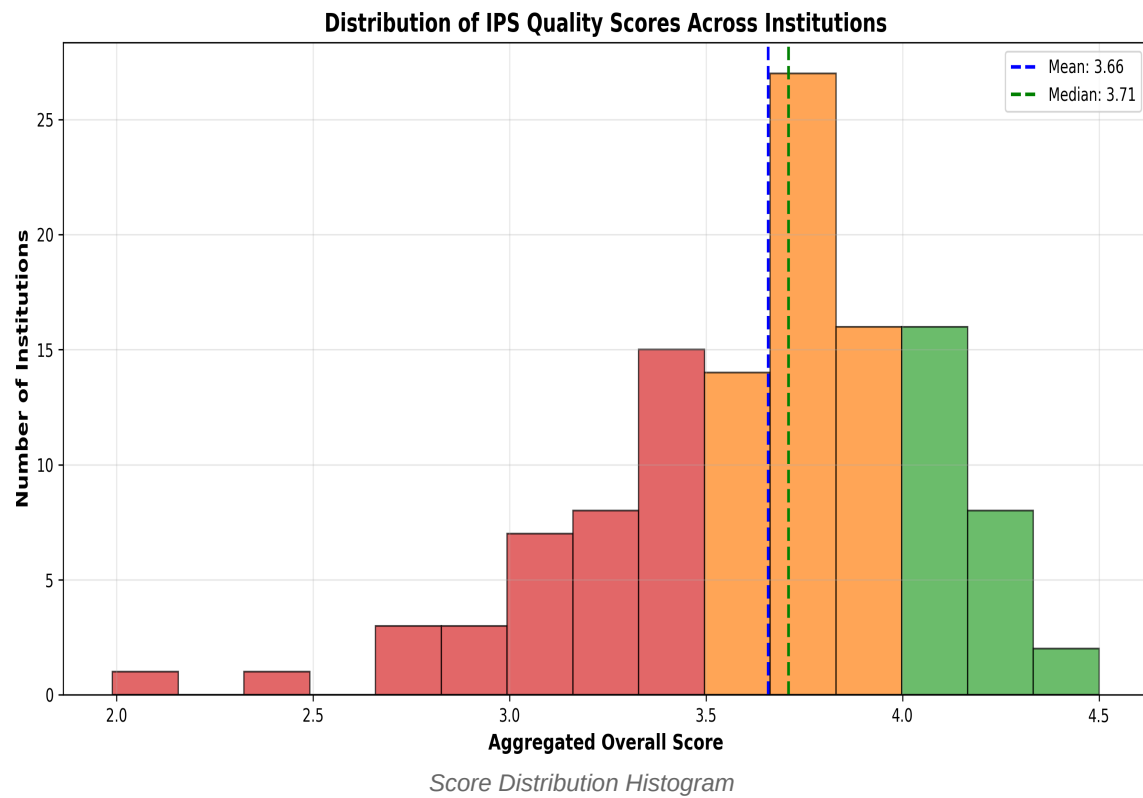
4.2 Performance Tiers

26 plans (21%) reach the Strong tier, an overall score of 4.0 or above. 57 (47%) sit between 3.5 and 4.0, and 38 (31%) fall below 3.5 (Table 5). The judge's anchors are demanding: a document reaches the top tier only when its core provisions are quantified and enforceable, not merely present.

Table 5: Plans by performance tier

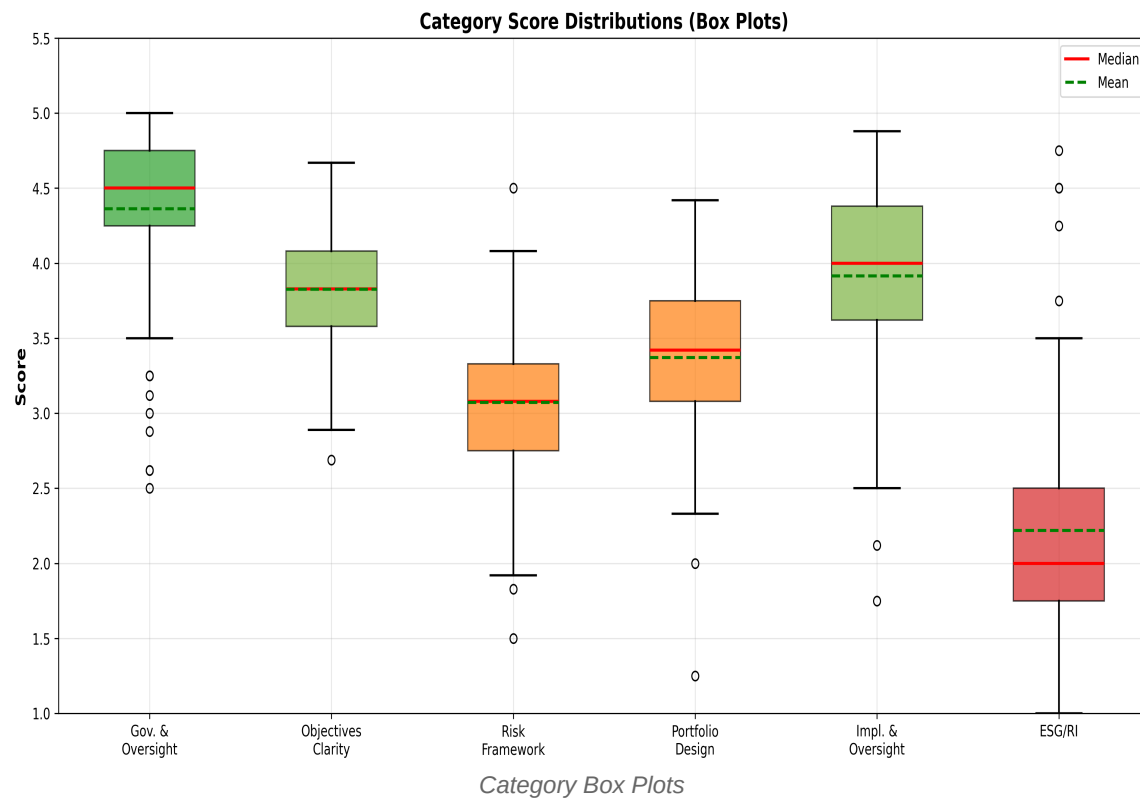
Tier	Score band	Count	Share	Mean overall
Strong	4.0 and above	26	21%	4.16
Middle	3.5 to 3.99	57	47%	3.75
Lower	below 3.5	38	31%	3.18

Figure 3: Overall Score Distribution



The histogram shows the overall scores with the tiers colour-coded. Two thirds of the plans lie within one standard deviation of the mean, between about 3.2 and 4.1.

Figure 4: Category Score Distributions (Box Plots)



The box plots show each category's median, mean, quartiles and outliers. Governance & Oversight is tight and high; Responsible Investing/ESG is the widest, running from documents that say nothing to a handful that treat it as a policy in its own right.

4.3 Plan Type and Size

Documentation quality varies little with what kind of plan wrote the document, or how big it is. By plan type the group means run from 3.6 for statewide systems to 3.8 for municipal systems (Table 5b), a spread well inside the spread within any one type. By size, the rank correlation between a plan's assets and its overall score is about 0.1: bigger plans score a little higher on average, but size explains almost none of the variation, and every size band holds both strong and weak documents (Table 5c). A well-specified policy is a drafting choice, not a function of resources.

Groups of fewer than 10 plans are shown by count only. Plan type and size are listed beside each plan's name in the Library, so a statistic over two or three plans would be those plans' own scores under another label.

Table 5b: Overall score by plan type

Plan type	Plans	Mean overall	Median overall	At 4.0 or above
Statewide systems	49	3.58	3.62	10 (20%)

Teachers' systems	15	3.62	3.73	4 (27%)
County systems	14	3.72	3.75	2 (14%)
Municipal systems	20	3.80	3.81	7 (35%)
Police and fire systems	16	3.76	3.75	3 (19%)
School-employee systems	5	-	-	-
University systems	2	-	-	-

Table 5c: Overall score by assets under management

Assets	Plans	Mean overall	Median overall	At 4.0 or above
Under \$10bn	39	3.65	3.71	5 (13%)
\$10bn to \$30bn	42	3.57	3.69	8 (19%)
\$30bn to \$75bn	20	3.76	3.87	7 (35%)
\$75bn and over	20	3.75	3.77	6 (30%)

5. Institutional Rankings

5.1 Top Performers

The ten highest-scoring documents, with the category in which each is strongest:

Plan	Overall
Plan 39	4.50
Plan 117	4.44
Plan 137	4.33
Plan 113	4.29
Plan 58	4.27
Plan 16	4.25
Plan 76	4.25
Plan 109	4.23
Plan 136	4.21
Plan 150	4.19

5.2 Pattern Analysis: What the Strongest Documents Share

Table 6 sets the mean category score of the five strongest documents beside that of the five weakest.

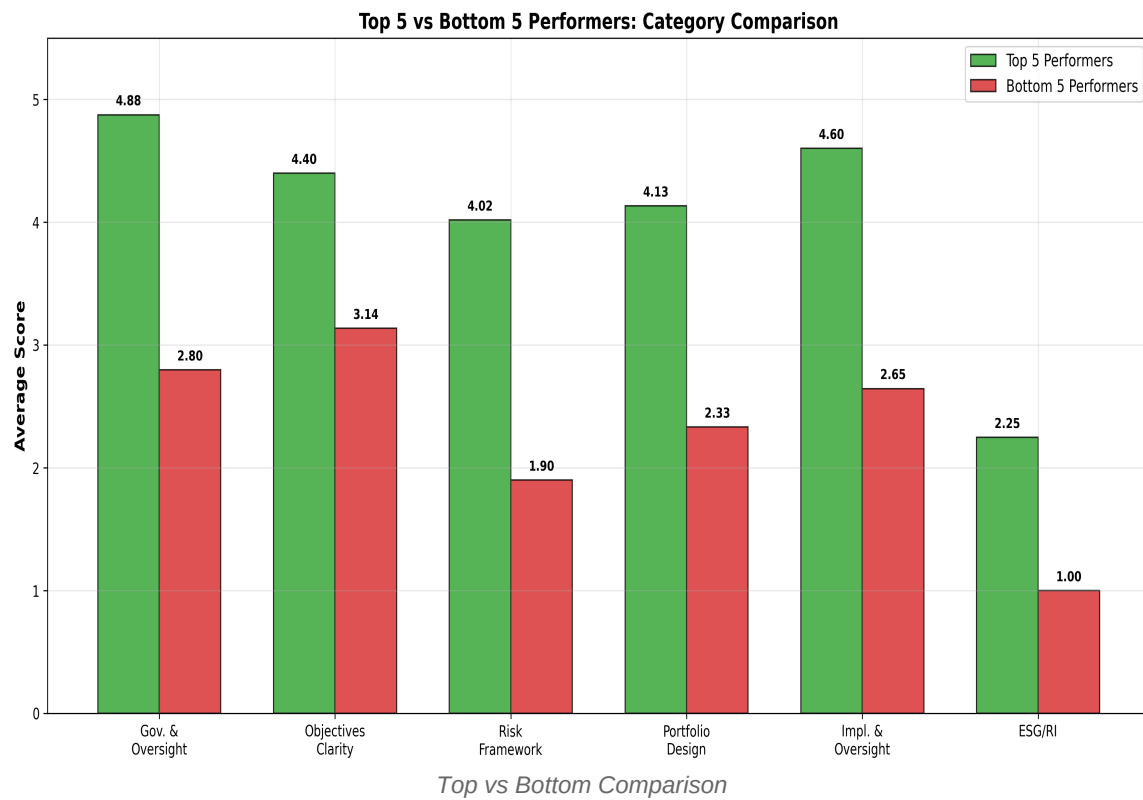
Table 6: Mean category score, top five versus bottom five plans

Category	Top five	Bottom five	Gap
Risk Framework	4.02	1.90	2.12
Governance & Oversight	4.88	2.80	2.08
Implementation and Oversight	4.60	2.65	1.96
Portfolio Design	4.13	2.33	1.80
Objectives Clarity	4.40	3.14	1.26
Responsible Investing/ESG	2.25	1.00	1.25

The gap is widest in Risk Framework, where the top five average 4.0 against 1.9 for the bottom five. It is narrowest in Responsible Investing/ESG, which is weak in both groups: the strongest documents in the sample are not, in the main, the ones that have engaged with ESG. What they share is depth of enforceability in the core governance and portfolio domains - explicit asset-class targets and ranges, numeric risk limits with a named monitoring body, specific rebalancing triggers, defined review cadences.

Notable finding: Three of the ten highest-scoring plans come from a single state. Enforceable, decision-binding policy language appears to cluster among one plan that share drafting standards, advisers or a statutory framework.

Figure 5: Top vs Bottom Performers by Category



5.3 Institutions Requiring Significant Improvement

The five lowest-scoring documents, with the core category in which each is weakest (ESG, optional and weak in all five, is left out of the comparison):

Plan	Overall
Plan 151	2.78
Plan 88	2.77
Plan 146	2.71
Plan 156	2.49
Plan 78	1.99

6. Areas of Strength

6.1 Governance & Oversight

Table: Governance & Oversight

Mean	Median	Min	Max	Plans at 4.0 or above	Plans at 2.0 or below
4.36	4.50	2.50	5.00	100 (83%)	0 (0%)

Public pension plans document who decides what. Board authority, delegation to staff and consultants, the reporting cycle and the review cadence are the best-specified parts of almost every document, and the category is the only one in which most plans reach the Strong anchor.

6.2 Implementation and Oversight

Table: Implementation and Oversight

Mean	Median	Min	Max	Plans at 4.0 or above	Plans at 2.0 or below
3.92	4.00	1.75	4.88	64 (53%)	1 (1%)

Manager selection, monitoring and termination, and the benchmarks against which each sleeve is measured, are set out in operational detail by about half the sample, and the category has the second-widest spread: where it is weak, the benchmark is unnamed or the termination decision has no trigger.

7. Areas of Weakness

7.1 Responsible Investing/ESG

Table: Responsible Investing/ESG

Mean	Median	Min	Max	Plans at 4.0 or above	Plans at 2.0 or below
2.22	2.00	1.00	4.75	4 (3%)	61 (51%)

Critical finding: 61 plans (51%) score 2.0 or below in ESG, meaning the document either says nothing about responsible investment or confines it to proxy voting. At the other end, 4 plans reach 4.0 or above, so a fully specified ESG policy does exist in the sample; it is rare. The category is the widest-spread in the assessment because it runs from absent to well-specified, where every core category runs from weak to strong. ESG is optional in the rubric and never moves the headline score; its weakness is reported, not penalised.

7.2 Risk Framework

Table: Risk Framework

Mean	Median	Min	Max	Plans at 4.0 or above	Plans at 2.0 or below
3.07	3.08	1.50	4.50	4 (3%)	4 (3%)

Risk tolerance is usually stated in words, rarely in numbers. Few documents give a volatility target, a drawdown or funded-status threshold, or a liquidity buffer with a stress test behind it, and fewer still name who monitors the limit and what happens when it is breached. This is the clearest expression of the enforceability gap in Section 3.3.

7.3 Portfolio Design

Table: Portfolio Design

Mean	Median	Min	Max	Plans at 4.0 or above	Plans at 2.0 or below
3.37	3.42	1.25	4.42	11 (9%)	2 (2%)

Strategic asset allocation is generally documented, but allowable ranges and the treatment of illiquid drift are often missing. Rebalancing policies omit triggers or authority. The spending and distribution policy is the single most common critical gap in the sample: most documents defer it to the actuarial valuation or the plan statute rather than state the fund's own commitment.

8. Critical Gaps Analysis

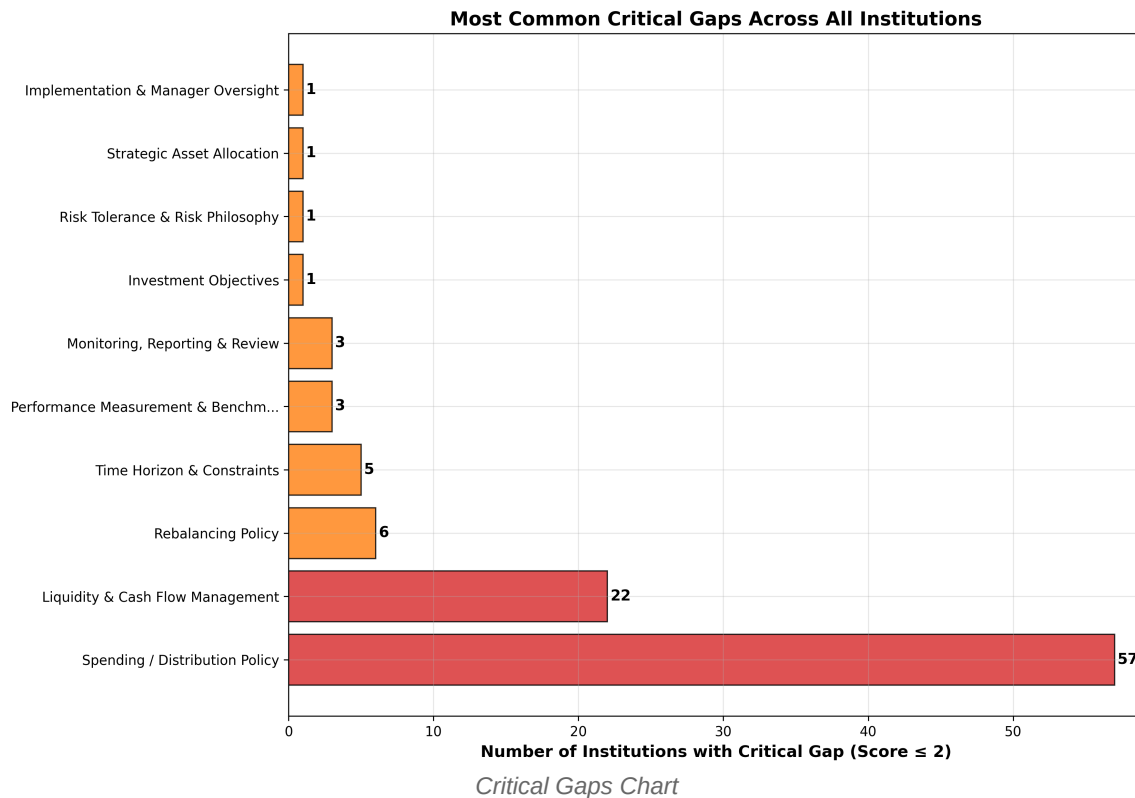
64 plans (53%) carry at least one critical gap, a mandatory domain scored at 2 or below, and 100 such gaps are recorded in total. They are not spread evenly.

8.1 Most Common Critical Gaps

Domain	Plans with a critical gap
Spending / Distribution Policy	57
Liquidity & Cash Flow Management	22
Rebalancing Policy	6
Time Horizon & Constraints	5
Performance Measurement & Benchmarks	3

Monitoring, Reporting & Review	3
Investment Objectives	1
Risk Tolerance & Risk Philosophy	1
Strategic Asset Allocation	1
Implementation & Manager Oversight	1

Figure 6: Critical Gaps Frequency



Spending / Distribution Policy is the most frequent critical gap, present in 57 of the 64 plans that have one. The gaps concentrate in exactly the domains that require a quantified, operational commitment: spending and liquidity mechanics, rebalancing triggers, and the horizon and risk-limit provisions that govern them.

8.2 Case Study: The Plan With the Most Critical Gaps

One plan carries more critical gaps than any other document in the sample; the domains are listed below with their scores. Reading them together shows the shape of a weak IPS: the policy is present as a description, absent as a commitment.

Domain	Score
--------	-------

Risk Tolerance & Risk Philosophy	1.75
Time Horizon & Constraints	1.75
Liquidity & Cash Flow Management	1.00
Spending / Distribution Policy	1.00
Strategic Asset Allocation	1.50
Rebalancing Policy	1.25
Performance Measurement & Benchmarks	1.75
Implementation & Manager Oversight	1.75

Focus Area: Risk Tolerance & Risk Philosophy

Strong policy language in this domain includes:

- Clear triggers or thresholds for action
- Explicit authority and accountability for the decision
- A defined review cadence and documentation requirement
- Procedures for exceptions and for extreme market conditions

Each of these turns a description into a decision rule. That is the difference the judge measures, and it is what separates the documents in Section 5.1 from those in Section 5.3.

Figure 7a: Top 20 Institution x Category Performance Heatmap

Figure omitted from the public edition.

The heatmap shows the twenty highest-scoring plans across the six categories.

Figure 7b: Bottom 20 Institution x Category Performance Heatmap

Figure omitted from the public edition.

The heatmap shows the twenty lowest-scoring plans across the six categories.

9. Category-Specific Insights

9.1 Governance & Oversight

Table: Governance & Oversight

Mean	Median	Min	Max	Plans at 4.0 or above	Plans at 2.0 or below
4.36	4.50	2.50	5.00	100 (83%)	0 (0%)

The strongest category. Board structure, delegated authority and the monitoring cycle are set out clearly in most documents; the few weak scores come from policies that describe the board's role without saying what it decides, or that set no review cadence at all.

9.2 Objectives Clarity

Table: Objectives Clarity

Mean	Median	Min	Max	Plans at 4.0 or above	Plans at 2.0 or below
3.83	3.83	2.69	4.67	43 (36%)	0 (0%)

Near-universal competence. Purpose, scope and objectives are stated in nearly every document. Scores fall where the return objective is left qualitative, or the philosophy does not connect to the allocation that follows.

9.3 Risk Framework

Table: Risk Framework

Mean	Median	Min	Max	Plans at 4.0 or above	Plans at 2.0 or below
3.07	3.08	1.50	4.50	4 (3%)	4 (3%)

A gap between qualitative statements and quantified tolerance. The strongest documents give a volatility or tracking-error target, a drawdown or funded-status trigger, and a liquidity buffer with a stress test behind it; most give a paragraph of intent.

9.4 Portfolio Design

Table: Portfolio Design

Mean	Median	Min	Max	Plans at 4.0 or above	Plans at 2.0 or below
3.37	3.42	1.25	4.42	11 (9%)	2 (2%)

Targets are usually documented; ranges, rebalancing triggers and the treatment of illiquid drift less so. The spending and distribution policy, the most common critical gap, is typically deferred to documents outside the IPS.

9.5 Implementation and Oversight

Table: Implementation and Oversight

Mean	Median	Min	Max	Plans at 4.0 or above	Plans at 2.0 or below
3.92	4.00	1.75	4.88	64 (53%)	1 (1%)

About half the sample sets out manager selection, monitoring and termination, and sleeve-level benchmarks, in operational detail. Where the category is weak the benchmark is unnamed, or the termination decision has no stated trigger.

9.6 Responsible Investing/ESG

Table: Responsible Investing/ESG

Mean	Median	Min	Max	Plans at 4.0 or above	Plans at 2.0 or below
2.22	2.00	1.00	4.75	4 (3%)	61 (51%)

The weakest category and the widest. Half the sample says nothing beyond proxy voting; a handful state an ESG policy with objectives, a stewardship process and reporting, and treat it as a policy in its own right.

10. Strategic Asset Allocation: What the Policies Set

10.1 The Allocation Table

102 of the 121 documents (84%) set out a target allocation. The rest refer to a schedule kept outside the IPS, give ranges without a target, or name asset classes without weights. The best-practice anchor asks for a complete row for every asset class - a target, an allowable range and a benchmark - and that row is the exception: 61 of the 102 tables give a range for every asset class and only 18 name a benchmark for every one (Table 7). A typical table carries 7 asset classes. Where the benchmark is missing, the document has set a weight it cannot measure: Section 9.5 finds the same gap from the implementation side.

Table 7: What the target allocation table contains

Shares are of the 102 documents with a target table.

Feature	Documents	Share
A range for every asset class	61	60%
A benchmark for every asset class	18	18%
More than one portfolio governed	5	5%

10.2 The Typical Policy Portfolio

The median document puts about 46 per cent in public equity and 23 per cent in fixed income, and spreads the rest across private equity, real estate, real assets, private credit and absolute-return strategies (Table 8). The interquartile spread is wide in every class, so the typical portfolio is a description, not a norm. A range is typically about ten points wide. The illiquid-drift problem of Section 7.3 is what happens when a fall in public markets pushes a private allocation past the top of one, and the document says nothing about it.

Table 8: Target weight by asset class, per cent of the fund

Asset class	Documents with a target	Lower quartile	Median	Upper quartile	Median range width
Public Equity	87	36.0	46.0	52.0	10.0
Fixed Income	85	18.0	23.0	29.0	9.0
Private Equity	71	8.5	11.0	15.5	10.0
Private Credit / Debt	52	3.9	5.9	10.0	7.0
Real Estate	64	7.8	9.5	10.0	10.0
Real Assets / Infrastructure	68	4.0	8.0	14.0	7.0
Hedge Funds / Absolute Return	36	5.0	6.0	10.2	8.0
Cash	59	0.0	1.0	2.0	5.0
Multi-Asset / Other	50	5.5	14.2	24.0	10.0

11. How the Policies Organise the Allocation

11.1 Four Structures

Behind the target table sits a choice about what the board is allocating between. The documents fall into four structures (Table 9):

- **Asset-class policy portfolio.** The traditional table: a target and range for equity, fixed

income, real estate, private equity and so on. The asset class is the unit of decision.

- **Asset classes plus a functional sleeve.** The same table with one or more sleeves defined by

what they do rather than what they hold: risk-mitigating strategies, crisis risk offset, a diversifying sleeve, an inflation-sensitive or real-return sleeve.

- **Role-based.** The top level is a set of roles - growth, income or stability, real return or

inflation protection, diversification or risk reduction - with asset classes nested under each. The GRID pattern (Growth, Real return, Income, Diversification) is the best-known version; the documents here use several close relatives of it.

- **No allocation in the IPS.** The targets sit in a separate schedule or exhibit.

A **total portfolio approach** - one passive reference portfolio fixing the fund's risk, with every investment competing for capital against it rather than against an asset-class budget - is a fifth possibility. No document in the sample adopts it in full. Its building block, the reference portfolio, does appear: 7 documents measure the fund against a passive reference portfolio, as a test of whether the policy portfolio's complexity is paying for itself, while keeping an asset-class table as the thing the board sets.

Table 9: Allocation structure and overall score

Structure	Documents	Share	Mean overall	Median overall	At 4.0 or above
Asset-class policy portfolio	83	69%	3.65	3.71	18 (22%)
Asset classes plus a functional sleeve	21	17%	3.74	3.72	4 (19%)
Role-based (growth, income, real return, diversifying)	12	10%	3.91	3.83	4 (33%)
No allocation in the IPS	5	4%	-	-	-

11.2 What the Structures Have in Common

The asset-class table remains the norm (69% of documents). The role-based documents score higher on average (3.9 against 3.7), but the group is small and the comparison controls for nothing, so the safer reading is that the plans which have rethought the structure of their allocation have tended to rewrite the rest of the document with it.

Across the role-based and hybrid documents the same handful of roles recurs (Table 10). Growth is the anchor of every role-based structure that names one; the defensive side is where the vocabulary varies most - stability, safe assets, risk reduction, crisis risk offset, principal protection - for what is in substance one job: to hold up when growth assets fall.

Table 10: Roles named in role-based and hybrid documents

Role	Role-based documents	Hybrid documents
Growth	11	0
Stability, safe or risk-reducing assets	8	0
Diversifying	6	9
Real return or inflation protection	7	7
Income	2	3
Risk mitigation	1	5
Liquidity	2	0

Three further features cut across the structures:

- **Total-fund leverage.** 10 documents allow policy weights to sum past 100 per cent, or set an asset-allocation leverage target or range. Others define asset-allocation leverage only to prohibit it.
- **Risk-based allocation.** 2 documents allocate by risk rather than capital, balancing the contribution of economic risk factors instead of fixing dollar weights.
- **Reference portfolios.** The 7 documents of Section 11.1 that test the policy portfolio against a passive alternative.

The role-based structures also cluster: 4 of the 12 come from county systems in a single state, the pattern Section 16 lists under peer diffusion.

11.3 What the Structure Changes

A role-based structure moves the hard question up a level. An asset-class table asks the board how much to hold of each kind of asset; a role table asks how much growth the fund needs and how much protection it can afford, and leaves the choice of instruments to the level below. That is a better question, but it is only as good as the definition of each role. The strongest documents of this kind say what each role is for, which assets may fill it, and the benchmark or test that shows whether it did its job - a crisis-risk-offset sleeve measured by its behaviour in equity drawdowns, not by its return in a rising market. Where that test is missing, a role is a label on an asset class and the structure has changed nothing.

12. What Changes When a Policy Is Rewritten

The corpus is refreshed monthly. Since the current judge configuration was fixed, 33 plans have had their published document replaced and both versions scored. 3 of those replacements were not rewrites - 1 corrected the instrument (the stored file was the wrong document); 1 re-saved the same text; 1 replaced a scan

with a text copy of the same version. They set the noise floor: the same policy text, scored again, moved by less than a tenth of a point. The other 30 are revisions: a newer version of the same policy, adopted by the plan's board.

Revision is not, on this evidence, improvement. 10 of the 30 revisions raised the overall score by more than a twentieth of a point, 7 lowered it by as much, and 13 held (Table 11). The mean change is under a tenth of a point, and the median revision does not move the score at all. Most revisions update an allocation table, a benchmark or a manager provision and leave the rest of the document as it was. Gains of half a point or more are rare (2 of 30), and they come from documents rebuilt rather than amended.

Table 11: Overall score before and after revision

Outcome	Revisions	Share
Improved (by more than 0.05)	10	33%
Held (within 0.05)	13	43%
Fell (by more than 0.05)	7	23%
Mean change	0.06	-
Median change	0.00	-

Where revisions do move a score, they move it in the domains a plan revisits anyway: Risk Tolerance & Risk Philosophy, Strategic Asset Allocation and Performance Measurement & Benchmarks gain most (Table 12). Spending policy and liquidity, the two weakest core domains, barely move: a rewrite updates what the document already covers rather than adding what it lacks. The gaps of Section 8 will not close by the normal cycle of amendment; they need to be put on the agenda of a revision.

Table 12: Mean change in domain score across revisions

Domain	Mean change	Revisions that raised it	Revisions that lowered it
Risk Tolerance & Risk Philosophy	+0.15	10	4
Strategic Asset Allocation	+0.14	11	5
Performance Measurement & Benchmarks	+0.12	8	2
Time Horizon & Constraints	+0.09	7	5
Scope, Purpose & Coverage	+0.07	12	5
Rebalancing Policy	+0.07	8	9
Liquidity & Cash Flow Management	+0.07	11	9
Monitoring, Reporting & Review	+0.04	9	7
Spending / Distribution Policy	+0.02	8	6
Responsible Investment / ESG	+0.02	10	5
Implementation & Manager Oversight	+0.02	8	7

Investment Philosophy	-0.01	8	10
Governance & Decision-Making	-0.01	4	3
Investment Objectives	-0.03	6	8

13. Lessons From the Newest and Strongest Documents

13.1 What the Strongest Documents Say That Others Do Not

Table 13 compares the language of specific practices in the 26 Strong-tier documents with the sample as a whole and with the 30 revised documents of Section 12. It is a keyword scan of the text, not a score: it records that a document uses the words of a practice, so it over-counts a passing mention and misses a practice described in other words. Used to compare groups, where both errors fall evenly, it shows which practices travel with quality.

Table 13: Documents using the language of each practice

Practice	All documents	Strong tier	Revised documents
Active-risk (tracking-error) budget	54 (45%)	22 (85%)	13 (43%)
Review cycle for the IPS	89 (74%)	25 (96%)	23 (77%)
Drawdown or tail-risk measure	39 (32%)	14 (54%)	12 (40%)
Effective date stated	40 (33%)	14 (54%)	14 (47%)
Unfunded commitments addressed	17 (14%)	9 (35%)	4 (13%)
Leverage policy	15 (12%)	7 (27%)	4 (13%)
Illiquid drift in rebalancing or ranges	28 (23%)	9 (35%)	3 (10%)
Liquidity stress test or coverage measure	4 (3%)	3 (12%)	2 (7%)
Liquidity tiers	4 (3%)	3 (12%)	2 (7%)
Private-markets commitment pacing	34 (28%)	9 (35%)	7 (23%)
Climate risk named	21 (17%)	6 (23%)	4 (13%)
Dated interim targets or ranges	12 (10%)	4 (15%)	2 (7%)
Funded-status reference	41 (34%)	9 (35%)	6 (20%)
Revision or amendment history	18 (15%)	2 (8%)	6 (20%)

The pattern is clear at both ends. The practices that most separate the strongest documents from the sample are an active-risk budget, a review cycle for the IPS, a drawdown or tail-risk measure and a stated effective date. Liquidity stress testing and liquidity tiers are rare everywhere - 12% and 12% of even the Strong tier - which is the domain-level finding of Section 3.4 seen from the other side: no group of plans has yet made

liquidity a policy.

13.2 Suggested Additions to Best Practice

The rubric's anchors describe current best practice. The newest and strongest documents suggest where it should go next. Each item below is a candidate for the next rubric version and for any plan's next revision:

1. An active-risk budget as standard. 85% of Strong-tier

documents set a tracking-error budget against 45% of the sample. Every policy that permits active management should say how much active risk it buys, and who watches it.

2. A tail-risk limit beside the volatility target. A drawdown, CVaR or funded-status trigger states

the loss the board is not prepared to take, which a volatility figure does not.

3. Liquidity as a policy, not a paragraph. Tiers of liquidity, a coverage ratio over benefit

payments and capital calls, and a stress test that sizes the buffer - the practices almost no document yet has, in the domain where the critical gaps are.

4. Private markets governed as private. A commitment pacing plan, unfunded commitments counted

in liquidity planning, and a stated treatment of illiquid drift when public markets fall.

5. Every role defined and tested. Where a document allocates by role, or carries a functional

sleeve, it should say what the role is for and how its performance is judged (Section 11.3).

6. A reference portfolio as a yardstick. Measuring the policy portfolio against a simple passive

alternative tells the board whether its complexity is paying; few documents yet do it.

7. A leverage policy wherever leverage is possible. Define asset-allocation leverage, cap it, and

name who approves it - or prohibit it in terms.

8. Dated, versioned documents. An effective date on the cover and a revision history at the back.

54% of Strong-tier documents state an effective date against 33% of the sample; a reader, a regulator or an automated monitor cannot otherwise tell which version governs.

9. Temporary deviations on a sunset. Where a range is widened for a transition, the document should

say until when, as the dated interim ranges in the stronger documents do.

14. Implications and Recommendations

14.1 Key Takeaways

1. **Governance is documented well.** Public pension plans maintain well-structured policies with clear authority and reporting; this is the sector's established strength.
2. **Enforceability is the differentiator.** What separates the strongest documents from the weakest is not breadth of coverage but the presence of quantified, binding provisions across the core governance and portfolio domains. The quantification lens is a full point below the structure lens.
3. **Risk and portfolio mechanics are the operational gap.** Risk tolerance, liquidity, spending and rebalancing are the domains where a description most often stands in for a commitment, and where the critical gaps concentrate.
4. **ESG is a systematic gap, with proof that it can be closed.** 51% of plans score 2.0 or below; the few that reach the Strong anchor show what a specified policy looks like.
5. **The spread is wide.** About 2.5 points separate the strongest document from the weakest, so the improvement available to most plans is large and the models for it already exist within the sector.
6. **Revision does not close the deepest gaps.** Rewrites move the domains a plan revisits anyway; spending and liquidity, the weakest core domains, need to be put on the agenda of a revision deliberately (Section 12).
7. **The allocation is moving from asset classes to roles, slowly.** Most documents still set an asset-class table; the role-based minority score higher, and the best of them define each role and the test that shows it worked (Section 11).

14.2 Recommendations by Priority

High Priority

1. Quantify the risk framework (for plans scoring below 3.5 in Risk Framework)

- State risk tolerance as a numeric limit: volatility, tracking error, drawdown or funded-status threshold
- Name the body that monitors the limit, the frequency, and the action a breach requires
- Set a liquidity buffer and describe the stress test that sizes it

2. State the portfolio mechanics (for plans with a critical gap in Portfolio Design)

- Give every asset class a target and an allowable range, and say how illiquid drift is treated

- Set rebalancing triggers, the authority to act, and the frequency of review
- State the spending and distribution policy in the IPS, even where the statute or valuation governs it

Medium Priority

3. Complete the implementation chapter (for plans scoring below 3.5 in Implementation and Oversight)

- Name a benchmark for every sleeve and the standard for outperformance
- Specify manager selection criteria, the monitoring cycle and termination triggers

4. Formalise ESG integration (for plans scoring below 3.0 in ESG)

- State whether and how ESG risk enters investment decisions, and who is accountable
- Set out the proxy-voting and engagement framework and the reporting on it
- The 4 documents at the Strong anchor in ESG are the models to read

Ongoing

5. Maintain the governance chapter

- Keep a fixed review cycle for the IPS itself, annual or biennial
- Record every change and its rationale

14.3 Best Practices From the Strongest Documents

- **Coverage:** every mandatory domain is addressed with substantive content
- **Quantification:** measurable limits and targets stand beside every qualitative statement
- **Operational detail:** strategy is connected to implementation through stated processes
- **Named accountability:** each limit has a monitor and each decision has an owner
- **Currency:** the document shows a review cycle and recent amendment

15. Limitations

15.1 Methodology Constraints

- **Document-based assessment:** the scores reflect the written policy, not practice. A plan may run

robust processes its IPS does not describe.

- **Model-based scoring:** scores are produced by an LLM judge (Section 2.3), not a human panel. The judge is constrained by explicit anchors, a structured schema and evidence checks, but can still misread an atypical document. Until the expert validation (Section 2.3) is run, the scores should be read as the judge's assessment, not as ground truth.

- **Expert validation not yet run:** the planned validation uses a single rater with a test-retest ceiling on a stratified sample, not a multi-rater panel over the full universe, and no ratings have been collected yet.

- **Evidence verification is flag-only:** cited quotes are checked against the source, but the result is reported rather than enforced, and the matcher under-counts quotes reconstructed from tables.

- **Wrong-document risk:** the corpus is gathered from what each sponsor publishes, and some sponsors publish a different instrument in the place of an IPS; a single plan is withheld from scoring for that reason. A document of the wrong kind with a convincing cover page would not be caught automatically.

- **Point-in-time:** policies are amended. The assessment captures the published document at the analysis date; the corpus is refreshed monthly.

15.2 Sample Limitations

- **Coverage:** the 121 plans are drawn from the two hundred with the most assets, so the findings describe the large-plan sector and may not hold for small plans.

- **Selection:** a plan enters the sample only if its sponsor publishes the IPS. Plans that do not publish are absent, and their documents may differ systematically from those that are.

15.3 Rubric Limitations

- **ESG scoring:** the rubric may under-score plans whose responsible-investment programme is real but documented outside the IPS.

- **Equal weighting:** mandatory domains carry equal weight in the overall score; some may matter more for a given plan type or circumstance.

- **Best-practice drift:** the anchors reflect current practice and will need revisiting as it evolves.

Section 13 suggests where.

15.4 Limits of the Added Analyses

- **Allocation structure is a judgement:** each document's structure was drafted from its target table and asset-class definitions by a keyword classifier and then checked by hand. Borderline cases - a single functional sleeve, a top level that mixes roles and asset classes - could reasonably be placed either side of a line.
- **Practice markers are a keyword scan:** they record the words of a practice, not its quality, and are used only to compare groups.
- **Revisions are few and recent:** the before-and-after comparison covers one refresh cycle and plans whose sponsor happened to revise; re-scoring noise on unchanged text is small but not zero.
- **No causal claims:** plan type, size and allocation structure are compared, not controlled for; a difference between groups is an association.

16. Conclusion

This assessment of IPS quality across 121 U.S. public pension plans finds a sector with strong governance fundamentals and a wide gap between describing a policy and committing to one. The mean overall score of 3.7 reflects competent documentation; the strongest documents show what a fully specified policy looks like, and they do so within the same constraints every other plan faces.

Key Findings

Strengths:

- Governance structures and oversight are well established (4.4 average)
- Objectives and philosophy are clearly articulated (3.8 average)
- Implementation and manager oversight are set out in operational detail by about half the sample (3.9 average)

Gaps:

- Risk frameworks lack quantified limits (3.1 average)
- Portfolio mechanics are under-specified, and spending policy is the most common critical gap (3.4 average)

- ESG integration is systematically weak (2.2 average)
- Across every domain, quantification and enforceability is the weakest lens (3.0 average)

The Path Forward

Closing the gaps identified here is not a documentation exercise. A specified, enforceable IPS:

1. Strengthens fiduciary accountability and the plan's legal position
2. Keeps decisions consistent across board transitions
3. Gives members, sponsors and legislators a transparent standard to hold the plan to
4. Sets clear expectations for staff, consultants and managers
5. Preserves institutional memory

Future Research Directions

- **Expert validation and a noise band:** run the designed blind validation (Section 2.3), and re-score unchanged documents repeatedly so every score, and every change between versions, carries an error band
- **Longitudinal analysis:** follow each revision as the monthly refresh continues, and test whether the domains this paper flags - spending, liquidity, risk limits - are the ones plans improve
- **Funding and contributions:** whether funded status, contribution discipline or benefit cash-flow pressure relate to documentation quality, using public plan-level data
- **Allocation structure and outcomes:** whether role-based, reference-portfolio and leveraged plans differ in realised risk, drawdowns and liquidity events from asset-class plans
- **Private-markets governance:** whether pacing plans, counted unfunded commitments and illiquid-drift rules are associated with fewer forced sales or range breaches in stressed markets
- **Peer diffusion and shared templates:** how structures and enforceable language spread among plans that share consultants, statutes or a state system, measured by text similarity between documents
- **The total portfolio approach:** criteria for scoring a document that sets risk through a reference portfolio rather than an asset-class table, before the first U.S. plan adopts it in full
- **Beyond U.S. public pensions:** the same rubric applied to endowments, sovereign funds and non-U.S. pension funds, where role-based and total-portfolio structures are more common
- **Outcomes:** whether documentation quality relates to governance events or investment outcomes

Final Observations

The spread of about 2.5 points between the strongest and weakest documents demonstrates that excellence is achievable within current constraints. The top performers, strongest in the core governance and portfolio domains, are proven models that other plans can adapt. Plans scoring below 3.5 should prioritise a systematic rewrite around quantified commitments; those in the middle tier can target the weak categories identified here; even the strongest have an unaddressed frontier, most often ESG.

An excellent IPS is both compass and constitution: it guides decisions and binds them. The sector has built the compass. The work that remains is the constitution.

Appendix: Assessment Rubric Summary

Rubric Overview

Rubric ID: ips_rubric_v1_3

Purpose: evaluate the quality, enforceability and governance clarity of Investment Policy Statement documentation. Scores reflect the written policy, not investment skill or outcomes.

Scoring Scale

Score	Level	Description
5	Best practice	Fully specified, enforceable, peer-aligned
4	Strong	Clear, operational, minor gaps
3	Adequate	Basic structure with gaps
2	Weak	Present but vague or non-operational
1	Deficient	Missing or purely aspirational

Assessment Domains (14 scored, 12 of them core)

The rubric defines 15 domains; the 14 that apply to a public pension plan cover **99 fact elements**:

1. **Scope, Purpose & Coverage**
2. **Investment Objectives**
3. **Investment Philosophy** - optional

4. Risk Tolerance & Risk Philosophy
5. Time Horizon & Constraints
6. Liquidity & Cash Flow Management
7. Spending / Distribution Policy
8. Strategic Asset Allocation
9. Rebalancing Policy
10. Performance Measurement & Benchmarks
11. Governance & Decision-Making
12. Monitoring, Reporting & Review
13. Implementation & Manager Oversight
14. Responsible Investment / ESG - optional

Tax Considerations is conditional on an institution type that does not apply to a public pension plan, so it is not scored and its fact elements are not counted above.

Aggregation Method

- **Lens scores:** the judge scores four quality lenses per domain against the written anchors
- **Domain score:** the mean of the applicable lens scores
- **Overall score:** the mean of the 12 core domain scores. Core means mandatory, or conditional

and applicable. Investment Philosophy and Responsible Investment / ESG are optional: both are excluded; tax considerations is not applicable and is not scored

- **Critical gap:** a mandatory domain scored at 2 or below
- **Category grouping:** the 14 domains are grouped into six categories for reporting; a category score

is the mean of its domain scores

Key Assessment Principles

1. **Document-based:** only what the IPS states is evaluated, never the plan's practice
2. **Enforceability:** higher scores require specific, operational, binding language
3. **Completeness:** every required element must be present for a top score
4. **Governance clarity:** roles, authorities and processes must be explicit

5. **Quantification:** numeric targets, thresholds and ranges outrank qualitative statements

End of Report