Measuring Public Pension Health

New Metrics and New Approaches

About the Author

Tom Sgouros has worked for over 30 years as a policy consultant specializing in public budgeting, finance, taxation, and other technical issues of public policy. He has consulted to political campaigns and office-holders, to activists and media outlets, and has been invited to testify about public finance issues to legislatures in several states. He was Senior Policy Advisor to the Rhode Island General Treasurer 2015-2016, and is now a fellow at The Policy Lab at Brown University. He is also a member of the research faculty in Computer Science, where he works on data science, visualization, and information theory.

Special thanks to Sarah Thang, Matthew J. Lyddon, David Yokum, and The Policy Lab at Brown University for providing project and financial management support.

Printed in the United States of America by a member of the Allied Printing Trades Council.

PENSION ACCOUNTING WORKING GROUP

Table of Contents

Executive Summary

A pension system is a complex organism, and any accounting of that system inevitably embodies choices about which aspects to measure, how to measure them, and how much emphasis to give each measurement. On the theory that any enterprise benefits from the regular examination of its assumptions and conventions, NCPERS, with the support of Arnold Ventures, hosted the Pension Accounting Working Group project. The Working Group, composed of public pension experts from across the country, met to discuss the existing accounting rules, to suggest new metrics for assessing the health of a pension system, and to consider new ways to think about old metrics.

Current accounting practices and the metrics defined by them are for the most part uniformly applied across the universe of public pension plans. However, what the metrics convey is not uniformly understood, and they are easy to misinterpret. The Working Group sought to develop new metrics to generate insights by adding context, for example by comparing a liability to a state's ability to levy taxes, assessing the context of contributions by their effect on the balance sheet, or considering risk when assessing the value of investment assets.

The Working Group also identified a lack of standardization in the reporting of pension results as a problem to address. Pension systems are complex, but a complex presentation encourages readers to make their own summary. The result is that observers will typically pick one or two metrics they think are most important and ignore the rest that make up the bulk of a valuation or financial report.

This report describes a "scorecard", a standardized summary of pension valuation results (shown on next page), as well as three new metrics, of varying degrees of novelty, to appear on it:

- The Scaled Liability is a measurement of pension liabilities against the size of the economy that supports these liabilities.
- The UAL Stabilization Payment (USP) is an objectively defined cash flow policy standard comparable to the funding ratio, an objectively defined balance sheet policy standard.
- Risk-Weighting Assets is a proposed method to assess the value of a plan's assets, taking into account its capacity to endure the downside risk it has taken on through its allocation of investments.

We assess the utility of these measures by calculating them for different systems, testing them against real data and real plans.

In addition, this report includes a discussion of the practice of computer simulation of pension systems, including stress testing, a widely used technique for risk assessment, but also sensitivity testing and projections. The discussion includes suggestions about increasing the comparability of stress tests and the ways in which such tests may or may not acquire meanings useful to plan managers and policy makers.

Rhode Island ERS (State Employees + Teachers)

POLICY					
Benefits		**Funding**		**Investments**	
Employee participation	●	Annual employer share	●	Investment strategy	●
Income replacement	●	UAL sources	●	Risk discussion	●
COLA terms	●	COLA funding	●	Allocation motivation	●
Other Benefits	●	Employee contribution	●	Benchmark defined	●

ACTION					
Benefits		**Funding**		**Investments**	
Benefit replacement	10yr 16%	USP % payroll	28.3%	Global equities	42%
	30yr 53%	ADC % payroll	31.3%	Fixed-income	24%
COLA	suspended until UAL>80%	Actual Contribution	31.3%	Real estate	7%
		Normal Cost	8.1%	Hedging	9%
SS participation	some	Experience study	●	Private equity	14%
		Assumed return	7%	Cash	4%
		Assumed inflation	2.5%	Investment mgmt fees	●
		Wage inflation	3%	Sharpe ratio	●

CONDITION					
Benefits		**Funding**		**Investments**	
Active state employees	N=10,803	Total liability	$18.89 billion	Assets/Benefits	8.05
	Age 49.2	Actuarial assets	$6.89 billion	Risk-weighted assets	6.92
Active teachers	N=13,372	Market Assets	$7.73 billion	Market returns *1-year*	
	Age 46.8	UAL as % payroll	26%	Net	2.2%
Retired state employees	N=9,270	POB debt	$0	Bench	11.2%
	Age 74.3	Scaled liability	0.4%	Market returns *5-year*	
Retired teachers	N=10,441	Net cash flow	21%	Net	0.1%
	Age 74.2	Extra contribution?	No	Bench	9.8%
Actual FY21 COLA	0.0%	Layered Amort?	●	Market returns *10-year*	
				Net	8.5%
Sponsor Fiscal Health				Bench	9.8%
Budgeted general revenue	$4.43 billion			Market returns *since 1995*	
Per capita income	$37,504			Net	7.7%
Poverty rate	10.6%				
GO Bonds M/SP/F	Aa2/AA/AA				

A mock-up of the proposed pension funding scorecard, for the Rhode Island Employee Retirement System.

Introduction

A pension system is a complex enterprise. Money flows in via payments and investments, and money flows out via benefits and expenses. Billions of dollars must be taken in, invested, and paid out every year. Many systems involve thousands, tens of thousands, even millions of members, each with their own priorities, needs, and lives. Those members depend on that money arriving on time, as promised.

Accounting was invented, and accounting rules exist, to measure the condition and prospects of any enterprise. Because a pension system is so complex, the accounting for it can hardly help but emphasize some aspects of a plan's circumstances and understate others. Those choices of emphasis are capable of shaping people's understanding of a system's state, and thus indirectly shaping actual policy and decisions made by its managers.

Because a pension system is so complex, the accounting for it can hardly help but emphasize some aspects of a plan's circumstances and understate others.

In 2019, NCPERS published a report whose thesis was that the pension fund accounting rules might themselves have an indirect role in the development of the condition of public pension funds across the country.[1] The report pointed out that the traditional rules emphasize risks that public plans do not face while ignoring or soft-pedaling risks that they endure every day. For example, a public pension plan need not worry about the disappearance of its sponsoring employer, as a private plan must. A city might go into bankruptcy, but there is no provision for a city to be liquidated. Whatever happens to the city government of Chicago, there will be a Chicago in its place afterward. Full funding is therefore not a necessary defense against liquidation of the sponsor, because a public plan does not face that risk.

One may argue that a fully funded plan remains desirable for other reasons, such as providing a path to lower the costs of providing pension benefits or a bulwark against volatility. Additionally, other aspects of plan design, such as the method used to allocate liabilities or the discount rate, can acknowledge the permanent nature of a public pension plan. But the risks a public plan faces are different from the risks a private plan faces, so the management of those risks, and the insurance against them, must be different.

On the other hand, while a pension plan does face substantial investment risk, there is no place in the rules to account for a portfolio's risk. A portfolio filled with risky assets is valued in an identical manner to a much more conservatively invested collection of assets. If you assume that the market price accurately reflects the risk, there is no problem.[2] But if you suspect that a market price reflects an imperfect assessment of the risk according to the appetites of only some market participants (who may not be pension funds), it is less clear. Almost by definition,

1 Sgouros, Tom, "The Case for New Pension Accounting Standards" NCPERS, 2019, https://www.ncpers.org/files/The%20Case%20for%20New%20Pension%20Accounting%20Standards_May%202019.pdf

2 This is the "Efficient Market Hypothesis," a tenet of much modern economic theory, but one that assumes a certain uniformity and rationality of market participants, among other mismatches between theory and reality. For a critique, see Andrei Shleifer "Inefficient markets: an introduction to behavioral finance," Oxford University Press, 2000, or Robert J. Shiller, "Irrational Exuberance," Princeton University Press, 3rd ed, 2015.

too much risk is unlikely to provide a secure retirement for members and low costs for its sponsor, even if there is considerable debate about what is "too much." The social science dictum, "what you measure is what you get," seems to be true here. If yield is what matters in the short term – which is the case when all eyes are on the market value of assets – then risk may get less attention, even if in the long term it is more important.

The perspective of that earlier report was influenced by the work of Amos Tversky and Daniel Kahneman in their inquiries about how, exactly, people make decisions.[3] The idea behind their work was to look at the ways in which the presentation of a decision is fundamentally linked to the factors one takes into account while making it, and therefore inextricably linked to the outcomes as well. This work has currency today through the work of Richard Thaler and the development of the psychology of choice, or "prospect theory."[4] The point of this work is not merely to admire the factors that go into people's decisions, but to explore the ways in which those factors can be changed to encourage better decisions.[5]

It is one thing to complain that a set of rules may encourage certain decisions, as the 2019 report did, but it is something entirely different to propose what to do about it. Through the generosity of Arnold Ventures, the Policy Lab at Brown University, and NCPERS, a Working Group was convened in February 2021 to consider additions to the important metrics used to assess a pension plan's funding. The Working Group did not set out to tear down the existing edifice, but to propose new metrics that represent additional ways to look at some aspects of a pension plan.

The Working Group identified a lack of standardization in the reporting of pension results as a problem worth attention. Pension systems are complex, but a complex presentation may inadvertently encourage readers to make their own summary. Many industry observers, including reporters and members of the public, but also some managers and trustees, will typically pick one or two metrics they think are most important – often the funding ratio alone – and ignore the rest of the results that make up the bulk of a valuation or financial report. The headline will often become about the condition of the system, without consideration of how it got there or what is being done about it.

It is one thing to complain that a set of rules may encourage certain decisions but it is something entirely different to propose what to do about it.

In order to encourage users to consider more than one or two metrics in their evaluation of a pension plan, it makes sense to promulgate a succinct standard presentation of a set of important metrics: a "Pension Funding Scorecard." Such a display can be designed to illustrate the point that a system's health is not only dependent on its condition but also on the policies in place and the actions taken by its management, and do it in a compact, standard, and readily digestible fashion.

---

3 See, for example, Kahneman, Daniel. 2003. "Maps of Bounded Rationality: Psychology for Behavioral Economics." American Economic Review, 93 (5): 1449-1475. https://www.aeaweb.org/articles?id=10.1257/000282803322655392

4 See, for example, "Mental Accounting and Consumer Choice," Richard H. Thaler, Marketing Science, 27(1), January-February 2008, pp15-25.

5 The canonical account of the intersection of this work with public policy is "Nudge: Improving Decisions About Health, Wealth, and Happiness," Richard H. Thaler and Cass R. Sunstein, Penguin, 2009.

Rhode Island ERS (State Employees + Teachers)

POLICY					
Benefits		**Funding**		**Investments**	
Employee participation	●	Annual employer share	●	Investment strategy	●
Income replacement	●	UAL sources	●	Risk discussion	●
COLA terms	●	COLA funding	●	Allocation motivation	●
Other Benefits	●	Employee contribution	●	Benchmark defined	●
ACTION					
Benefits		**Funding**		**Investments**	
Benefit replacement	10yr 16%	USP % payroll	28.3%	Global equities	42%
	30yr 53%	ADC % payroll	31.3%	Fixed-income	24%
COLA	suspended until UAL>80%	Actual Contribution	31.3%	Real estate	7%
		Normal Cost	8.1%	Hedging	9%
SS participation	some	Experience study	●	Private equity	14%
		Assumed return	7%	Cash	4%
		Assumed inflation	2.5%	Investment mgmt fees	●
		Wage inflation	3%	Sharpe ratio	●
CONDITION					
Benefits		**Funding**		**Investments**	
Active state employees	N=10,803	Total liability	$18.89 billion	Assets/Benefits	8.05
	Age 49.2	Actuarial assets	$6.89 billion	Risk-weighted assets	6.92
Active teachers	N=13,372	Market Assets	$7.73 billion	Market returns *1-year*	
	Age 46.8	UAL as % payroll	26%	Net	2.2%
Retired state employees	N=9,270	POB debt	$0	Bench	11.2%
	Age 74.3	Scaled liability	0.4%	Market returns *5-year*	
Retired teachers	N=10,441	Net cash flow	21%	Net	0.1%
	Age 74.2	Extra contribution?	No	Bench	9.8%
Actual FY21 COLA	0.0%	Layered Amort?	●	Market returns *10-year*	
				Net	8.5%
Sponsor Fiscal Health				Bench	9.8%
Budgeted general revenue	$4.43 billion			Market returns	*Since 1995*
Per capita income	$37,504			Net	7.7%
Poverty rate	10.6%				
GO Bonds M/SP/F	Aa2/AA/AA				

The Pension Funding Scorecard includes some innovative new metrics, as well as a proposal to incorporate the abstractions of policy into a quantitative framework, highlighted here.

To facilitate the production of scorecards, perhaps even scorecard apps, we also call for the widespread adoption of an electronic reporting standard for pension results. The Public Plan Database (PPD) has already done much to elevate the level of research and discussion of public pension plans.[6] However, due in part to the effort of data collection, the PPD covers only a fraction of the extant pension plans in the country. Standardization of reporting would facilitate the PPD mission as well as facilitate the analysis of pension plan results across the country.

Key to the scorecard is the distinction between a plan's condition, the actions taken by its managers, and the policies it pursues. These are three distinct, but independently vital, considerations. After all, two systems in the same funding position are in very different conditions if one has sound policies to improve and the other does not. By separating the presentation of these metrics, the scorecard emphasizes the distinction between what is the case, what is being done about it, and what is the ultimate goal.

In addition to an innovative way to quantify abstractions like policy, this report presents three new metrics to include in the scorecard, each of which might be thought of as a way to incorporate context into a measurement:

- The Scaled Liability is a measurement of pension liability against the size of the economy that supports it. Metrics like this, that use economic strength as a proxy for tax capacity, are already widely used to assess sustainability. Nonetheless, it is useful to identify a standard way to make the comparison, especially for smaller governments where the economic statistics are not as readily available.
- The Unfunded Actuarial Liability (UAL) is a balance sheet metric widely used to assess a plan's condition. The UAL Stabilization Payment (USP) is an objectively defined cash flow policy standard that shows how expensive it is to maintain that funding level. This is not a statement of what is good policy, simply a benchmark against which to measure it. That is, once a USP is defined, it is informative to know how it compares to actual payments, or to the payments defined by a plan's actuaries.
- A Risk-Weighted Asset Value is a measure of asset value that takes into account a plan's capacity to endure the downside risk of a bear market, given its current cash flow and allocation of investment assets.

The report assesses the utility of these measures by calculating them for different systems, testing them against real data and real plans. We do not go into the behavioral questions of choice architecture and incentives suggested by the 2019 report,[7] though this report can be read as setting up measures that can be tested in that fashion. The question of which measures produce better outcomes is one that can be tested experimentally, but in order to do that, one needs alternatives to test.

---

6 Public Plans Data. 2001-2020. Center for Retirement Research at Boston College, MissionSquare Research Institute, National Association of State Retirement Administrators, and the Government Finance Officers Association. https://publicplansdata.org/

7 Though see "Nudging the nudger: Toward a choice architecture for regulators," Susan E. Dudley and Zhoudan Xie, Regulation and Governance, 14 June 2020, https://doi.org/10.1111/rego.12329; "Nudges for nudgers," Nature Energy volume 3, p701, 2018. https://www.nature.com/articles/s41560-018-0255-4

In addition, this report includes a discussion of the practice of computer simulation of pension systems, including stress testing, a widely used technique for risk assessment, but also sensitivity testing and projections. The discussion includes suggestions about increasing the comparability of stress tests and the ways in which such tests may or may not acquire meanings useful to plan managers and policy makers. This chapter also suggests how experience with stress testing, if carried out in a systematic way, can lead to the substitution of less expensive metrics that may provide the same information, such as the risk-weighted asset valuation proposal highlighted here.

WORKING GROUP

The contents of this document are the responsibility of the author and the author alone. The Working Group members listed below participated in discussing and debating these concepts – and the author is forever indebted to them for their contributions, suggestions, and criticisms – but members of the group do not necessarily agree with this document as a whole, or with any particular piece of it.

Tom Sgouros, co-chair, The Policy Lab at Brown University
Scott McCarty, co-chair, Arizona Public Safety Personnel Retirement System
Hank Kim, facilitator, National Conference on Public Employee Retirement Systems
JP Aubry, Boston College Center for Retirement Research
Michael Belarmino, Government Finance Officers of America
Keith Brainard, National Association of State Retirement Administrators
Michael Cohen, CalPERS
Dan Doonan, National Institute for Retirement Security
David Draine, Pew Charitable Trust
Kelly Fox, CalPERS
Bernard Gallagher, Center on Budget and Policy Priorities
Michael Kahn, National Conference of Public Employee Retirement Systems
Jim Kane, National Education Association
Jim Link, Public Finance Management
Todd Tauzer, Segal

Liabilities in Context

Summary: *Security for a pension plan is ultimately dependent on the financial strength of the plan sponsor. It makes sense, then, to try to incorporate that strength into a measurement of a pension liability. Others have done exactly that, such as Michael Kahn for NCPERS and Lenney et al. for Brookings. This Working Group report suggests a convenient way to standardize this comparison for states, counties, and municipalities and discusses the results for several plans.*

Ultimately, what stands behind any pension plan is the financial strength of the plan sponsor. In the event of devaluation or depletion of a plan's assets, or even just underfunding, the ability of the sponsor to make up the difference is the ultimate guarantee for a retiree's pension. A city whose pension assets are allowed to dwindle away still owes the commitments made to its retirees. In that position, a city with a large economy relative to the size of its pension liability will be much more secure than one with a smaller economy. It makes sense, therefore, to find a way to measure the size of the pension commitment against this economic strength.

Ultimately, what stands behind any pension plan is the financial strength of the plan sponsor.

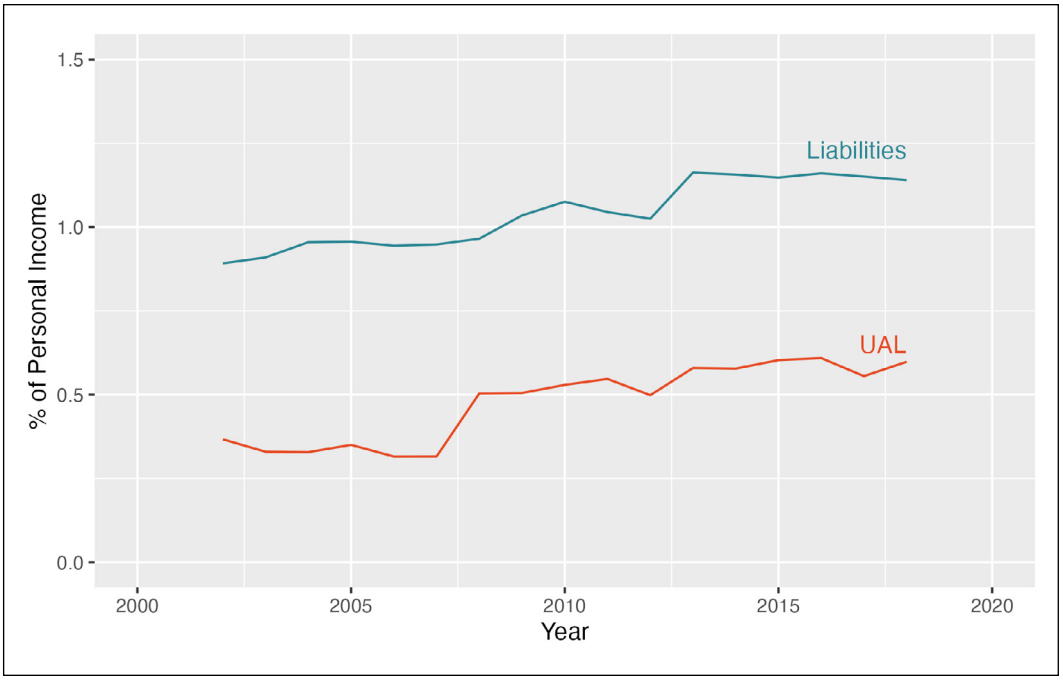
This is tricky territory. The affordability of some governments' expenses is a difficult concept, with no widely accepted definition. Debates over the subject have a history stretching back to the beginning of this republic when Alexander Hamilton wrote, "[T]here can be no common measure of national wealth; and of course, no general or stationary rule, by which the ability of a State to pay taxes can be determined."[8] He was not the first to observe this problem, nor the last. The debate since then has been vigorous, to say the least. It's fair to say that disputes over the question of tax capacity have not infrequently been decided by armed conflict, including here in the United States. And yet one of the common complaints about pension liabilities is that they are not affordable. It makes sense to attempt to meet that argument on the ground on which it is made, even though little can be dispositive in such uncertain territory.

Aggregate statistics are suggestive, implying the collective liability of the nation's public pension plans – the present value of all the future pension payments they will pay to their employees – amounts to only a small percentage of the nation's economic activity, especially when spread out over the decades over which it is to be paid.[9] The figure below compares the collective liability of all state and local public pension plans with 30 years of personal income. The lower line is the liability reduced by the assets on hand to make those payments, also called the "unfunded" portion of the liability.

8 Federalist Papers, No. 21, 12 December 1787.

9 See, e.g., Michael Kahn, "Enhancing Sustainability of Public Pensions," NCPERS, 2022. https://www.ncpers.org/files/ncpers-enhancing-sustainability-of-public-pensions-2022.pdf

The upward trend of the unfunded liability seems a cause for concern, but upon closer examination, what emerges are three relatively stable periods, punctuated by the dramatic asset losses of the 2008–09 recession and the adoption in 2014–15 of the new Government Accounting Standards Board (GASB) rules for valuing pension liabilities. This is not a picture of uncontrolled growth, even if the growth of the unfunded portion may be a cause for concern.



Liabilities and unfunded liabilities as a percentage of personal income. The lines slope upward, but this is not a picture of out-of-control growth, but more like periods of stability punctuated by increases that correspond to external events.

The unfunded liability values amount to a fraction of a percent of the national income. For comparison, the cost of the Pentagon is about 4 percent of US national income. Elementary and secondary education costs around 3.1 percent of national income, and at the state level it varies from around 2.5 percent of state GDP to more than 4 percent.[10] The cost of all state and local governments' expenditures, excluding pension costs, currently runs at about $3.7 trillion a year in the United States, approximately 17 percent of national personal income. These numbers vary considerably at the state level, as do net federal contributions to each state, so they should be taken only as a rough guide to the scale of these expenses.

Lenney *et al.* use a similar approach in their work, making the case that if the growth of the pension liability can be constrained to be less than or equal to the growth in the economy governed by the plan sponsor, the plan should be considered sustainable, even if it appears not to be fully funded.[11] The Federal Reserve includes

10 https://ncses.nsf.gov/indicators/states/indicator/public-school-expenditures-to-state-gdp

11 Jamie Lenney, Byron Lutz, Finn Schüle, and Louise Sheiner. "The Sustainability of State and Local Government Pensions: A Public Finance Approach." BPEA Conference Draft, Spring, 2021. https://www.brookings.edu/bpea-articles/the-sustainability-of-state-and-local-government-pensions-a-public-finance-approach/

a comparison of net pension liability with measures of GDP as well as state revenues in their "Enhanced Financial Accounts," a component of the Z.1 Financial Accounts of the United States reporting.[12]

For both of these reasons, it appears sensible to formulate a standard comparison between a pension plan and the economic strength of its sponsor. There are a variety of measures of an economy available: gross domestic product, personal income, household money income. There is even a purpose-built metric for state tax capacity published by the US Department of Treasury, called "Total Taxable Resources," or TTR. This is an analysis of a score of different variables for each state, modeling the application of the different taxes used across the country.[13] The federal government uses it to equalize distributions of aid that depend on a state match, and it is explicitly meant to be a measurement of the tax capacity of that state.

These kinds of intergovernmental comparisons can only ever be roughly suggestive, but they will be more instructive if made on a common basis.

Though they differ in their particulars, trends in one of these measures are generally reflected in similar trends among the others. GDP and personal income are different concepts, but they follow similar trends over time. Even the TTR is not very different from the others.[14]

Because of the uncertainties that stem from Hamilton's observations, one must acknowledge that these kinds of intergovernmental comparisons can only ever be roughly suggestive, but also note that they will be more instructive if made on a common basis. Further, because the differences among these measures are not profound, it makes sense to choose among the various possible measures on grounds of convenience. Total Taxable Resources is available only at the state level, and while it is designed as an estimator of a state's tax base, it is only available with a two-year lag, the amount of time it takes the Treasury Department to collate and analyze the data. Because TTR is only available at the state level, unlike with simpler economic measures, it would be challenging to extrapolate to the county or city level.

Money income, published by the Census Bureau, is available at geographic levels as small as a census tract. It is a good estimator of a government's spending capacity or tax base, though of course this is somewhat indirect for governments that do not have an income tax and instead rely on property or sales taxes. However, though they are compiled from data across different survey series to be more timely, Census Bureau estimates take time to compile and are only available as two-year moving averages. The IRS also

12 https://www.federalreserve.gov/releases/z1/dataviz/pension/comparative_view/table/

13 See https://home.treasury.gov/policy-issues/economic-policy/total-taxable-resources. Also Compson, Michael L. "Historical Estimates of Total Taxable Resources for U.S. States, 1981–2000." Publius 33, no. 2 (2003):55–72. http://www.jstor.org/stable/3331187. and Kincaid, J. (1989), Fiscal Cap

14 See Appendix A.

publishes estimates of gross income for areas as small as ZIP codes, and these are available as annual estimates, but there is a significant lag due to having to wait for people to file their returns.[15]

Personal income, published by the federal Bureau of Economic Analysis (BEA), uses a somewhat more generous definition of income, for example including income received on behalf of individuals as well as income strictly received by them. Though it might be argued that because its treatment of taxable income is the same as nontaxable it is not as good an estimator of fiscal capacity, it is widely used to estimate economic strength. It is also only available down to the county level. However, the BEA makes it available quarterly for states and annually for counties, within a few weeks of the close of the relevant period.

To estimate personal income for areas smaller than a county, one can use the Census money income results to derive a relative proportion of income for areas within a county, and apply that to the BEA data. The proportion does change over time, so the scheduling mismatch between the Census Bureau and BEA means this is not a perfect solution. However, the fraction changes relatively slowly, so it can help us approach a consistent measure of an economy, such that trends over time are meaningful. It also permits comparisons between geographic regions of different sizes, and this approach is used in the results to follow.

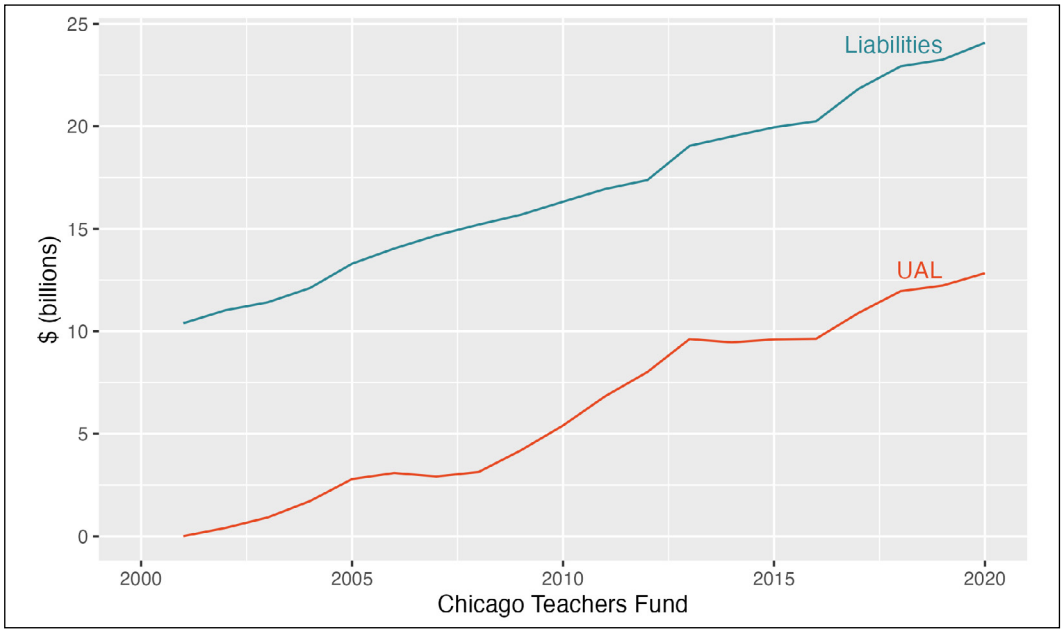
One obvious issue is that pension systems overlap. A citizen of Chicago is also a citizen of Illinois, so analyzing state and local plans individually will produce lower numbers than the aggregate data in the figure on page 10. We have not tried to correct for this in the current presentation, but this will help explain why even the outliers in the following graphs are at a lower level than the aggregate measure.

The pension liability shown here is to be paid out over several decades. It is hardly appropriate to portray a multiyear debt against a single year of income. Thus, the graphs that follow compare liabilities to the measure of income, projected over the next 30 years, with a 2.5 percent growth rate. This is an arbitrary choice of factor, since most of what is of interest in this metric is its progress over time, and in comparison to other governments.[16]

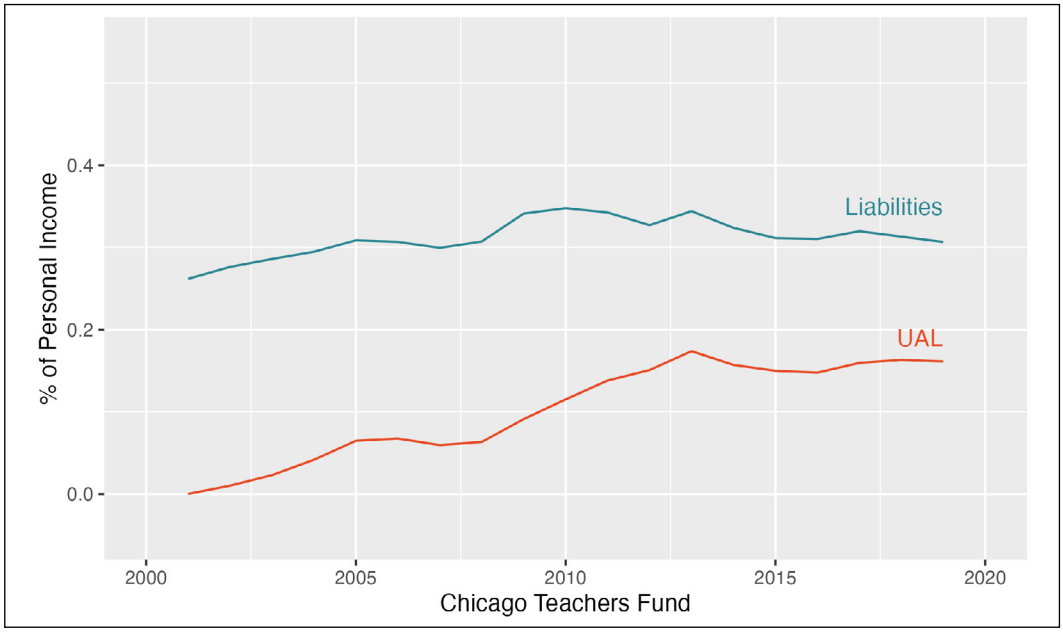
15 There is a useful discussion of several of these issues in the Census Working Paper, John Ruser, Adrienne Pilot, Charles Nelson, 2004, "Alternative Measures of Household Income: BEA Personal Income, CPS Money Income, and Beyond" https://www.census.gov/content/dam/Census/library/working-papers/2004/demo/CPS-BEA.pdf

16 This also follows the usage of Kahn, 2022, op. cit.

The image below shows the growth in total liability, as well as the unfunded portion, for the Chicago Teachers Fund.[17]



The image below shows the same information, but as a proportion of personal income.

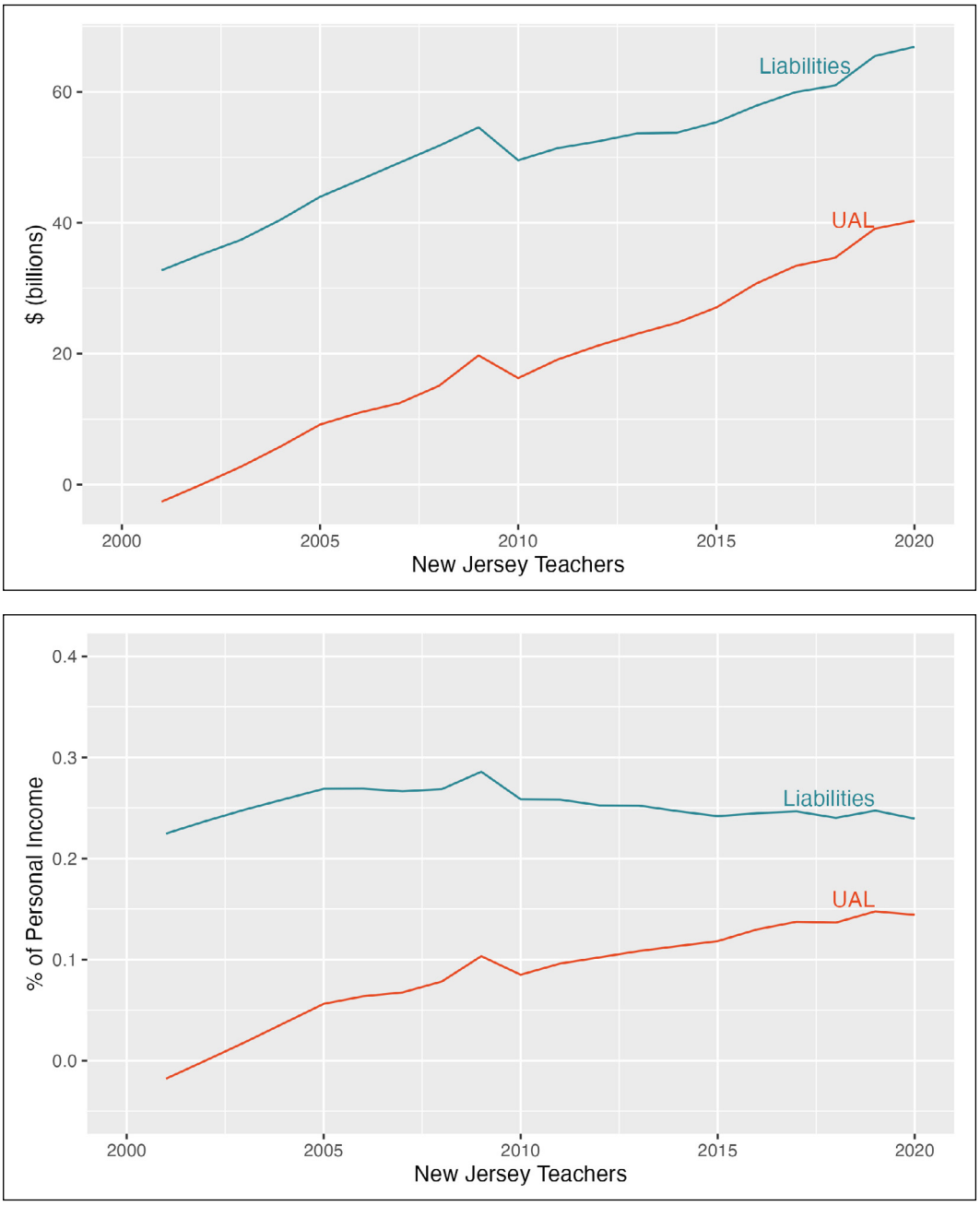


Since personal income has grown quickly, the growth of the liability does not appear nearly as dramatic. The unfunded portion of the liability has grown, implying that the real problem is less the growth of the liability than the lack of growth of the assets to pay for it. Indeed, a secondary merit of this image is that it makes it clearer that the ratio of assets to liabilities has declined from 100 percent to less than 50 percent over the time period shown.

17 Pension data here and throughout this report is from Public Plans Data. 2001-2020. Center for Retirement Research at Boston College, MissionSquare Research Institute, National Association of State Retirement Administrators, and the Government Finance Officers Association. https://publicplansdata.org

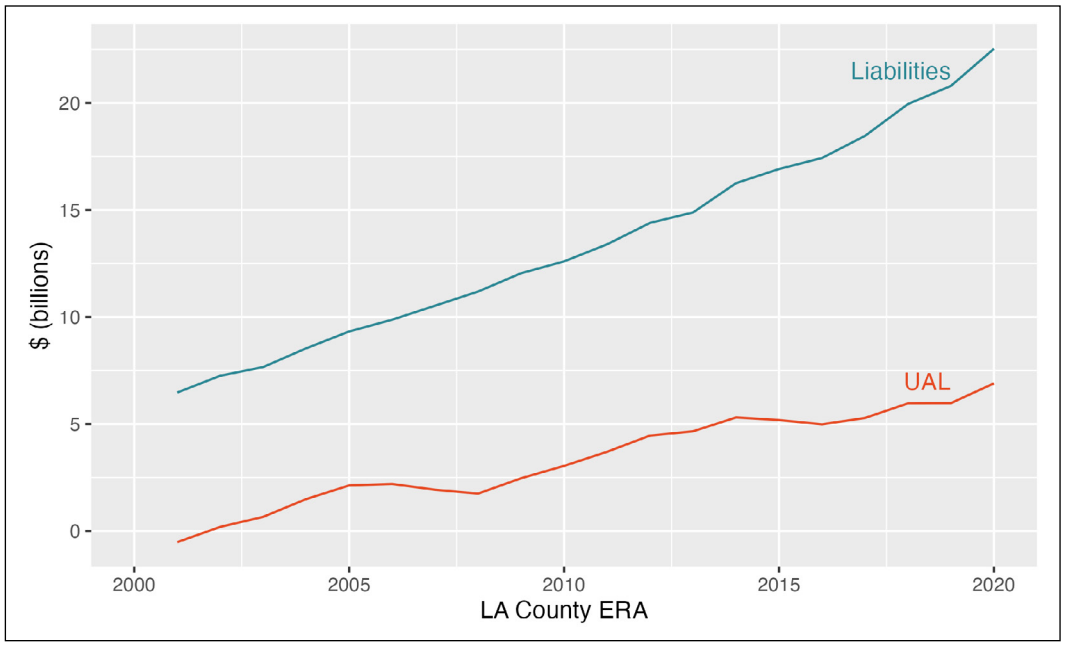
Again, plotting these values as a proportion of the state's personal income makes it clear that an out-of-control liability is not at all the main issue, which is instead the failure to keep up with it. In this case, the funding ratio has fallen from 100 percent in 2000 (when the UAL line was at zero) to less than 40 percent.

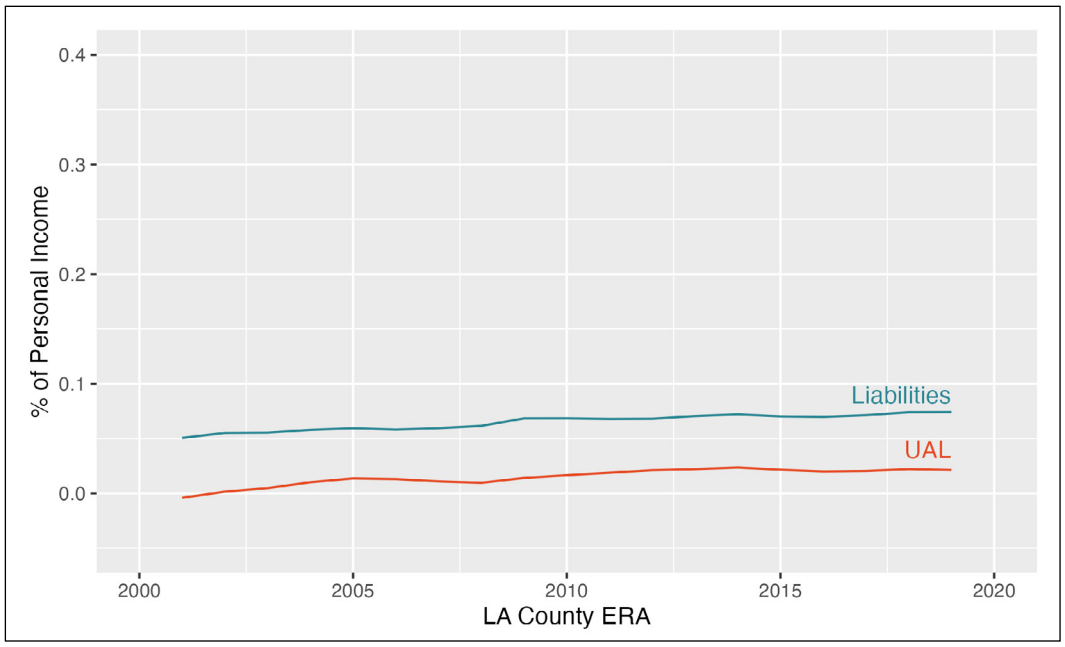
The following two images show the same data for the New Jersey Teachers' Pension and Annuity Fund. Again, the growth as a proportion of New Jersey's economic strength is much less dramatic, also implying that the problem is not merely the growth of the liability.



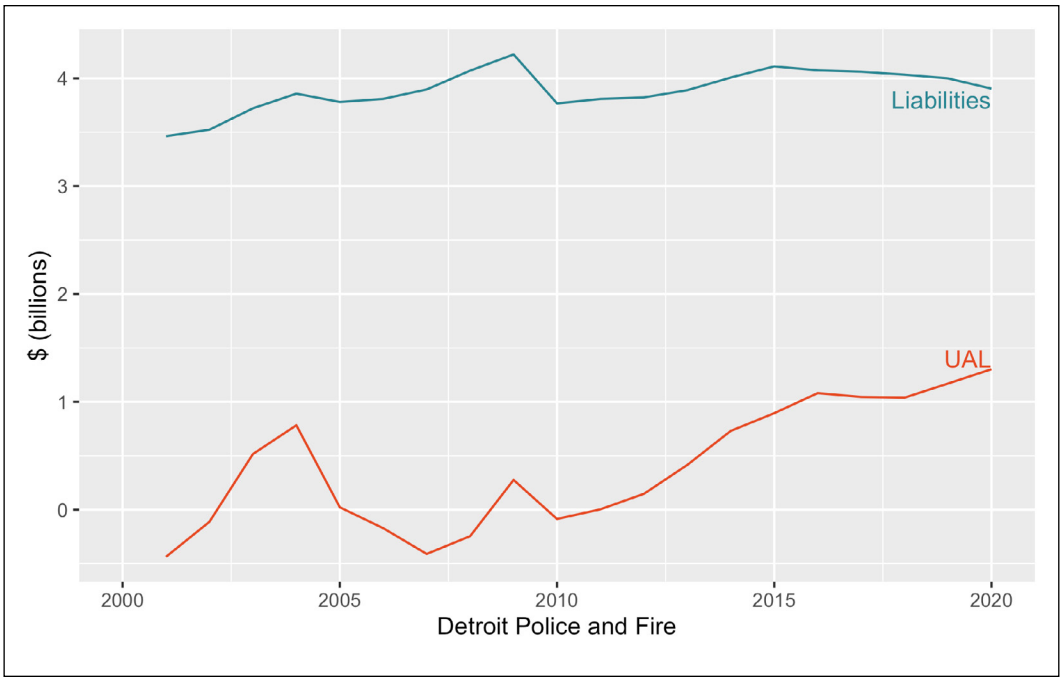
Mainly through neglect, the funding status of both the Chicago and New Jersey plans has decayed over the past two decades, but the funding gap is less than a quarter percent of the economic output of these areas.

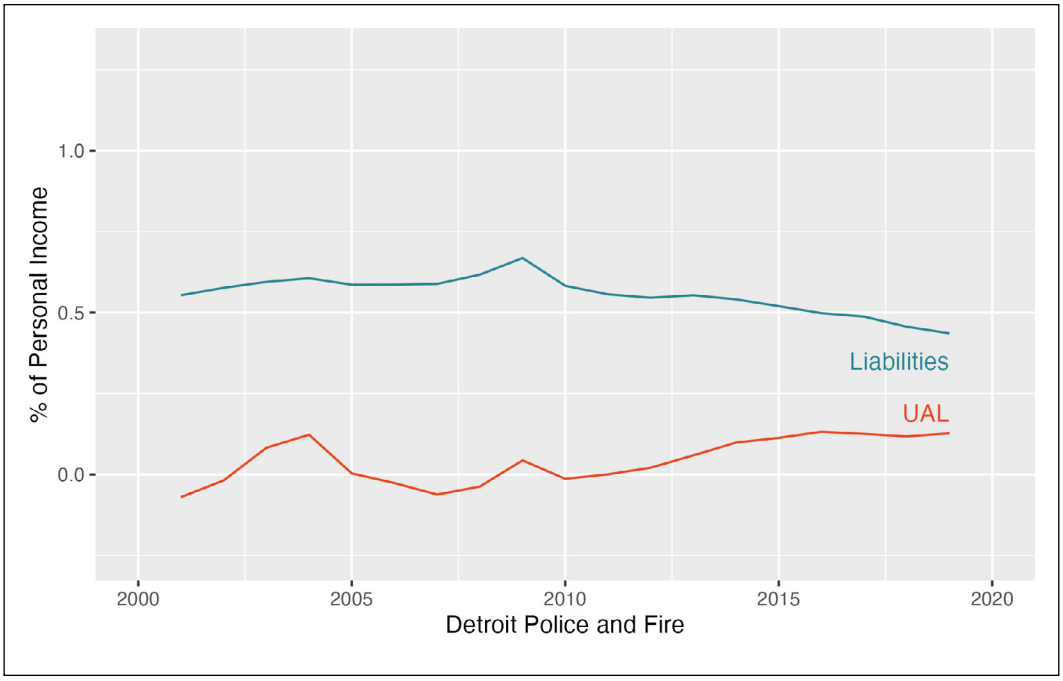
As the charts below show, the Los Angeles County Employees Retirement Association (LACERA) funding position has declined from 100 percent to about 70 percent over the period displayed, but though the new figure shows that the total liability has risen relative to the area economy, it is still a small fraction of the area economy. Though obviously one cannot use this high-level analysis to account for the preferences of voters and city officials, the implication here is that the economic strength of the taxable area can absorb the burden of the full liability more easily than can many other places.



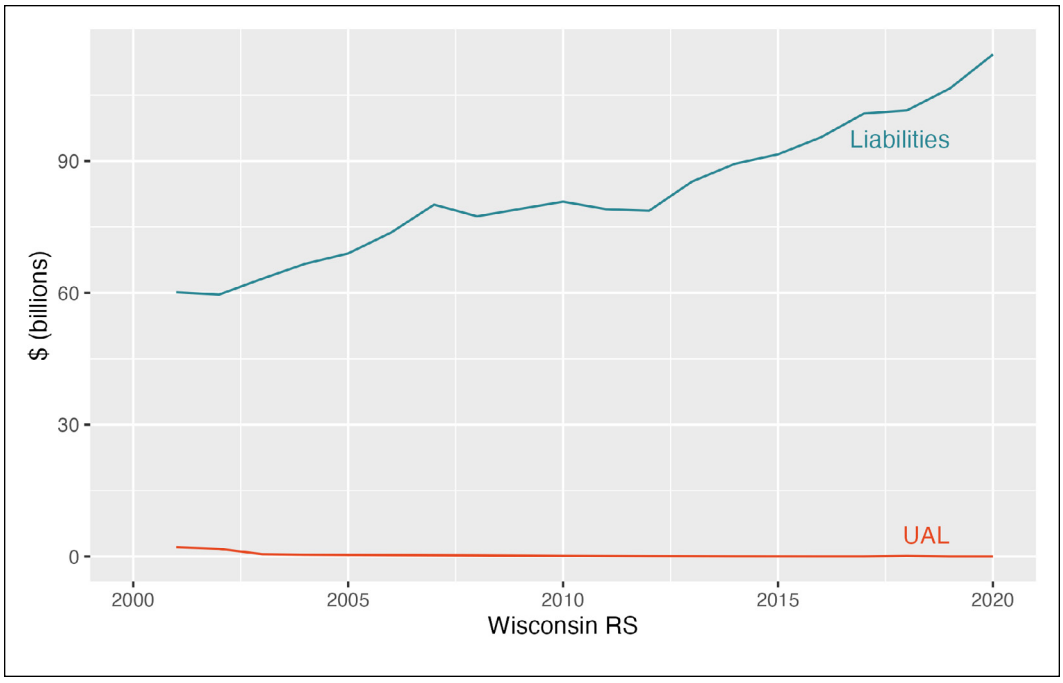


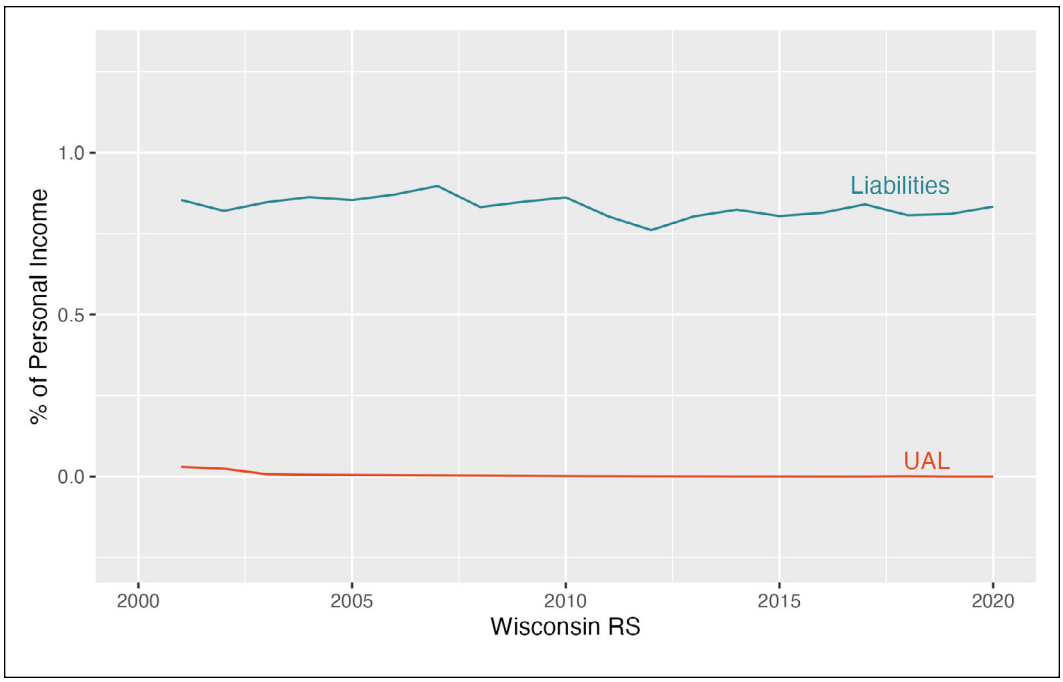
By contrast, the funded status for Detroit Police and Fire was consistently higher than that of LACERA during the study period. Even in 2013, on the eve of its bankruptcy, the funding ratio was 89 percent. However, the full liability for its pension systems represented a much larger bite of its economy and could be considered cause for concern. For example, the same degree of asset volatility experienced by LACERA and Detroit could produce very different outcomes, simply because the liability without the assets is so much larger relative to the city's financial capacity.





Perhaps it is surprising that Wisconsin's pension plan appears to be in a similar condition to Detroit's when scaled to the size of its economy, but it, too, has promised pension benefits of quite a substantial size in a relative sense – almost 1 percent of the total state personal income. Since the plan is fully funded, this does not appear to be a problem, and Wisconsin has unusually extensive risk-sharing provisions that will reduce the liability in hard times. However, the level indicates a high degree of reliance on the assets themselves, something plan policy makers should bear in mind.





The comparison between Wisconsin and Detroit also highlights the risk of comparisons. The state of Wisconsin's system is expansive, encompassing the majority of public employees throughout the state, and has those risk-sharing provisions, while the Detroit system serves only the municipal public safety employees. Pension plans differ substantially across many different axes, and comparisons between one plan and another must be made with great circumspection. Metrics like these can be revealing when comparing results from a single plan over different points in time, but are often no more than suggestive when comparing one plan to another.

CONCLUSION

The comparison between pension liabilities and the size of the economy governed by a plan sponsor can be revealing about a number of issues: whether a plan's benefits are "affordable," whether benefits or payments are more likely to be the cause of an underfunded system, and whether there is a path to sustainability. The measurement is a rough one, and will always be more suggestive than dispositive, but seeing its progress over time can be quite revealing about capacity and sustainability.

Every pension plan has – as an implicit and unvalued asset – the sponsor's promise to pay into the plan in the future.

This is especially important because every pension plan has – as an implicit and unvalued asset – the sponsor's promise to pay into the plan in the future. Those future payments matter, or rather the confidence the plan has in those future payments matters. A plan sponsored by a fiscally strong government is in better shape than a plan whose sponsor is facing hard times, even with the same liabilities and assets. Though future contributions are not supposed to be used to pay past benefits, they are potentially available as a financial cushion to weather market downturns and other volatility risks. A plan that can count on them can endure more volatility than one that cannot, and the fiscal capacity of the sponsor plays a role in that confidence.

A comparison to aggregated personal income is a convenient proxy for fiscal capacity and is also widely available for different levels of government. Though it is not readily available in the United States for areas smaller than a county, it is relatively easy to use Census Bureau money income measures to generate a version of personal income for cities, and the result is a consistent and easy-to-calculate measure that can be compared among different levels of governments.

DISCUSSION

The discussion within the Working Group about this measure was largely about how valuable it would be and what it actually means. Most of the members had already done or seen comparisons like this, so there was nothing unusual about it. There was widespread agreement that standardizing an economic capacity measure was a sensible step.

The primary point of discussion was around the implicit equation of economic strength with fiscal capacity as well as the choice of personal income as a way to measure economic strength. This has two components. First, the actual revenue derived by some governments may not be directly applicable to a measure of income. For example, comparing the LACERA liability to the personal income of Los Angeles might be misleading because the city does not levy an income tax, but instead derives its revenue from property and sales taxes.

However, as noted above, personal income is merely one of many ways to measure the size of an economy and wealth of a community. The best measure of property tax capacity would obviously be the amount of taxable property, but this would foreclose comparisons to governments that do not tax property and would demand an additional estimate of sales tax capacity and corporate taxes and so on. In the end, it seemed that uniformity was a worthwhile enough goal to tolerate some imprecision in the equation between personal income and fiscal capacity. After all, even if taxes are assessed via property and sales taxes, they are still paid out of income.

The second component refers to the basic objection that the question of tax capacity is ill-defined. Capacity is not only the ability to pay taxes, but the willingness of the populace to pay them. The connection between the size of an economy and the ability of the overseeing government to meet these obligations seemed to some to be unsubstantiated.

Another point that arose was whether questions of sustainability like this are properly a part of pension accounting or funding discussions. To some, these concerns belong to policy makers outside the pension plan: to the mayors, not the fund managers. But in a world where elected officials sometimes serve as pension trustees and fund managers advise legislatures, the delineation between the two groups is not very distinct.

The question of explicitly valuing a government's promise to pay in the future provoked extended discussion. Several people suggested that this would be a "phantom" asset. One example that came up was the city of Jacksonville valuing future tax revenue and recognizing it as an asset on the pension plan balance sheet.[18] But what was happening there was more akin to double-counting, since that future revenue would presumably have already been available to pay pension costs. The question of whether and how to recognize the value of an ongoing promise to pay pension premiums is simply asking why this promise–unlike virtually all others a government makes–is deemed to have no value in this context.

18 See "Pension Brief: Are Asset Transfers a Gimmick or a Sound Fiscal Strategy," Todd Tauzer, Todd Kanaster, Carol Spain, S&P Global Ratings, February 19, 2019.

UAL Stabilization Payment

Summary: *How large a payment is necessary to put a pension plan in the same funding position at the end of a year as it was at the beginning? The answer is the "UAL Stabilization Payment," or USP, a number that can be derived from past experience, assuming plan assumptions are met. The USP can be defined consistently and objectively, and that makes it a useful standard against which to measure actual policy. It is especially useful because there are few objective ways to evaluate policy with respect to cash flow from the balance sheet. It can also give a sense of how aggressively the actuarially determined contribution (ADC) will reduce unfunded liabilities.*

Many of the measures used to assess the condition of a pension plan focus on the balance sheet – the size of the liability and the quantity of assets – sometimes at the expense of cash flow considerations. But a plan's external cash flow, the difference between contributions and expenses, can be an important factor in a plan's survival. A plan whose sponsor cannot afford to keep its condition from deteriorating is in trouble, even if its balance sheet looks healthy.[19] This is more than theory; it was not distant liabilities that tipped cities such as Stockton, California, Central Falls, Rhode Island, and Detroit into insolvency, but the inability to maintain a positive cash flow.

Unlike the balance sheet, however, there is not a clearly defined cash flow standard against which to measure a plan's performance. Furthermore, like a plan's funding status, there are a variety of factors affecting the cash flow, only some of which have to do with policy choices. A mature plan with many retirees, for example, will be in a different cash flow situation than a younger plan where few have yet retired. The role of a cash flow standard is typically played by a statement of how contributions relate to the ADC. But without knowing what policy goal the ADC is calculated to achieve, this is not as informative as it seems.

Despite all this, it is possible to define an objective standard for external cash flow by simply asking whether a plan is in better, worse, or the same shape at the end of a year as it was at the beginning. The sponsor payment necessary to maintain a plan in its current financial position can be an indicator of changes in plan policy and underlying conditions, as well as a useful measure of exposure to risk for a plan sponsor. As a funding ratio is to a balance sheet, this stabilization payment is to a statement of revenue: a way to measure a fund's condition in a sense that is related to policy outcomes. This is not to say that maintaining the status quo is the best policy. We do not evaluate merit; we only note that it is possible to define it objectively. It can therefore serve as a benchmark for gauging policy and outcomes in the same way that people currently use the funded position.

We therefore seek to define a UAL Stabilization Payment as the payment necessary to leave a plan in the same condition at the end of a year as it was at the beginning, assuming the investment target is met. This is very similar to the concept of a "Tread Water" payment, promoted by Moody's,[20] though we seek to refine the treatment of liabilities by measuring an independent accrual rate for the total liability. Similarly, recent S&P

19 And cash flow is linked to investment risk, too. See the Risk Weighting and Other Shortcuts chapter on page 37.

20 See, e.g., "Pension Risks Growing for U.S. State and Local Governments," Tom Aaron, Moody's presentation to Southern Municipal Finance Society, September 2016. https://www.nfma.org/assets/documents/SMFS/smfs_pension_moodys.pdf

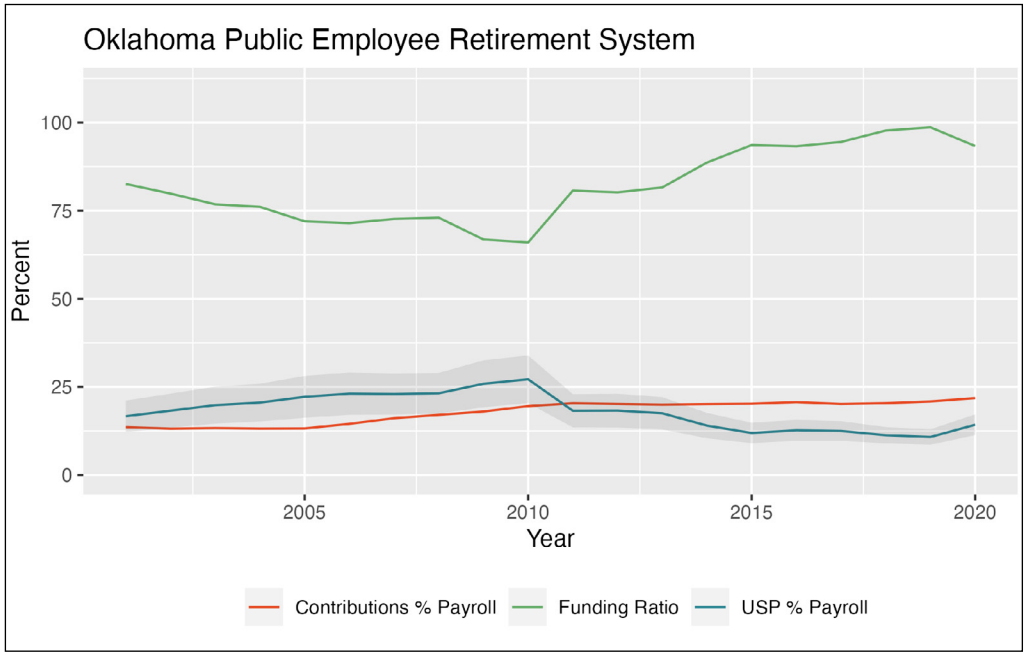
guidance on evaluating pension system health introduced the concept of "Minimum Funding Progress" (MFP) by adding 1/30 of the net pension liability to the Tread Water amount, to derive a payment that will leave a plan in slightly better condition at the end of a year than at the beginning.[21]

Details about how to calculate the USP are in Appendix B, including a discussion of what the data show about the rate at which the unfunded liability compounds and how the assumed rate of return is not a reliable predictor of the total liability accrual rate.

RESULTS

Using data from the Public Plan Database, we made graphs of USP versus payroll for the years 2001-2020, to illustrate its behavior, indicating with shading an estimate of the potential error in USP value. We selected plans in which our calculation of the liability accrual rate – using a combination of the benefit growth and payroll growth – was close to the assumed rate of return. The concerns about using the assumed rate of return as a predictor for liability accrual are significant, but they matter most at low funding levels.

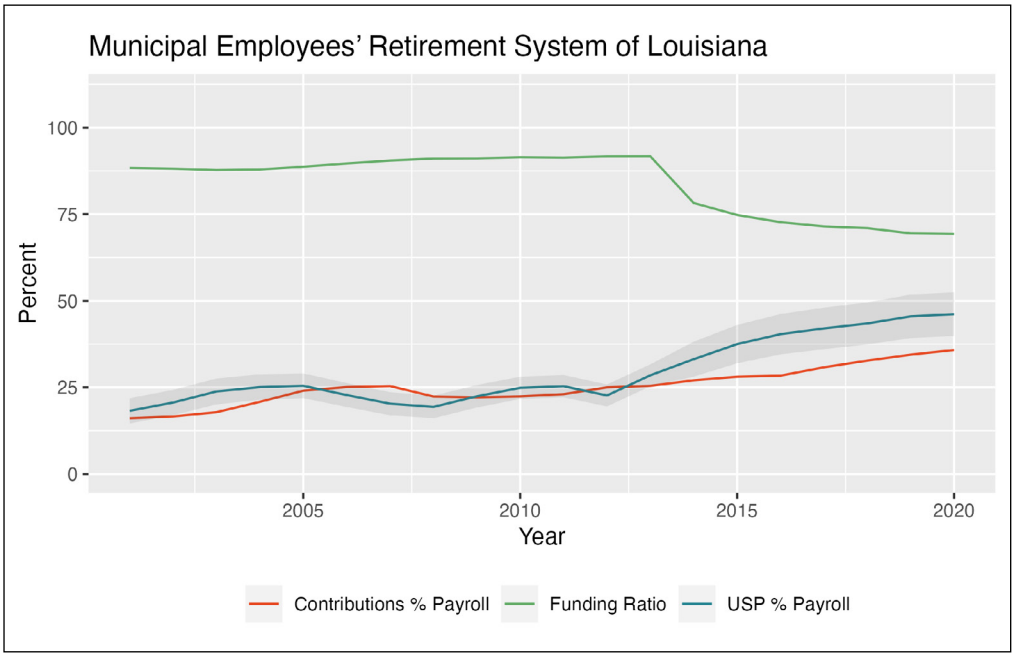
During the period that contributions in the Oklahoma Public Employee Retirement System (PERS) were below the USP, from 2000 until 2010, the PERS saw a steady decline in the funding ratio. In 2010, the USP went down, largely through a substantial decline in service costs, to a level below the annual contributions. As a result, the funding ratio began to climb again.



In the early part of this graph, contributions are below the USP, and so the funding ratio declines. In 2010, policy reforms reversed the situation and the funding ratio begins to climb again. The shaded area indicates the estimated uncertainty in the USP. See Appendix B.

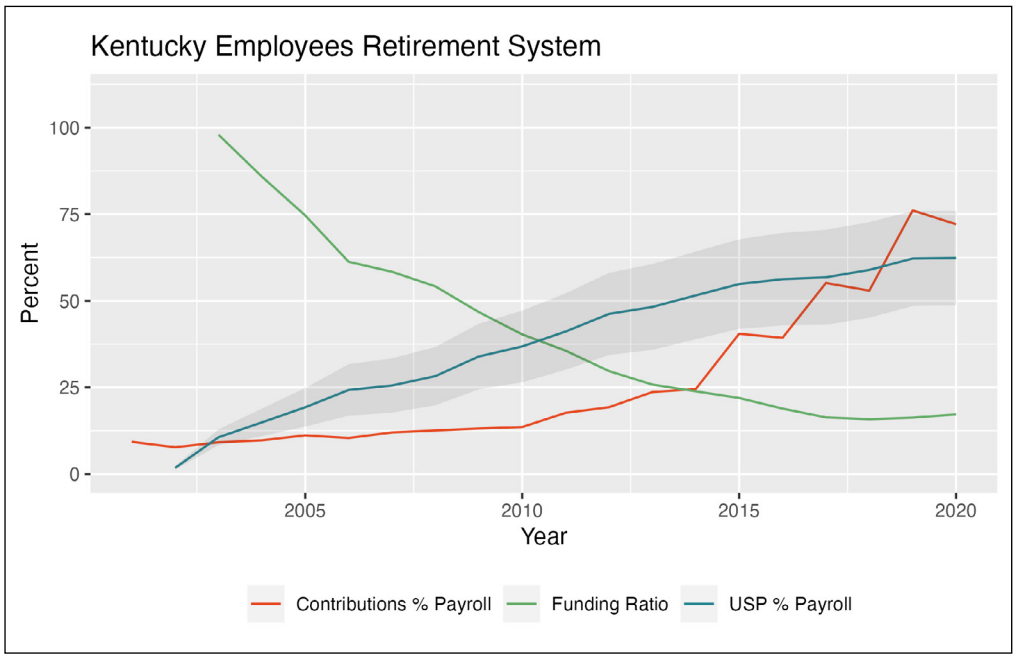
21 "Assessing U.S. Public Finance Pension and Other Postemployment Obligations for GO Debt, Local Government GO Ratings, and State Ratings," Carol H. Spain, Todd D. Kanaster, Eden P. Perry, and Robin L. Prunty, S&P Global Ratings, October 7, 2019.

In the Municipal Employees' Retirement System of Louisiana (MERS), one sees the opposite story. In 2012, a substantial rise in both service costs and benefits moved the USP well above the contribution levels, causing a decline in the funding ratio. The interesting thing about the MERS experience is that the system appears to have paid 100 percent of its ADC since at least 2014. That is, the ADC has been set well below the USP for several years. Again, one can know that a system made its ADC, but without knowing what policy is behind that ADC, it is difficult to know what conclusion may be drawn, even if the full ADC is paid every year.



In Louisiana, while the USP and contributions were approximately equal, before 2013, the funding ratio remained steady. When contributions fell beneath the USP, the funding ratio declined. Note above that as the funding ratio declines, the potential for error in a USP calculation using the assumed rate of return increases. This is indicated by the widening shadow around the USP line.

In Kentucky, contributions to the Employee Retirement System (ERS) were well below the USP for years, causing a steep decline in the funding ratio. Only when contributions climbed to surpass the USP did the funding ratio level off. Note that the potential error in the USP estimate changes dramatically, since it is lower at high levels of funding. The UAL Stabilization Payment can be used to assess a plan sponsor's employer and employee contribution to a plan, and can provide an objective yardstick with which to measure cash flow, comparable to using a funding ratio to assess a balance sheet.



Before 2014, contributions to the Kentucky ERS were routinely below the USP rate, causing a significant decline in the funded position. This decline also makes the USP more expensive. Once contributions climbed to the neighborhood of the USP, the decline in the funding ratio began to level out.

DISCUSSION

Some of the discussion around the USP and related "Tread Water" measurements was about how these provide different information than simply knowing whether a plan has made its ADC, since the naive observer may not know what the funding policy behind those ADC calculations might be. This is consistent with the observation above that the USP is an objective measure that is related to policy. It is not necessarily good policy or bad, but it can be objectively defined, and that makes it a useful benchmark.

The USP is an objective yardstick to measure cash flow, comparable to using a funding ratio to assess a balance sheet.

Suggestions that the USP should incorporate growth of the economy, per the Lenney et al. paper, provoked significant discussion. One comment was that such a measure should not be called a "stabilization" payment, since the UAL would only be stabilized with respect to the growing economy, but that both might be growing together.

Some discussion ensued over the evidence that the assumed rate of return seems to be a poor choice to use to discount the liabilities when calculating the USP. A frequent suggestion was that while the proper discount rate is the assumed rate of return, it is made inaccurate by changes not anticipated by the various plan actuaries. For example, employment growth is not incorporated into normal cost estimates, so ample employment growth will tend to make inaccurate the liability "roll-forward" equation on which the USP is based.[22] Some of these factors are discussed in Appendix B.

22 Of course, new hires may be adequately covered by the contribution rate. This is a comment about the accuracy of the roll-forward equation, not about how new hires will impact the system.

Stress Testing and Risk

Summary: *The computer simulation of pension results can be used for several different purposes, only one of which is "stress testing." The different uses have very different requirements for precision and initial data, so determining the purpose is important to satisfying the conditions for a successful exploration. We also consider a new classification of the kinds of risks faced by a pension plan, as an aid to formulating the questions a simulation might answer.*

Because the provision of pension benefits requires making commitments to be fulfilled decades in the future, there is a substantial degree of risk accruing to the entity that makes the promise. There is uncertainty to overcome in both the question of whether there will be funds enough at some date in the future, as well as the question of finding the least expensive way to provide those funds. Quantifying that risk is therefore important, both as a fundamental component of pension management, as well as a way to assess the wisdom of making the commitment in the first place.

Stress testing, the computer simulation of future circumstances under a variety of different scenarios, can provide a way to quantify what is otherwise a fairly indefinite question: How much risk is there, and what can be done about it?

Though stress testing is a vital tool, it is important to understand how to use it and how to understand the results.

Though stress testing is a vital tool, it is important to understand how to use it and how to understand the results. A useful analogy is to crash testing a car. One does not learn anything by crash testing a car at a walking pace, when the fender barely dents. However, neither does one learn much by smashing it into a wall at 200 miles per hour, when the pieces are shattered into bits. In a similar sense, a car's protective features are designed to address accidents one commonly sees in today's traffic. One does not learn anything about the performance of those features by crash testing the car into a biplane or a rhinoceros. To mean something, a stress test must simulate, both in scale and effect, a plausible set of the circumstances a plan might encounter.

CATEGORIZING MODELING

Summary: *Modeling can be used for several purposes, and being clear about the purpose can help define requirements for the model. We distinguish between stress testing, sensitivity testing, parameter search, and predicting the future.*

Part of the problem with stress testing is a certain blind-men-and-the-elephant quality to the enterprise. In the pension world, depending on who is speaking, "stress testing" can mean significantly different things, which have in common mainly the use of computers to simulate the performance of a pension system. Most prominent among these are the following elements:

- Sensitivity testing. Many quantities associated with a pension plan – the liability, the unfunded liability, various demographic averages – are the solutions to complicated differential equations whose dependencies are difficult or impossible to characterize with a formula. Using a computer model is a good way to test the sensitivity of one variable to another.
- Parameter search. System behavior is specified by some collection of parameters: the assumed and actual rates of return, the level of premium payments and payroll, the growth of payroll and population, and so on. One can think of such a collection as a point in some multidimensional parameter space. One can use a model to search for a point or a neighborhood in that space that satisfies some condition (e.g., minimizing payment volatility or maximizing asset accumulation) by evaluating the plausibility of outcomes resulting from different choices of points.
- Projecting a possible future. One can use a model to establish boundaries to probable outcomes based on policy choices. Of course, this is what actuaries have been doing since the profession was invented, but originally the techniques were only deterministic. Computers allow one to use stochastic techniques that are totally impractical without them.
- Stress testing. This entails forecasting possible results, given some simulated catastrophe. This is a general question about resiliency and sustainability, not a specific question about a possible future.

There are common features among these activities. For instance, in testing sensitivities, one might set some variable to a stress level to identify what depends on it. For example, one might hold down the expected return on assets for a long time to identify the exact dynamics of such an event: Which variables move first, which move later, which do not move at all? But unlike a stress test, this need not entail a claim that such a level is a probable or even plausible event. The manipulation of circumstances allows one to identify a connection between variables, but whether that connection could have real-world consequences cannot be inferred without further restricting conditions.

Depending on who is speaking, "stress testing" can mean significantly different things. But we can be more precise.

Similarly, one might predict possible futures under an array of different policy choices, and one might also stress the system under that same array of choices. But in the first case, one is asking, "What is likely to happen?" while in the second, "What kinds of things are likely to happen in case of a disaster?" These are similar, but different, questions. One is specific, the other general. One is making a specific prediction – conditional on policy choices – while the other's purpose is to produce general insights into system performance under stress.

Though these are different enterprises, they all require a way to represent the state of a pension plan and a way to propagate that state forward in time. After that, the different questions imply different requirements. To have confidence in a projection of future conditions generally requires confidence in the description of the present from which it is projected. In a complex world, detail is vital. For example, a prediction of portfolio value requires a detailed description of the components of an asset portfolio, because the various pieces

tend to move in slightly different ways. Without that detail, the prediction is more likely to go astray. On the other hand, testing the sensitivity of the funding ratio to the actual yield requires no such detail. In this case, the only requirement is likely that the proposed values remain in the range of non-absurd. The questions are different, so the requirements are different.

The requirements for initial conditions are necessary to give meaning to the results.

Given a representation of the state of a pension system and an accurate description of system dynamics, the following would hold true:

- Sensitivity testing requires little more than some rough boundaries beyond which variable values should be considered absurd.
- Predicting the future requires a higher degree of accuracy in its description of initial conditions. This usually entails more detail because of the internal dynamics of asset markets.
- Stress testing requires a scenario against which to test and a description of initial conditions only complete enough to meet the demands of the scenario. It also entails a claim that the scenario is at least plausible, that there is some kind of validation claim to be made.

Each of these requirements is necessary to give meaning to the results. A prediction of the future has no meaning if the initial conditions used bear an uncertain relation to the real world; sensitivity testing in a nonlinear system has no meaning if variables are set to implausible values far outside the normal range; and stress testing results have no meaning if the scenario is implausible.

In the background to these questions is a theoretical issue that has been lurking around discussions of computer models since the days of punch cards. In a 1969 paper, Allen Newell, a pioneering computer scientist and psychologist, described it as the "power/generality problem."[23] Succinctly, the better a model is at mimicking reality, the more limited the circumstances in which it can be applied. Though there is not a comprehensive theory behind this statement, observations from many different fields, ranging from cognitive modeling to weather forecasting, all point in its direction. It is possible to make a general model of hurricane dynamics to help understand the behavior of hurricanes in a range of scenarios, or to make a model of the weather to help understand how some specific hurricane is going to progress, but it seems impossible to do both at once.

For a pension plan, this limitation says that the better a model is at predicting the future for some given pension plan, the harder it will be to apply it to any other plan – including the same plan at a different point in time. An accurate simulation of circumstances is attainable for some given place and time, but the search for accuracy leads to a sacrifice in comparability.

This seems a less damaging outcome to the enterprise when one also considers that it is a complex world and that the probability of any specific stress test scenario occurring in the future is close to zero. The capacity of

23 Newell, Allen. "Heuristic programming: Ill-structured problems." Progress in operations research 3 (1969, reprinted 1993).

the universe to surprise us seems inexhaustible, and predicting what will happen far out on the fat tails of the probability distributions has eluded the best minds of the world for a very long time.

If scenarios cannot mimic the possible futures but only resemble them, perhaps that is an acceptable loss if it comes with a concomitant increase in comparability. The way to learn whether a policy has been effective in changing a condition detected by a test is to run the same test again after the policy is in place and has had time to have an effect. The way to learn whether some stress test results are meaningful is to compare them to other results of the same test, either from the same system or different systems. This is how testing works in medicine, for example, along with in engineering, environmental monitoring, public opinion surveys, and more. For any field in which testing is widely used, consistency is the key to assigning a meaning to test results. In some branches of medicine, there are even standardized patients on which to test therapies. Diabetes researchers, for example, can choose to use a certain strain of mouse, or a suite of numerically simulated standard patients, to seek approval for a second round of clinical testing.[24] This is not consistency for its own sake, but for the sake of comparability.

To advance the comparability of stress testing for public pension plans, what is needed is a widely accepted framework for such tests.

To advance the comparability of stress testing for public pension plans, what is needed is a widely accepted framework for such tests. This should include named example scenarios and standards for validating those scenarios, as well as agreement on how best to interpret the results of a test. With this kind of consistency, plan managers can learn from past tests, and potentially from other plans' experience as well. They can compare the results from one scenario against another and perhaps use multiple tests as an ensemble, if that seems revealing. They can use the results, in confidence, to adjust policy. All these possibilities depend on a degree of consistency. Without that, it will always be difficult to compare results of one test to another, even within a single plan's history, and thus always difficult to learn any lasting lessons from a test.

Two further considerations must be taken into account for any discussion of stress tests to be complete: politics and money.

Dramatic stress test results are often politically useful to those who believe that public pension plans are generally a bad idea. Without a well-defined methodology and standards for application and interpretation, stress test results can often be interpreted to imply certain disaster ahead. When results become fodder for counterproductive demagoguery, they do not help secure anyone's retirement.

Another important issue is that stress tests cost money. For a large plan, this cost may be negligible, but not all plans are large. A widely accepted framework for testing can help in two ways. First, clarity on what kinds

24 https://tegvirginia.com/software/t1dms/

of tests are suggested under different conditions will be helpful in economizing. Repeated stress testing may not add additional information if no policy changes have taken place, or if conditions have not changed enough to be worth the trouble.

The second way in which standardization can help is that a widely used framework for testing will help people learn from outcomes over time. Comparison of a system's results on a single test from one year to another will allow plan managers to determine whether policy changes make a difference, and will allow them to compare the results of a test with results from the world. When a pension plan adjusts its asset composition in response to some stress test, it is only an application of that same test that will reliably tell whether plan resilience has been improved.

Another possible outcome of experience with a test is that over time it may become apparent that some tests provide no more actionable information than some other metric that is much less expensive to calculate. For example, a comparison between benefit payments and an asset value discounted for risk might turn out to be as useful a measure of liquidity risk as a much more elaborate stress test. But the world will only learn such lessons if stress testing is comparable over time and from one plan to another. Only then will be it be feasible to compare stress test outcomes with some other metric in a meaningful way.

CATEGORIZING RISK

Summary: *Placing risks in categories can help analysts see commonalities between them, as well as differences. For example, there are several different ways that a pension system can miss its investment targets, but a proper categorization makes it clear that these can have very different mitigation strategies.*

The ultimate risk facing a pension plan is arriving at a month without having the money with which to pay benefits. But there are a number of different possible paths to that point – was it market conditions? tax revenue? political considerations? demographic changes? – and each different path suggests different tests to assess the risk properly, as well as different policies to manage that risk.

Two risks with the same result might suggest two very different solutions.

The American Academy of Actuaries has a practice note, ASOP 51, that contains a long list of the different risks associated with a pension plan.[25] One can take the project a step farther than those authors by suggesting categories into which those risks might be placed. There are commonalities to risks that can suggest common strategies to analyze or mitigate them, and those can serve as a basis for classification. This is not merely an exercise in tidiness, but can help in the construction of stress tests and the interpretation of their results as well as the management of the risks themselves.

25 https://www.actuary.org/sites/default/files/2020-07/ASOP_51_Practice_Note.pdf

There are certain risks that seem more related to the plan itself than to the sponsor, even if one stipulates that the debts and assets of a pension plan are the debts and assets of the plan sponsor. For example, the risk of a liquidity crisis that might be caused by assets remaining illiquid when they need to be spent, or inadequate controls on investment decisions, would seem to belong to the plan and its management rather than to the sponsor. Poor investment decisions are a different sort of risk than the risk of a down market, even if they both produce inadequate income. These are two problems that might have the same result but suggest a very different set of solutions. If the problem is poor decisions, one can change investment strategies or change strategists. If the problem is a down market, those are not as useful solutions. Distinguishing between risks to the plan and risks to the sponsor would seem to separate these two.

Distinguishing risks in this way produces three separate categories of risk.

Contribution risk, the possibility that the plan sponsor might not make good on its commitments to the plan, also seems to fit the category of a risk to the plan and not to the sponsor. Indeed, it is actually a risk to the plan perpetrated (willingly or not) by the sponsor, not at all a risk to the sponsor. This risk is most vivid in the case of a private pension system, where the plan sponsor might be liquidated and thus disappear one day, but it is also felt by public plans, whose sponsors are occasionally known to skip a contribution or two, or even several years' worth. Here then, are two categories of plan risk: operational and contribution.

Other risks in the ASOP 51 list – e.g., interest rates, asset values, liability growth – seem larger, so to speak, than these concerns. These seem more like risks directly to the plan sponsor, a different category. Beyond them, there is also political risk, that the expense of a pension system might be deemed too high, causing the sponsor to close or modify the plan. This is a risk as real as the other categories, but unlike them, it is a risk to the plan members, not to the sponsor, suggesting a third broad category

Distinguishing risks in this way produces three separate categories of risk:

- Risks to plan: These are operational risks, like mismanagement of funds or bad investments, as well as contribution risk. The operational risks could also be labeled internal and the contribution risk external.
- Risks to sponsor: These could be interest rate risks, investment risks, or demographic risks.
- Risks to plan members: These could be political risks that might cause a plan to close, or benefits to be cut.

The risks to the plan sponsor are the ones that best lend themselves to quantification via modeling and stress testing. Again, because of distinctions in how they are managed and detected, they might also be described as coming in three categories:

- Actuarial risk is when some actuarial prediction might be inaccurate. This includes such things as inaccurate mortality tables and underestimating payroll growth. It would also include misestimating the long-term rate of return on assets and unanticipated changes in plan benefits.

- Volatility risk is when the average values on which planning is based may be an inadequate characterizationof the environment the plan faces. This would include a sudden drop in asset values, even if such a drop had no impact on the long-term average because of a subsequent recovery. Similarly, a round of state employee layoffs under one governor would introduce volatility into the payroll growth record that may have impacts not reflected in the long-term average, even if the next governor hires them all back.
- Regulatory risk is a concept familiar to banking and insurance. There is no central regulator of public pension plans, but the concept can be applied to situations in which the accounting rules change or in which an actuary (or rating agency) changes its estimate of the relative importance of pension metrics. A change in the discount rate used to value the liability is an example. This differs from the risk of getting the rate wrong – diverging in both the consequences and the mitigation strategies – even though they are obviously related.

The purpose behind this kind of taxonomy is similar to the purpose behind the modeling approaches above. Categorizing the different risks helps clarify the common features of the different risks and suggests guides to their simulation and management. Two different actuarial risks, say an underestimate of retirement rates and an underestimate of mortality rates, will have similar effects. Both will increase the liability and both will be relatively slow-moving, also meaning they will take a relatively long time to diagnose. And both can be corrected by amortizing the differences over time. Volatility risks are different. They can occur more suddenly and be more severe, at least in the short term. Thus, preparation is a more important component of a management strategy for volatility risks, while actuarial risks can generally be addressed after they have been identified, over time.

Categorization can also help distinguish between risks that seem similar. For example, measures to protect against actuarial risk and volatility risk in investments are different, even though either risk might result in the same outcome: inadequate investment income. In fact, one usually assumes that decreasing volatility risk will decrease potential yield, so strategies to address these two risks may even conflict for some plans. Though they are both a kind of investment risk, and though both are a risk assumed by the plan sponsor, the difference in mitigation strategies implies that they should be considered different risks. Including the plan risk of simply having a bad investment manager, there are three different varieties of investment risk, all of which have distinct and possibly conflicting management strategies.

VALIDATION OF STRESS TESTS

Summary: *What gives any data point meaning is having another data point to compare it to, preferably more. Modeling results are no different. One might seek to acquire meaning by comparing modeled test conditions to real-world events, demonstrating that the data points are adjacent in some fashion to observations. Or one might run a similar test on the same plan at a different time, or on a different plan at the same time. Either way, it is the comparison that gives test results their meaning. Validating a stress test requires both justifying its plausibility and offering a basis for the comparison of its results.*

Just as numbers need context to become data, modeling results also need context in order to have a meaning. One way to provide this context is to consistently use the same test conditions, however outlandish the scenario. Context is provided, then, by comparison of one set of test results with another. For many purposes, a better approach is to do the same, but with a test that passes some measure of plausibility.

Just as numbers need context to become data, modeling results also need context in order to have a meaning.

Validation of stress testing scenarios is challenging. One can really only know with confidence that the probability of any given scenario actually happening is inverse to the degree of precision used to specify it. Baldly put, any scenario specified precisely enough to be of use in an extended asset-shock test will never happen. But that does not mean a certain degree of validation is not feasible.

For example, there are assertions one can make with confidence that can help to build stress scenarios, such as that small perturbations are more likely to occur than large ones. This is true in consideration of asset returns but also mortality deviations, inflation predictions, and any other distribution of random events that looks even vaguely like a normal curve. Consistent scaling of a plausible scenario is therefore a way to generate one plausible scenario from another, though one can scale down from historical examples with more confidence than one can scale up.

Another way to distinguish levels of plausibility is to note that some scenarios actually did happen while others find no historical justification. Scenarios that mimic actual events are set apart from superficially plausible scenarios that in reality cannot happen, perhaps for subtle reasons not captured by the data. This is a meaningful distinction and suggests that stress scenarios with a historical analogue will ultimately have more meaning than ones that do not, simply because they are more plausible.

As noted in the discussion of modeling categories above, it is important to acknowledge that the purpose of a stress test is not to simulate a future period in time. A stress test is not an exercise in soothsaying but an attempt to ask a "what if" question in relation to a plausible set of circumstances. Many different scenarios might be useful for an asset-shock test. Why not choose one that has actually happened, rather than risk testing on something that cannot ever happen for reasons that might elude notice?

Obviously, if one is interested in the effect of a shock on gross measures of a plan's assets, some detail of a historical downturn's dynamics may not be relevant. But if one begins with the known data and summarizes, one is in a more plausible scenario than a scenario that's made up from the beginning. Furthermore, consider a test like this, where a detailed set of metrics were compiled into a gross measure. If a test using the gross measure indicates that de-risking a portfolio would be merited, this would imply a follow-up that would profitably use the more detailed data that had been summarized for the first test. That way, one is asking the question, Give me more detail about this scenario, rather than inventing a whole new scenario and potentially confounding the results. Without that kind of consistency, how could an asset manager demonstrate that the de-risking would work to address the problem the first test detected?

The Federal Reserve uses a historical approach like this in publishing detailed stress testing scenarios for banks.[26] They enumerate "baseline," "adverse," and "severely adverse" scenarios, with enough detail to each one that bank examiners can apply the details to fairly complex asset portfolios. Sometimes, they also provide an "alternate severe" scenario. (There was one in place in 2021, to address possibilities related to the evolution of the COVID-19 pandemic.)

When tests have names, people can talk about them in a sensible way, understanding very specifically the scenario to which each refers.

The Fed scenarios are validated by a team of economists. The baseline scenario is their prediction of the likeliest future, while the others are about how that future might go south. Their policy is to apply historically validated trends to their baseline scenario (the "recession approach," as they describe it) to generate their adverse and severely adverse scenarios.[27]

In some ways, a bank is a fairly fragile enterprise – some of the risks a bank faces can be suddenly catastrophic – so the Fed is very specifically concerned that the baseline scenario mimic current conditions and how they are expected to evolve, and the adverse scenarios build off them. This, of course, presents problems when conditions change quickly, as they did in March 2020, when the Fed scenarios were suddenly rendered irrelevant by the onset of the COVID-19 pandemic. For a pension plan, many of the important events tend to unfold more slowly, and so the always-current nature of the Fed's baseline scenario is not as significant for a meaningful pension stress test. For pension stress test purposes, the important parts of the Fed framework seem to be that they are validated by historical analysis and they have names by which to refer to them.

There is nothing exclusive about these tests. Banks can conduct other tests if they want to, or if their shareholders demand it. In fact, if they do business in other countries, they probably have to do testing under those regimes, with other rules. But the Fed framework provides a consistent and readily comprehensible way to compare one bank with another, and also to compare one bank to itself at a different point in time.

For public pensions, it is not necessary to demand that all stress tests use one of a small number of example scenarios. The landscape is too heterogeneous, and nobody likes to be told what to do. However, the benefits of consistently using a set of widely adopted scenarios will likely be obvious enough that trustees and policy makers will suggest them if they can be developed, named, and publicized widely.

The names are not an unimportant detail. A name suggests a specification that a practitioner can look up, while the lack of a name is only an invitation to do something similar. The Fed tests have names so people can talk about them in a sensible way, understanding very specifically the scenario to which each refers. Using the Fed's Severely Adverse scenario means looking up what that means, while simply saying that one used a severely adverse scenario could mean any number of things.

---

26 https://www.federalreserve.gov/publications/dodd-frank-act-stress-test-publications.htm

27 https://ecfr.federalregister.gov/current/title-12/chapter-II/subchapter-A/part-252, but more helpfully, see https://www.federalregister.gov/documents/2017/12/15/2017-26858/policy-statement-on-the-scenario-design-framework-for-stress-testing

Another important conversation to have about a test concerns the issue of stochastic simulation versus deterministic modeling. The real world is full of random variation, so a realistic simulation will always have some kind of random component to it. The risk, though, is that one can over-randomize. Many real-world variables have connections that cause their randomness to be in sync, or in counter-sync, to other variables. Stock prices and bond prices are the classic example of trends that tend to counter one another, but there are many others relevant to pension simulation. Longevity and health improvements are likely not unrelated to retirement decisions, for example. Similarly, a collapse in asset prices is likely to be related to a collapse in tax revenue, and perhaps vice versa.

There are degrees of determinism. One can think of following a given stress-testing scenario as a deterministic exercise, in contrast to running a stochastic model. But this is misleading, since a well-specified scenario will have some of the historical random variation built into its record. Following the recorded asset price variations through the quarters of the period 2000–2002 may seem deterministic, but it is not as deterministic as simply setting an average rate of return over a period of years.

TESTING PURPOSE

Summary: *Stress testing should have a purpose: a question whose answer they are meant to provide, or at least indicate. Formulating the question, while being mindful of the different categories of risk and testing, will define what kind of simulation is necessary and what kinds of initial conditions are required to provide or indicate an answer.*

A test should be structured around a question, or set of questions.

The question for which the Federal Reserve bank stress test scenarios were developed is very specific: How well will some given banks perform over the next three to four years? How risky will those years be? Because economic conditions are constantly changing, the Fed scenarios are regularly updated to keep up with the changes. Fed economists attempt to maintain a constant, though subjective, level of severity for the adverse scenarios, but the scenarios they publish next year are always going to be different from the scenarios this year. One can use them to compare the performance of one bank to another, but it is more challenging to use them to compare the performance of a bank this year to the same bank next year. They are comparable only to within the degree one trusts the consistency of Fed economists. This is not a shortcoming, but is simply a reflection of the questions these stress tests were designed to answer.

This, of course, is the point of any well-planned pension stress test: to identify the factors, or to choose among the many factors, that could plausibly cause a fiscal strain on the plan sponsor. According to ASOP 51, a test should be structured around a question, or set of questions. The Fed scenarios are structured around a set of questions having to do with bank performance over the near future. If a pension plan wants to ask those questions, perhaps varying the policy choices to ask questions about their suitability, the tools are right there.

However, the Fed scenarios might not be the best to use if the question is, "Have plan policy changes reduced risk since the stress test we did five years ago?" In that case, the plan should apply a test as similar as possible to the one used five years back. Without that kind of consistency, there can be no confidence in a comparison of results.

Similarly, the Fed tests are not that helpful if one is asking, "How likely is a particular outcome over the next few years?" The Fed framework is deterministic, so is only going to offer the best guess of the Fed economists about how things will turn out in the short term.[28] A question of assessing probabilities calls for a stochastic approach instead.

Again, a stress test or other modeling exercise should be structured around a question or set of questions. In that respect, the tests may be as varied as the questions. But some questions are very commonly asked, and scenarios designed to help answer those questions, named in a way that makes it easy to refer to them, will go a long way toward creating meaningful risk evaluations.

DISCUSSION

An early point of discussion of stress testing had to do with whose responsibility the tests actually are. The point was made that stress tests are mainly used to assess risk to the sponsor, while the cost of the tests is borne by plans. For large plans, this is barely an issue, but not all plans are large. Being clear about the taxonomy of risk makes it clear that in some cases the plans are being asked to finance tests measuring exposure of the sponsors.

Frequently, the point was raised that plans are very heterogeneous, and that a standardized set of stress scenarios may not be relevant to a particular plan. This is true, but the point of standardization is not to force some plan to do a stress test that is not appropriate, but to give more information to the plans that do use that test. The point is to have tests from which one can learn over time, by making the results comparable.

One comment suggested that "for plan design decisions, showing projected normal cost across a range of future return scenarios (incorporating any risk-management provisions included in current or proposed plan design) would be valuable. Pennsylvania's Independent Fiscal Office did something that included this for their 2017 reforms." This does sound like a useful exercise, but according to the classification of modeling exercises, one might call this sensitivity testing instead of stress testing.

Another comment about stress testing suggested that a sensible framework for stress tests should be clear when such a test is warranted. Focusing on answering a specific question seems a good policy, and if conditions have not changed since the last year such a test was done, perhaps the question was answered already and does not need to be answered again.

---

28 Though they do include a market volatility index, so one can potentially use their data to manufacture a stochastic model. https://www.federalreserve.gov/newsevents/pressreleases/files/bcreg20210212a1.pdf

"A primary or overarching objective of a stress test should be to identify the factors that would result in an untenable or unsustainable plan condition, chiefly the inability of the plan sponsor to pay promised benefits, but also, perhaps, a contribution threshold the plan sponsor is unwilling to cross. Identifying and measuring such objectives would serve two core purposes. It would (1) require the plan sponsor to identify what level of plan cost is unacceptable; and (2) identify individual and collective scenarios – market performance, insufficient contributions, benefit levels, and so on – that are projected to result in an untenable or unsustainable plan condition, which would enable the plan and its sponsor(s) to take actions intended to avoid such a situation."

"I would agree that this is a key purpose of stress testing. However, even for jurisdictions where this kind of analysis will show no risk of an untenable or unsustainable outcome, a stress test can be useful for planning (i.e., how much would employer contribution rates rise in a recession?) or policy choices."

"Stress test reports should contain language clarifying the likelihood of each scenario actually occurring, emphasizing that a stress test is not a projection and advising that results of stress tests should be interpreted in this context." This is exactly the point of validating a stress test scenario, as well as the point of distinguishing between a stress test and a projection.

Risk Weighting and Other Shortcuts

Summary: *Numerical simulations are a fantastically useful tool, but they can be expensive to conduct. Furthermore, testing is in some ways a backward-looking enterprise, even for projections of the future. That is, one conducts the test and sees the results. For a complex system, it is not always obvious how to inform decisions with those results in a way that will improve outcomes, without conducting the test all over again. Experience with testing in other disciplines has led many to conclude that simpler observations can provide the same value, with the further advantage of providing a metric one can "manage to." We present a method to risk-weight asset values for a pension plan as a possible companion for certain asset stress tests, and ultimately perhaps as a substitute for them.*

One learns more about any tool with use, and the same will be true of stress testing, if some degree of consistency can be asserted. As an example from the Fed, regulators have evidently learned from experience that their Adverse scenario was not very revealing, compared to the Severely Adverse, and so it is apparently not published anymore.

Stress testing, the numerical simulation of a disappointing or even catastrophic funding scenario, is one way to estimate the risk facing a pension plan. But there are other, potentially easier, ways to measure the same kind of risks. Under the Basel banking rules, for example, banks use a risk-weighting algorithm to estimate the value of their capital under adverse conditions. Some assets, like cash and high-quality bonds, are valued at face value, while others are discounted according to their risk. In the Basel III rules, some assets are discounted 100 percent.[29]

> Many fields use a simple metric weighted toward poor outcomes in place of a more complex estimation technique.

Other fields take a similar approach, using a simple metric weighted toward poor outcomes in place of some more complex estimation technique. Testing soil for drainage is a simple measurement, but it is typically done only once, during the wettest season, in order to measure a worst-case scenario. The worst-case number is the important value, and the change of the measurement over the course of an annual cycle is considered a detail, even though other factors, such as the duration of the high-water mark or the depth of the low-water mark, are also relevant to good drainage. Flood planners will typically use the water height alone as a metric for flood impact, ignoring other measures like inrush force and instantaneous pressure. And maximum wind strength, rather than the dynamics of wind direction and duration, is generally used as a simple metric for hurricane damage.

Perhaps more directly applicable to a discussion of pension risks, the concept of value at risk (VaR) was developed as a quick way for investment managers to perceive the downside risk of their portfolio. Like the

29 Bank for International Settlements, "High-level summary of Basel III Reforms," December 2017. https://www.bis.org/bcbs/publ/d424_hlsummary.pdf

others, it is a shorthand way to measure the risk of losses, using a quick look at variance instead of a detailed model of the circumstances.

In each of these cases, the more complex measure would provide more information than the simpler one, but the added cost and complexity has been deemed in the relevant field to be not worth the trouble. This may well turn out to be the case for pension stress tests if other simpler metrics can be found that indicate similar conditions for less trouble.

Metrics of this sort are already in place in the pension world. The display of the sensitivity of the net pension liability to choice of discount rate, mandated by GASB 67 is, in a sense, a presentation of the effect of lower-than-anticipated investment returns, exactly what many stress tests seek to uncover. It is not hard to find other examples. Here, for example, is a presentation of the possible effect on the contribution rate of a 10 percent loss in assets, given different values of the Asset Volatility Ratio (asset value divided by payroll).[30]

AVR	Unsmoothed Amortization	Unsmoothed Amortization
5.0	2.94%	0.59%
6.0	3.52%	0.70%
7.0	4.11%	0.82%

This table was used to present the possible outcomes of the Teachers Retirement System of Georgia experiencing a 10 percent loss of assets, given different values of the Asset Volatility Ratio (asset value divided by payroll).[30]

This is a simple table presenting the predicted outcome of a specific catastrophic scenario under different conditions. In other words, it is an easy-to-calculate table that may turn out to be as revealing as a number of possible stress tests. However, one can only learn whether that is the case by asserting some consistency to the practice of stress testing.

RISK-WEIGHTING PENSION ASSETS

Summary: *All investment assets carry some risk, but not all investors are equally able to weather those risks. We seek to define a measure of risk that takes into account the external cash flow of the pension plan as a measure of how resilient the plan is against asset shocks.*

Many stress tests are devoted to the quantification of investment risk. To construct a simpler measure of investment risk, one can begin with the observation that a plan that cannot avoid liquidating assets during

30 Teachers Retirement System of Georgia, 2020 valuation report, page 14.
https://www.trsga.com/wp-content/uploads/GA-TRS-6-30-2020-Valuation-Report-FINAL.pdf

a temporary downturn is in a position of greater risk than a plan that can wait out the bad times. Positive cash flow – when employer and employee contributions are enough to pay current benefits – therefore provides some insulation against investment risk, and the more negative the cash flow, the greater the risk that temporary losses will be locked in by the need for liquidity.[31]

A system with a weaker cash flow position is less able to afford investment in riskier assets.

This suggests that a system with a weaker cash flow position is less able to afford investment in riskier assets. It also suggests that the appropriate measure of risk would take into account the investment horizon. A positive cash flow position can endure short-term volatility while waiting for the long-term payoff, while a negative position may need less volatility in the near term. Second, consider that the variance around the long-term trend of most investment classes will tend to decline over time, as the ensemble of random monthly growth increments regresses to the long-term mean for that investment class. This is nothing more than a consequence of the central limit theorem of statistics.[32] As a result, most risk measures, such as volatility, or VaR, have a time duration, whether implicit or explicit.

A discount applied to an asset based on the statistics of the variation is a very similar idea to the standard VaR calculation, but instead of asking, "What is the maximum amount I am likely to lose over this period?" it is asking, "What is the effect of a not-implausible stretch of bad luck?" For convenience, we choose a 1-in-5 chance, partly because losing a bet at those odds is reasonable, and because that is roughly the likelihood of a loss event occurring more than one standard deviation from the mean.[33]

We propose, then, to discount the value of an asset by a weight interpolated between a short-term standard deviation and a long-term one, according to the plan's cash flow. A positive cash flow will use the long-term discount, and a very negative cash flow will use the short-term one.

Since part of the point of doing risk weighting is to assess the level of risk taken on in the portfolio allocation, it would make sense to assess the risk of asset classes. For banks, the Basel III framework contains fixed assessments of the risk of different asset classes, and its point is not to make an explicit prediction of what will happen, but to distinguish between assets that provide long-term security and those that do not. A pension system has a different set of priorities than do the loss-resilience needs of a bank, but a similar principle, using preset discount values for different asset classes, might be appropriate.

This provides a risk-based measure of asset value. But again, in the analysis of data, a number means nothing in isolation. Once again, meaning is only acquired by a number when it is compared to another number. That might

31 This is also related to the question of whether or how to value the promise to pay in the future, discussed in Chapter 2, since the value of that promise is also relevant to whether a plan may be required to liquidate assets in the event of some future downturn.

32 There is a school of thought that says that investments cannot be assumed to be normally distributed in the long term, and certainly the available evidence for private equity and hedge fund investments cannot refute the claim. But it remains a useful approximation to experience, and thus useful for the purpose of risk assessment. See the example below and Appendix C for more.

33 For a normally distributed set of events, there is about a 40 percent chance of an event occurring more than one standard deviation away from the mean. For investments, this means a 20 percent chance of gains larger than that and a 20 percent chance of losses larger than that.

be a comparison to the same measurement at a different time, or it might be a comparison to another value measured in similar conditions. As the example below shows, it is interesting to compare risk-weighted assets with the unweighted assets as a gross measure of portfolio risk. But one can also compare risk-weighted assets to benefit payments, to make what might amount to a worst-case liquidity measure. The unweighted version of this comparison is already often used as a rough indicator of a plan's prospects and the relative importance of the assets to making payments. What follows is a highly simplified example of how such a risk-weighting might work.

The table below shows an assortment of plans from the Public Plan Database, with a range of cash flow positions, from the positive to the substantially negative. The table shows annual benefits and contributions, along with net cash flow exclusive of investment income, as well as the benefits, both as a percentage of assets. (Dollars are in millions.)

Plan Name	Assets	Benefits	Contributions	Cash Flow	Assets/ Benefits
Detroit Police and Fire	2,601	291	47	-9.39%	8.94
New Jersey Teachers	26,583	4,478	3,138	-5.04%	5.94
Rhode Island ERS	6,509	809	557	-3.86%	8.05
Illinois Teachers	54,891	6,927	5,901	-1.87%	7.92
Arizona Public Safety	8,725	912	1,096	2.11%	9.57

Again, a plan with positive cash flow can afford to wait out a temporary investment downturn or a momentary loss of value by some asset, while a plan with a significant negative cash flow might be forced to liquidate an investment at its nadir. Thus, we can discount the asset values by the short-term or long-term standard deviation of the investment value, and we can scale the discount by the external cash flow's position in the range of 0 to -10 percent. (Positive values are set to zero, and negative values are limited to -10 percent.) For example, a net cash flow of -5 percent of assets would put the discount rate halfway between the minimum and maximum values for that class of asset.

From their annual reports, this is the rough allocation of assets for these plans. (We use an artificially small number of categories for illustration purposes. See Appendix C for a more complete set of asset categories applicable to most investment portfolios.)

Plan Name	Fixed Income	Equity	Real Estate	Other
Detroit Police and Fire	29	38	13	20
New Jersey Teachers	30	46	11	13
Rhode Island ERS	31	44	9	16
Illinois Teachers	24	52	16	8
Arizona Public Safety	19	38	5	38

These are the estimated standard deviations of investment returns. The "other" class is highly discounted in the near term because many investments in private equity or hedge funds are not liquid at all in the near term. Real estate investments are also generally illiquid in the short term. (We assume that "real estate" in these portfolios translates to private real estate investment vehicles. This may not be accurate in all of these cases, but again, this is a simplified example, to show how the risk-weighting is to work.)

Asset Class	Short-Term Discount	Long-Term Discount
Fixed Income	6.72%	1.91%
Equity	17.0%	2.1%
Real Estate	99%	3.14%
Other	99%	1.71%

Discounting the assets in these plans according to these weights and the allocations gives the following weighted results for both the assets themselves and also the benefit coverage of the assets. The first column specifies the percent of assets that will be valued with the short-term discounts; the others will be discounted by the long-term value above. See Appendix C for more details.

Plan Name	Short-Term Discount	Assets	Weighted Assets	Assets/Benefits	Weighted Assets/ Benefits
Detroit Police and Fire	93.9%	2,601	1,543	8.94	5.30
New Jersey Teachers	50.4%	26,583	21,520	5.94	4.81
Rhode Island ERS	38.6%	6,509	5,522	8.05	6.83
Illinois Teachers	18.7%	54,891	50,166	7.92	7.24
Arizona Public Safety	0%	8,725	8,553	9.57	9.38

The table shows that the asset coverage numbers are rearranged to some degree. For example, the assets for Detroit Police and Fire have gone from covering nine years of benefit payments to six, and the New Jersey Teachers has gone from covering six years to five. The Illinois Teachers fund, by contrast, has only declined from 7.92 years of coverage to 7.24, due to a more conservative portfolio.

What this measure is doing is akin to simulating a bad investment scenario over the next few decades, roughly what an asset stress test might do, at greater expense. Not only is this a possible substitute for the more expensive test, but this would be a metric to which policy makers can manage. That is, they can see immediately the effects of portfolio composition choices on the metric, and choose investments accordingly.

DISCUSSION

The discussion about risk weighting centered on what is the best measure of risk for some asset: Expected volatility? Expected downside volatility? VaR? Liquidity risk? What's most likely workable in cases where the portfolios are quite complex and/or make extensive use of derivatives? The measure outlined here tried to adapt well-known concepts like value at risk (VaR) to the realities of managing a pension plan.

Another subject of conversation related to risk-weighting assets was about what to do with the resulting number. The obvious answer would be to compare it to liabilities, but because such a measure would by definition produce a lower asset value than is currently the case, political considerations make adoption of risk weighting difficult in the estimation of the funding ratio, even though it might be a helpful representation of a plan's risk position.

Standardization of Reporting

with Jim Link

Standardization of the reporting of pension results allows managers and observers to see results in a timely way and can also expand the range of indicators. That is, the important factors to know about a plan are not just measurements of its condition but also the actions that are underway and the policies in place. We propose a standardized "scorecard" that encapsulates not only the condition of a pension system, but which can also include metrics that accommodate these somewhat more abstract features of a plan's management.

A troubling feature of managing a pension system is the time lag between actions and their consequences. A choice to raise benefits or lower contributions may take years, even decades, to have a perceptible effect on the gross measures of system health. Even a disastrous policy choice may have no repercussions obvious to the non-expert for years after it is made. The opportunities for a kind of policy moral hazard – actions with no consequences – are significant.

A choice to raise benefits or lower contributions may take years, even decades, to have a perceptible effect on the gross measures of system health.

The usual reply is that choices to overpromise a benefit or underfund a pension will affect a government's bond rating in the future. There are at least two problems with this assertion. First, it is often the case that the systems in the most trouble have government sponsors whose bond rating is already poor. A city with an already terrible bond rating need hardly fear the threat of further downgrade. This, of course, is only relevant to a small proportion of pension systems. A second, possibly more significant, problem is that bond ratings are complicated judgments about a number of different fiscal and economic issues relevant to some government. Pensions are only a part of the equation, and this clouds any kind of causality there might be between pension policy decisions and bond rating changes. Even if a pension policy change results in a bond rating change, the connection is tenuous and uncertain. This is all to say that the link between policy choices and bond ratings is just too indirect to be a useful form of policy discipline.

To a certain extent, this seems to have been among the motivations behind the changes between the GASB rules of the 1990s and their replacements, GASB 67 and GASB 68. Requiring governments to acknowledge the net pension liability on their balance sheets was seen by many as a way to introduce some immediacy to the consequences of policy choices. But measurement of the funded position of a pension fund is a crude and often misleading indicator of system health and elides many important factors, such as whether policy is in place to address that funded position, whether that policy is being followed, whether economic conditions are improving, or even such basics as whether a system is open or closed.[34]

34 Sgouros, Tom. "Funding Public Pensions: Is full pension funding a misguided goal?" Berkeley, CA: Haas Institute for a Fair and Inclusive Society, University of California, Berkeley, 2017.

The current situation is thus that the consequences of pension policy changes are either distant and uncertain or immediate and potentially misleading. This is not an environment that encourages enlightened policy making. What is needed is an indicator as comprehensive as a bond rating, but less opaque, so that any government can predict with certainty the change that some policy choice will create in that indicator.

A further problem with the usual methods for assessing the health of a pension system is that the same numbers can imply different things, depending on context. A measurement of a system's condition does not carry with it a record of how it got there or what is expected in the future, and yet those are arguably more important considerations. A pension system in a financially weak position is in a weaker position if current policy will not improve it – or if a good policy is not being followed. On the flip side, a system in a weak position, but following strong policies to recover, is not necessarily in a bad situation. A useful indicator of pension health, therefore, must consider not just a system's condition, but also the policy in place by the management and the sponsoring government, as well as the actions actually taken.

The problem with a complex report about a complex system is that it invites any reader to come up with their own summarization of it.

This chapter describes a specification for a "scorecard" to serve as such a metric: a way to measure policy, action, and condition.

Though there are some enhancements described below, what is suggested here is not very different from a summary of a pension system valuation report. Such a thing might seem to be redundant since, after all, valuation reports already exist. Nonetheless, there are points to make in favor of establishing a succinct, widely used, standardized summary form.

To begin with, valuation reports already exist, and have existed for a long time. Notwithstanding efforts to standardize their presentation, if they were a solution to the problems outlined here, would those problems remain as salient as they appear today? A concise summary whose presentation is consistent across pension systems would make it easier to compare systems, but more important, will make it easier for observers to develop proficiency at evaluating the health of a plan. The problem with a complex report about a complex system is that it invites any reader to come up with their own summarization of it. Many observers of public pension systems – reporters and the public, but also some managers and trustees – do exactly that, and seize on one or two numbers they feel are the most important (often the funding ratio) and give short shrift to the rest.

Public pension systems differ along so many different axes that it is sometimes difficult to draw any conclusions through comparisons. Obviously, no summary can do justice to the complexity of all the different tax structures, employment contracts, benefits policies, and investment vicissitudes out there. But this is not an argument against a standardized summary, rather an argument in its favor. Differences become clear when the standards are consistent. Obviously, experience with a standard may lead people to revise it, but that is an argument for humility and flexibility, not an argument against the standard.

Ultimately, the goal of a scorecard is to elevate the conversation about public pension plans, to say they are complicated systems with a lot of moving pieces, albeit also with a lot of commonalities. Nuances are important, but the value of consistent summarization is high. Just as data has to be marshaled into a consistent form for a machine learning application to make its findings, bringing pension data into a convenient and easy-to-read format will make it easier for industry professionals – and the mayors, legislators, reporters, and citizens who are their observers – to develop expertise in making evaluations of complex data.

SCORECARD SPECIFICATION

The pension scorecard is to be a tool to understand how policy choices impact long-term pension health. We describe in this section the important categories and components of the scorecard.

Pension systems are widely variable, each distinct, with different drivers and different demographic characteristics. Each is a somewhat unique being. That said, many of the core financial inputs and the math are consistent across systems. In order to better articulate and understand the forces that impact the disparate population of pension plans and their relative health, there is a need for a way to normalize inputs and results. The pension scorecard attempts to create a normalized view of one or more pensions by tracking three broad categories of information.

- **Policy.** Are reasonable and appropriate policies in place to create long-term pension health?
- **Action.** Are the actions of the pension plan managers and the policy makers of the sponsoring government consistent with policy and appropriate to support long-term pension health?
- **Condition.** What is the current and historic condition of the pension plan, and are these indicators trending positively or negatively?

Within each of the broad categories, the scorecard focuses on three areas that interact to produce a pension plan's health for good or bad. The areas of examination are Benefits, Funding, and Investments.

For the Condition section, some attention should also be spared for a fourth area: the fiscal health of the sponsor government. A system's security depends on that as much or even more than on the accumulated store of assets. Obviously, Policy and Action matter to a government's fiscal health as well, but questions of tax policy, debt management, revenue growth, economic development, and all the other factors involved in managing a government seem well beyond the scope of a pension scorecard.

Each of these areas track in some way across the categories, meaning that there is a policy, action, and condition feedback loop within each area. An argument can be made for any of the categories being the starting point, but in reality, they are a fully intertwined loop that is in constant reorientation.

POLICY

One of the difficulties in measuring the health of a system is that some aspects are somewhat intangible. How do you quantify good management?

Difficult is not impossible, however, and there are a few ways that one can identify good management practice and put a rough value on it. The mere existence of policy documents is one metric; whether they address the significant points is another. What is important is that the parties to the plan are discussing issues and reducing to writing the strategies and measurements to be used to achieve long-term plan health.

Since the primary emphasis is not on the idiosyncratic details of a system's management, it seemed adequate to suggest a rough three-point evaluation of policy that might lend itself to a red/yellow/green color scale that might help in a quick assessment of many questions. These are the categories and questions suggested.

- **Benefits.** Whether for the sponsoring employer or the pension system, a clear articulation of benefits is important. The policy should clearly state the following:
  - The participation required from employees
  - The expected income replacement desired for various groups of retirees
  - The terms and/or prerequisites of any cost-of-living adjustment (COLA) for retirement benefits
  - The same description for other benefits the system might provide (e.g., disability and survivor benefits)

  If the benefits policy exists with good descriptions of the key points, the score would be "green." If the benefits policy exists but some key elements are absent, the score would be "yellow."If no benefits policy exists, the score would be "red."

  For many systems, these policies will be a matter covered by a collective bargaining agreement, or even by statute, rather than by a written policy statement. These agreements and statutes can be judged by the same criteria.

- **Funding.** The funding policy is the document that defines how the pension plan will be funded by the employer and employees, as necessary. The policy should clearly articulate these important points:
  - There should be a method for determining annual funding and, to the extent there is a calculation involved, a method for determining the actuarial inputs and methods for conducting the calculation or the actuarial process used in the calculation.
  - If there is no calculation but rather a fixed dollar amount or an annual negotiation, the amount of contribution or the methodology for handling negotiation should be outlined. A fixed-rate system should also describe a policy for handling funding shortfalls.
  - There should be a description of how to handle unfunded actuarial liability (UAL) sources and treatment of surplus when it exists.
  - There should be a description of how to achieve funding for a COLA or any other non-periodic pension payment benefits. Are such things part of the valuation, or are they ad hoc?

- If there are employee contributions included in the plan funding, that amount, percentage, or split with the employer should be described.

If the funding policy exists with relatively explicit descriptions of the funding terms and methods and is designed to achieve actuarial full funding through self-correcting contribution amounts, the score would be "green." If the funding policy exists and is either ambiguous in its terms or is not designed to be self-correcting through an actuarial contribution calculation process, the score would be "yellow." If no funding policy exists, the score would be "red."

- **Investments.** The investment policy is the document that defines the parameters of the investments for the pension plan. It should describe the terms of the pension plan in the context of its risk and return expectation, along with the roles of various participants in the plan, such as the investment adviser, the board, an investment committee, or whatever exists.[35] The key points of a good investment policy are the following:
  - Discussion of investment strategy and targets, short- and long-term
  - Discussion of investment risk and mitigation strategies
  - Discussion of the motivation behind the investment policy asset allocation
  - Identification of one or more benchmarks against which to measure investment performance given the plan's asset allocation

  If the investment policy statement exists with relatively explicit descriptions of the risk and return goals of the pension plan, the investment strategies and asset allocation targets to be utilized, and the role of various participants in management of the pension plan, the score would be "green." If the policy statement exists but is generally ambiguous or missing one or more key elements, the score would be "yellow." If no policy statement exists, the score would be "red."

ACTION

Experience has shown that policy and its documentation are important in identifying good management, but without action to match, even the best policy has no effect. The metrics outlined here can be thought of as measurements of actions rather than as outcomes. The focus is to understand the decisions made on behalf of a pension system in the current year and over the preceding few years.

- **Benefits.** Details of the benefits provided by the pension and from other ancillary programs should include the following.
  - Benefit replacement rate. This is determined by using example careers of 30 years of service and 10 years of service, followed by a normal retirement at the end of the current fiscal year. The replacement rate – the ratio between final salary and initial pension – is a way to measure the generosity of benefits, as they are actually in practice.

35 The CFA Institute published this general guide to writing an IPS for institutional investors https://www.cfainstitute.org/en/advocacy/policy-positions/elements-of-an-investment-policy-statement-for-institutional-investors

- Cost of living adjustment. Is the COLA linked in some way to decisions outside the system (e.g., linked to inflation or investment returns)? The actual COLA value is part of the "Condition."
- Workforce participation in Social Security. This could be answered with "Yes," "No," or "Some."

- **Funding.** This provides a picture of annual contributions on several bases.
  - UAL stabilization payment (percentage of payroll). See UAL Stabilization Payment chapter on page 20
  - Actuarial defined contribution (ADC) rate (percentage of payroll). Comparing the ADC to the same yardstick as the USP and the actual contribution makes it easier to compare the three numbers.
  - Actual contribution rate (percentage of payroll). Whether the contribution was actuarially calculated or a fixed dollar amount, what percentage of payroll was covered by the contribution?
  - Normal cost (percentage of payroll). The normal cost accrued over the past year can provide a view of the trend in pension plan costs.
  - Experience study. It is important to measure how a pension system's actual results have diverged from the assumptions on a regular basis. This type of examination provides an effective method to determine adjustments to the pension valuation, but importantly it also provides a catalyst for discussing the most appropriate changes to assumptions going forward.

    As with the policy documents, it is a sign of good management to have an experience study done at regular intervals. If an experience study exists that is less than three years old, a "green" rating should be applied. If an experience study exists but is older than three years, a "yellow" rating should be applied. If no experience study has been done, a "red" rating should be applied.
  - Actuarial assumptions. An enumeration of the assumptions for the discount rate, the rate of inflation, and the rate of wage inflation can be important indicators of the risk appetite and the frequency with which assumptions are re-examined.

- **Investments**. These are actions that management chooses.
  - Investment management fees should present the dollars paid directly and indirectly during the fiscal year as a percentage of average assets under management (AUM). The scorecard should include both direct fees, paid via invoice, including fees drafted from a portfolio, as well as indirect fees, such as those paid within a mutual fund, electronic transfer fund (ETF), or other pooled vehicle in which fees are accrued daily within the fund.
  - Allocation of investments should show the broad categories of equities, fixed-income, real estate, hedging, private equity, cash, and other instruments.
  - The Sharpe Ratio measure of risk can be calculated from the data underlying the investment report and provides a measure of risk-adjusted return.

CONDITION

The last component to the scorecard is a more traditional collection of measurements to indicate the current condition of the pension system, providing an annual snapshot of the overall condition of a pension system along various axes.

- **Benefits**. This section records the conditions of the benefit structure: how many employees are in which benefit class, the median time to retirement, and other facts determined by policy choices.
  - Show the average age and number of employees in however many different benefit classes there are. Where there are more than a handful of classes, present a mean with an explanatory note.
  - Include the average remaining years to retirement as a measurement of the maturity of the employee population, if available.
  - Indicate what the COLA is, if one exists, for retirees in the given year.

- **Funding.** This will record facts about the overall funding of the system, again as a result of the policies and actions on which they depend.
  - The overall actuarial liability should be presented, as well as the actuarial and market value of the assets. If there is debt from a pension obligation bond, that should be presented as well, though separately.
  - The unfunded liability should be presented as a percentage of payroll.
  - A presentation of the scaled liability, as described in the Liabilities in Context chapter on page 9.
  - The net cash flow (percentage of assets), including all inflows and outflows, such as pension obligation and bond debt service, should be included. This is a useful measure of resilience when presented relative to total AUM.
  - Was an extraordinary contribution made to the plan? If so, in what amount and from what source (e.g., pension obligation bond, windfall, budget savings)?
  - If the system uses a layered amortization and the valuation report includes an analysis of the layers, this indicator is "green." Otherwise, it is "red."

- **Investments.** The basic investment measurements are fairly straightforward.
  - The scorecard should show the one-year, five-year, and 30-year net investment returns and compare them to the assumed rate of return. If 30 years' data are not available, some other long period can be used.
  - Reported benchmark performance, for the same reported periods, should be included, if available
  - Asset values, in proportion to benefit payments, should be included, as well as the amount by which this would be reduced by risk-weighting the asset values. (This is discussed the Risk Weighting and Other Shortcuts chapter on page 37.)

- **Sponsor Fiscal Health.** To understand the overall health of any pension, it is important to also understand the economic conditions affecting the employer sponsor. There are too many variables here to record more than a small fraction of them, but the following are a good beginning to any fiscal analysis and should be included as a gross measure of the sponsor's fiscal health.
  - The budgeted general revenue amount.
  - The per capita personal income. (See Liabilities in Context on page 9.)
  - The poverty rate, per the U.S. Department of Health and Human Services.
  - The general obligation bond rating.

EXAMPLE

Rhode Island ERS (State Employees + Teachers)

POLICY					
Benefits		**Funding**		**Investments**	
Employee participation	●	Annual employer share	●	Investment strategy	●
Income replacement	●	UAL sources	●	Risk discussion	●
COLA terms	●	COLA funding	●	Allocation motivation	●
Other Benefits	●	Employee contribution	●	Benchmark defined	●

ACTION					
Benefits		**Funding**		**Investments**	
Benefit replacement	10yr 16%	USP % payroll	28.3%	Global equities	42%
	30yr 53%	ADC % payroll	31.3%	Fixed-income	24%
COLA	suspended until UAL>80%	Actual Contribution	31.3%	Real estate	7%
		Normal Cost	8.1%	Hedging	9%
SS participation	some	Experience study	●	Private equity	14%
		Assumed return	7%	Cash	4%
		Assumed inflation	2.5%	Investment mgmt fees	●
		Wage inflation	3%	Sharpe ratio	●

CONDITION					
Benefits		**Funding**		**Investments**	
Active state employees	N=10,803	Total liability	$18.89 billion	Assets/Benefits	8.05
	Age 49.2	Actuarial assets	$6.89 billion	Risk-weighted assets	6.92
Active teachers	N=13,372	Market Assets	$7.73 billion	Market returns *1-year*	
	Age 46.8	UAL as % payroll	26%	Net	2.2%
Retired state employees	N=9,270	POB debt	$0	Bench	11.2%
	Age 74.3	Scaled liability	0.4%	Market returns *5-year*	
Retired teachers	N=10,441	Net cash flow	21%	Net	0.1%
	Age 74.2	Extra contribution?	No	Bench	9.8%
Actual FY21 COLA	0.0%	Layered Amort?	●	Market returns *10-year*	
				Net	8.5%
Sponsor Fiscal Health				Bench	9.8%
Budgeted general revenue	$4.43 billion			Market returns *since 1995*	
Per capita income	$37,504			Net	7.7%
Poverty rate	10.6%				
GO Bonds M/SP/F	Aa2/AA/AA				

A mock-up of the scorecard for the Employee Retirement System of Rhode Island.

ELECTRONIC STANDARDIZATION

Summary: *As with the standard reporting of the scorecard, standardized electronic reporting will facilitate the collection and interpretation of data about pension plans. The XBRL business information electronic standard is currently being adapted for US state and local government financial reporting purposes, including for reporting pension results. Electronic standardization of results will make generating the scorecard described above much easier, perhaps even dynamically, through a phone app or specialized web page.*

Though plans are all different from one another, it is wrong to say that one plan cannot learn from another's experience. The scorecard described here is a way to facilitate this learning, through standardization of a capsule report format.

There are other ways to accomplish a similar goal. To pick an example close at hand, to a large extent, the work of this report could not have been accomplished without the Public Plan Database (PPD), a tremendously valuable repository of annual pension data for more than 200 public pension plans. By providing the data all in one place, in a more or less standardized format, the PPD[36] makes analysis of plans simpler, but also facilitates the testing of new metrics through its wide range of available data.

XBRL is a mature standard for the electronic reporting of financial and business information. For over a decade, the US Securities and Exchange Commission has required its use by companies that report corporate financial results. It is also used for bank regulatory "call reports" to the FDIC, and the Federal Energy Regulatory Commission uses it for oil, gas, and electricity utility filings. The corporate taxonomy – the definitions and relations among the different quantities defined – is managed by the Financial Accounting Standards Board (FASB).

XBRL is a mature standard for the electronic reporting of financial and business information. Why not pensions, too?

The technology is straightforward and essentially involves labeling each number in a financial report with a tag that can be matched to a detailed definition of that quantity, and that links that number to others. For example, a total of some column of numbers can be labeled to belong to a fund or a category.

A newer version of XBRL, known as Inline XBRL, incorporates tags into an HTML document. The tags are hidden in a web page so they are invisible to a human reader, but can be viewed with a Java Script-based reader and can be automatically parsed by a simple application.

Not surprisingly, there are an increasing number of governments adopting the standard. All the municipalities in Spain and Finland, for example, currently use it to report financial results. Over on this side of the Atlantic, Florida is developing a taxonomy for reporting by its local governments, and a group at the University of

36 Op. cit. See page 7.

Michigan, in cooperation with the Michigan Treasury, is engaged in a similar development project aimed at all the municipalities there, beginning with the city of Flint.[37]

The Michigan project is including in its work a taxonomy for the reporting of pension results. If XBRL were universally used for the reporting of pension plan data, the comparison of one plan to another, the comparison of one plan to itself at a previous time, and the collection of aggregate data would be profoundly easier to accomplish. This would ease the work of policy analysts and potentially offer new avenues of inquiry and research. But it would also have the effect of improving the quality of information available to policy makers, enriching their understanding of the context of their decisions and leading to better decisions.

DISCUSSION

There was a fair amount of discussion about how exactly one could quantify good or bad management, but there were some common points. Many members agreed with the assertion that the mere existence of policy documents was a positive indicator, and that in judging policies, one can learn something of value just by focusing on big-picture concepts rather than on idiosyncratic details.

Quite a bit of discussion focused on the question of tax capacity and the difficulty of assessing it across different jurisdictions with different histories. One perspective is that the real tax capacity is a function of the income with which the population will pay it, regardless of the institutional, constitutional, and statutory limitations, none of which are irrevocably permanent. Others insisted that they must be taken into account.

Ultimately, perhaps the resolution depends on the answer to the question of who this scorecard is for. An analyst looking to compare the performance of one system with another will necessarily be less interested in the institutional and statutory limitations, while someone interested in the evolution of a particular system will be attuned to exactly those details.

37 See a working example of a balance sheet annotated with XBRL for the city of Flint at https://michsav.github.io/acfr/samples/54/ixbrl-viewerIDXUS.html

Conclusion

Though some of the ideas in this report may seem unorthodox, any field can benefit from periodic re-examination of its fundamental tenets. Accounting is no different. The world does not stand still, and accounting standards have long evolved with it, sometimes in response to varying conditions, and other times as a result of changes in behavior.

Just since 2000, there have been substantial changes in the accounting standards applied to banks, insurers, foundations, and corporations. The Basel III reforms brought comprehensive changes to how we value banks and how they value their assets. Private charitable foundations saw a shift from the "prudent man" to the "prudent investor" standard of value maintenance. And the arguments continue to rage among accountants about how corporations should evaluate and report their risks, how they should account for off-balance sheet transactions, and what value to place on stock options granted to their employees.

Pension crises are slower, and slow crises can be very difficult to recognize.

One difference between pension funds and these other organizations is time scale. Banks can become insolvent in a relative blink of an eye, and even the biggest corporations can quickly collapse under the wrong conditions. Pension funds operate on a different, generational, time scale, akin to the largest private foundations and endowments. In many cases, changes in accounting for banks, insurers, foundations, and others have been forced by the sudden onset of dire circumstances. The improvement of the Basel III framework over Basel II was motivated by the banking crisis of 2008, and the same crisis brought the final adoption of the prudent investor standard for the few holdout states that had not already adopted it. But pension crises are slower, and slow crises can be very difficult to recognize.

Documentary filmmaker John Marshall began his career with a critically acclaimed 1957 movie about the nomadic Bushmen of the Kalahari Desert. The movie followed several families as they hunted for small game, extracted water from desert plants, and otherwise eked a living out of a harsh African landscape. He and his audiences were fascinated, but it was not until years later that he understood that he had been filming a people suffering a decades-long catastrophe. Before he arrived with his camera and his interest in desert lore, ranchers had fenced the Bushmen off from the springs and watering holes they had been drinking from and hunting around for centuries, even shooting some who dared drink with the cattle. The bush tribes were starving, but the slow-moving nature of the crisis made it all but invisible to his eye – on that first trip. His return to the Kalahari Desert, 30 years later, made clear to him and his crew what had been unseen at the time.

Marshall learned that a slow-moving crisis can be difficult to detect from among the noise and distractions of other events and one's preconceptions, and the claim that the nation is currently undergoing exactly that has

been a staple of public pension criticism for over a decade.[38] Though it is possible that what we see around the country is a slow-moving pension crisis, that is no reason to quash the spirit of inquiry. On the contrary, it is all the more reason to look to the fundamentals of the enterprise, to ask about the interests of the members, the risks of the markets, and the incentives of the decision makers. Do the existing metrics provide the best picture of a pension system's current and future condition? Do they diagnose the right problems? Do they motivate the best decisions?

Do the existing metrics provide the best picture of a pension system's current and future condition? Do they diagnose the right problems? Do they motivate the best decisions?

The development of prospect theory, the psychology of making decisions, has meant the development of a conceptual framework and experimental tools with which some of these questions can be addressed empirically. Survey work and rigorous protocol analysis can reveal incentives and predilections among the population of legislators, trustees, and managers that may not be apparent to technicians. These questions will be important ones to answer going forward, and this report suggests alternatives worth consideration and empirical evaluation.

The metrics suggested in this report are suggestions meant to offer new perspectives and new insights. They are not a finished product, collectively or individually. There are unanswered research questions about how to calculate the USP, about the effects of stress tests on decision making, and about the weights to use in risk-weighting assets, among several other topics. But we find that they do raise interesting questions and believe that widespread application of these ideas will provide new intuitions and lessons about how best to manage these complex enterprises.

Through it all, the goal is the same: to find the optimal combination of expense and security, to maximize the benefit for plan members at a sustainable cost. When looking at the innovations of the early 20th century and the way life was changed for millions through the development of new technologies, people often focus on the material: the cars, the radios, and the airplanes that changed the world. But the less material innovations in finance – consumer-level insurance, banking, and pensions – were just as momentous, possibly more. These innovations democratized finance and helped create the widespread prosperity that has been our nation's hallmark for the better part of the past century.

The defined-benefit pension plan is among those innovations and, as an important contributor to that prosperity, well worth preserving. It is precisely that value that makes scrutiny so important. Currently at a low ebb in private business and under pressure in the public sector, pensions deserve a close look and a critical reconsideration of the methods we use to run and evaluate them, to preserve this valuable institution for the prosperity and security of this and future generations. If the ideas presented here can contribute to that preservation, they will be worth our time.

38 See, for example, https://www.nationalaffairs.com/publications/detail/how-to-avert-a-public-pension-crisis

Appendices

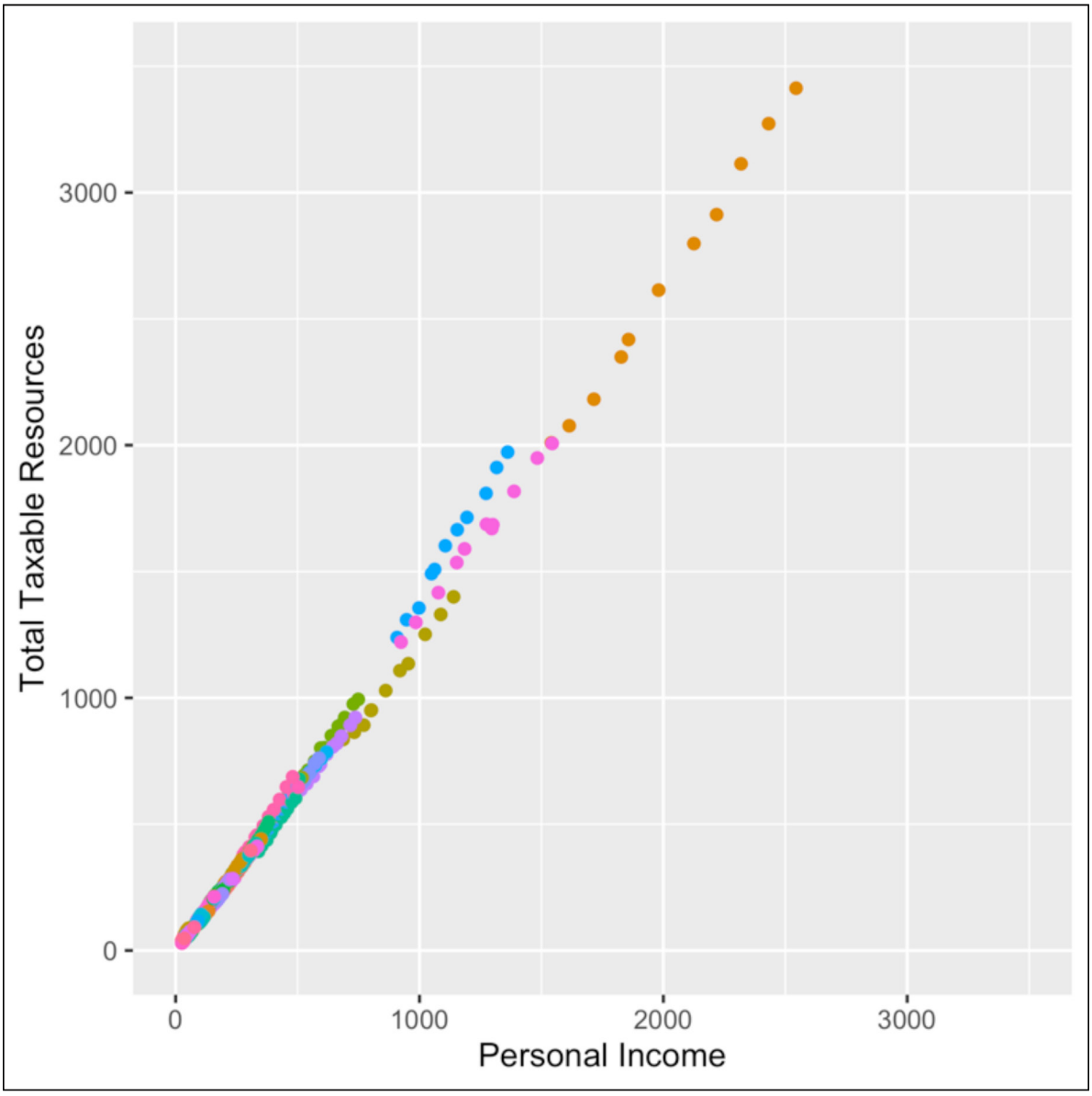
APPENDIX A: US TREASURY TOTAL TAXABLE RESOURCES (TTR)

The US Department of Treasury calculates "total taxable resources" (TTR) for the 50 states and the District of Columbia as a way to introduce a more rigorous approach to the question of tax capacity. It is used for allocating federal aid in a few different block grant programs. The metric takes into account several different variables – gross state product (GSP), contributions to Social Security, business income, and several others[39] – to make an estimate of income from which a "reasonable" level of tax might be collected. It does not account for differences in tax structure, and is confined to measurements of income that might be taxed.

The figure below shows TTR plotted against personal income for each state between 2009 and 2019. The data are very close to falling on a straight line, which means that the TTR, a sophisticated metric of tax capacity meant to answer the many objections to using the simple measure of personal income or gross state product, is not very different at all from the simple measure it was meant to replace. For the purposes of tracking a rough measure of tax capacity, personal income seems as good a measure as any other.

The TTR varies across the states, but the bulk of the states cluster around a value about 25 percent to 30 percent higher than personal income. For comparing one state to another, one might want to correct for this variation, but in a practical sense, for revealing trends in the comparison between pension liability and tax capacity, the two measurements are equally useful. Given the Bureau of Economic Analysis's timely provision of personal income data, its availability at the county level, and the ease with which it can be imputed to sub-county regions, the convenience of personal income is quite competitive with the theoretical benefit of the TTR.

39 See https://home.treasury.gov/policy-issues/economic-policy/total-taxable-resources



Total taxable resources versus personal income. This is 20 years of data for all 50 states. (Each state has a different color, reading from the right; California, Texas, and New York are the largest economies, so extend out to the right.) The points are very close to a straight line, implying that there is not a lot of difference between using TTR and personal income as a proxy for tax capacity in a comparison with pension liability. TTR does vary between the states, but what is important is the trend, not the exact value, and that is similar for either measure.

APPENDIX B: CALCULATING THE UAL STABILIZATION PAYMENT

We name this quantity "Unfunded Actuarial Liability (UAL) Stabilization Payment" because it differs significantly from Moody's "Tread Water" calculation, as outlined here.

Assuming no confounding policy change, such as an adjustment in the assumed rate of return or a change in benefit or employment policy, the rate of increase in a plan's unfunded liability is the solution to two paired equations describing the growth of the total liability and the appreciation of plan assets. One can write the following equation for the liabilities and assets:

$$L_{t+1} = (1 + \delta)L_t - B_{t+1} + NC_{t+1}$$
$$A_{t+1} = (1 + r)A_t - B_{t+1} + C_{t+1}$$

Where L_t and A_t are the values of the liabilities and assets at year t, B is the benefits paid in year $t+1$, NC the normal costs incurred that year, and C the contributions. The discount rate used to value the liability is δ, and r is the assumed rate of return on assets. These last two are typically assumed to be the same, as they are for the Tread Water and minimum funding progress (MFP) calculations.

There are hidden dependencies in the equation, since the liability change also depends on factors like the rate of employment growth, demographic changes in the member population, and benefit policy changes. The liability is not typically calculated with this equation, but is calculated each year by updating previous calculations of normal cost and aggregate changes. Because of the other dependencies, it makes sense to ask whether the assumed rate of return is the best estimator of the accrual rate for the liability.[40]

If the net pension liability is to remain unchanged during the course of a year, then $L_{t+1} - A_{t+1} = L_t - A_t$, so one can combine the two equations and get the following:

$$L_t - A_t = (1 + \delta)L_t + NC_{t+1} - (1 + r)A_t - C_{t+1}$$
$$C_{t+1} = \delta L_t - rA_t + NC_{t+1}$$

C_{t+1} constitutes the UAL Stabilization Payment (USP). A plan sponsor making this payment will see their plan end the year in the same shape as it began, assuming it meets the investment target.

One can take the idea a bit further. Lenney et al. make the case that the proper measure of the status quo is the maintenance of the unfunded liability as a proportion of the size of the economy governed by the sponsor.[41] If the unfunded portion of the liability grows no faster than the economy that supports it, then the growth is sustainable. If $\delta = r$ and g is the rate of economic growth, then the USP is the following:

40 The normal cost calculation does, of course, also depend on the assumed rate of return. But as we will see, that does not by itself imply that the assumed rate is a good predictor of the result.

41 Op.cit. See Liabilities in Context chapter on page 9.

$$C_{t+1} = (r - g)(L_t - A_t) + NC_{t+1}$$

If the size of the liability relative to the size of the economy is a more appropriate way to measure the load of a pension system, then this payment might be seen as a more appropriate way to measure the payment required to maintain the current condition.

There is, however, another avenue to explore. Returning to the original equation, at the top of the previous page, governing the rate of increase of a pension liability:

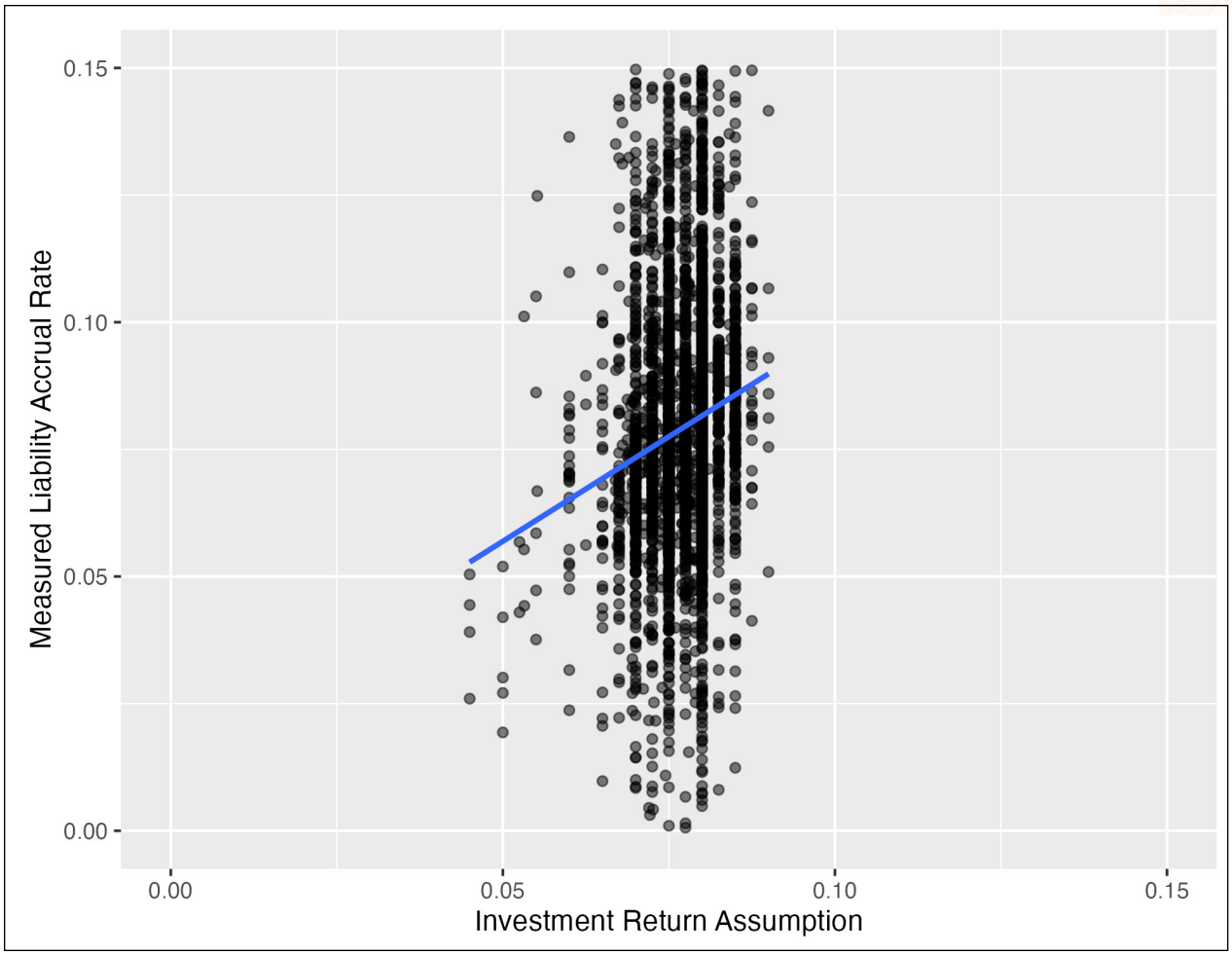
$$L_{t+1} = (1 + \delta)L_t - B_{t+1} + NC_{t+1}$$

Using successive years' liability estimates, it is possible to solve for δ, the accrual rate at which a liability increases, given the concomitant normal cost and benefit payments.

$$\delta = \frac{L_{t+1} - L_t + B_{t+1} - NC_{t+1}}{L_t}$$

Again, for the purpose of this roll-forward equation, δ is usually assumed to be equal to the assumed rate of return. But there is data with which to test this assumption.

Using data from the Public Plan Database (PPD), we calculated δ for 218 plans over the years 2000–2020. After discarding the values above 15 percent and below 0 percent, on the assumption that these probably represent some kind of significant policy change, they are presented in the figure on the next page, graphed against the assumed rate of investment returns for each plan and year.

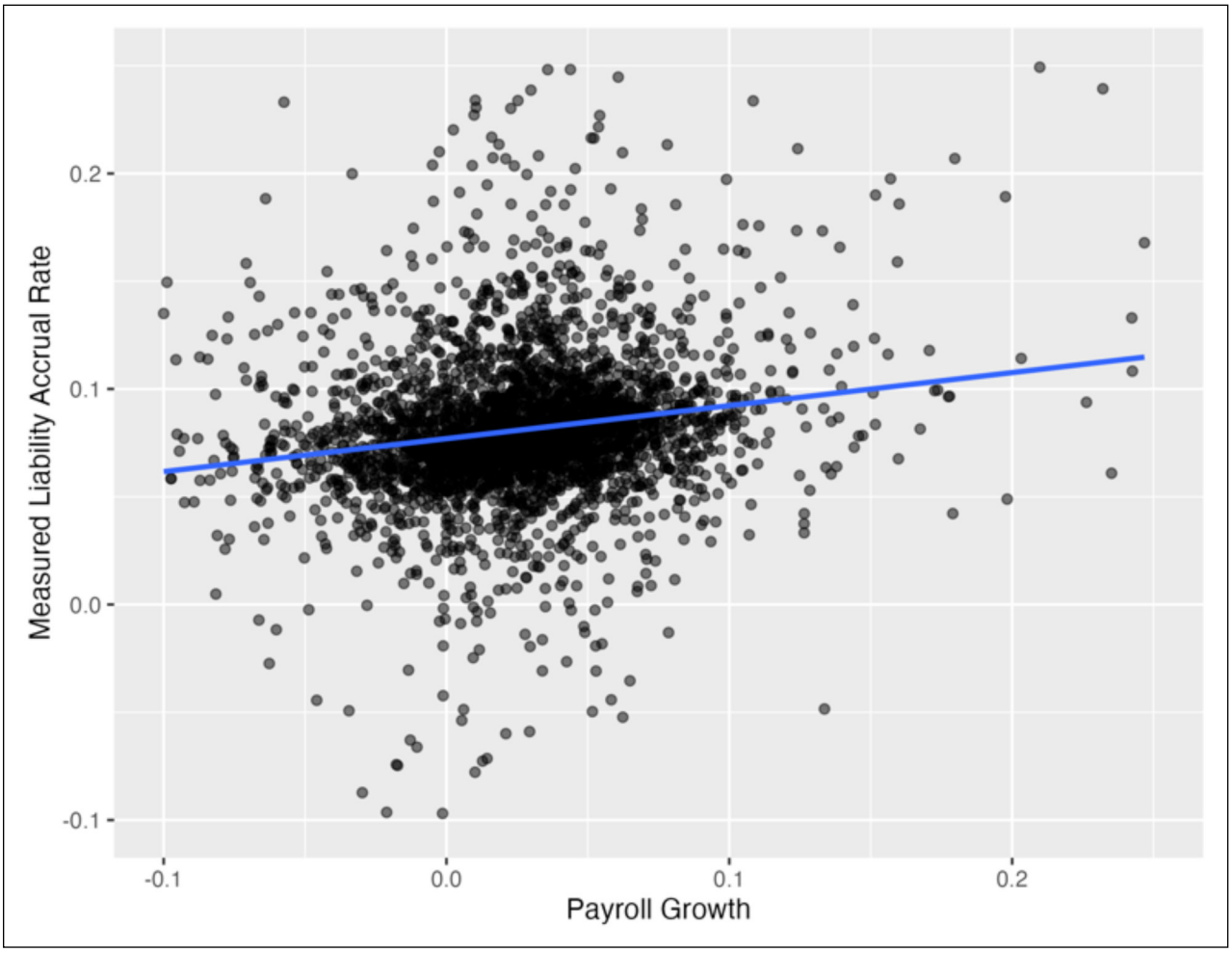


Assumed rate of return measured against δ, the real accrual of liability, for 218 pension plans in the Public Plan Database. The fit of the blue line is good, implying one can have confidence in the trend that a higher assumed rate corresponds to a higher accrual rate. However, the span of possible values is very large, so the blue line is a poor predictor of the actual value of the liability accrual rate.

The blue line is a linear fit to the data, and its slope is 0.82, fairly close to one, and meets the 99 percent standard of statistical significance. However, the variance is quite large, spanning the whole range of outputs for many values of the input. That is, if you know that the assumed rate of return for some given year is 7.5 percent, the range of possible accrual rates on the y-axis spans the entire range, from 0 percent to 15 percent. This is not uniformly random, of course, but the standard deviation there is about 2.5 percent, implying that after accounting for the discarded values, there is only a 60 percent chance that the actual accrual rate for that year is between 5.2 percent and 10.2 percent. In other words, the fitted line certainly shows a trend, but the values on the line are not good predictors of how fast the liability grows.

For a system with a relatively small unfunded liability, this difference may not be that significant in the USP calculation. With a funding ratio of 80 percent, for example, and assuming other typical values for the size of payroll and benefits, this might work out to a predicted error of plus or minus 3 to 4 percent of payroll for the USP. For a system at 50 percent funding, however, it would be more in the range of plus or minus 7 to 8 percent. Lower funding ratios will cause greater inaccuracy.

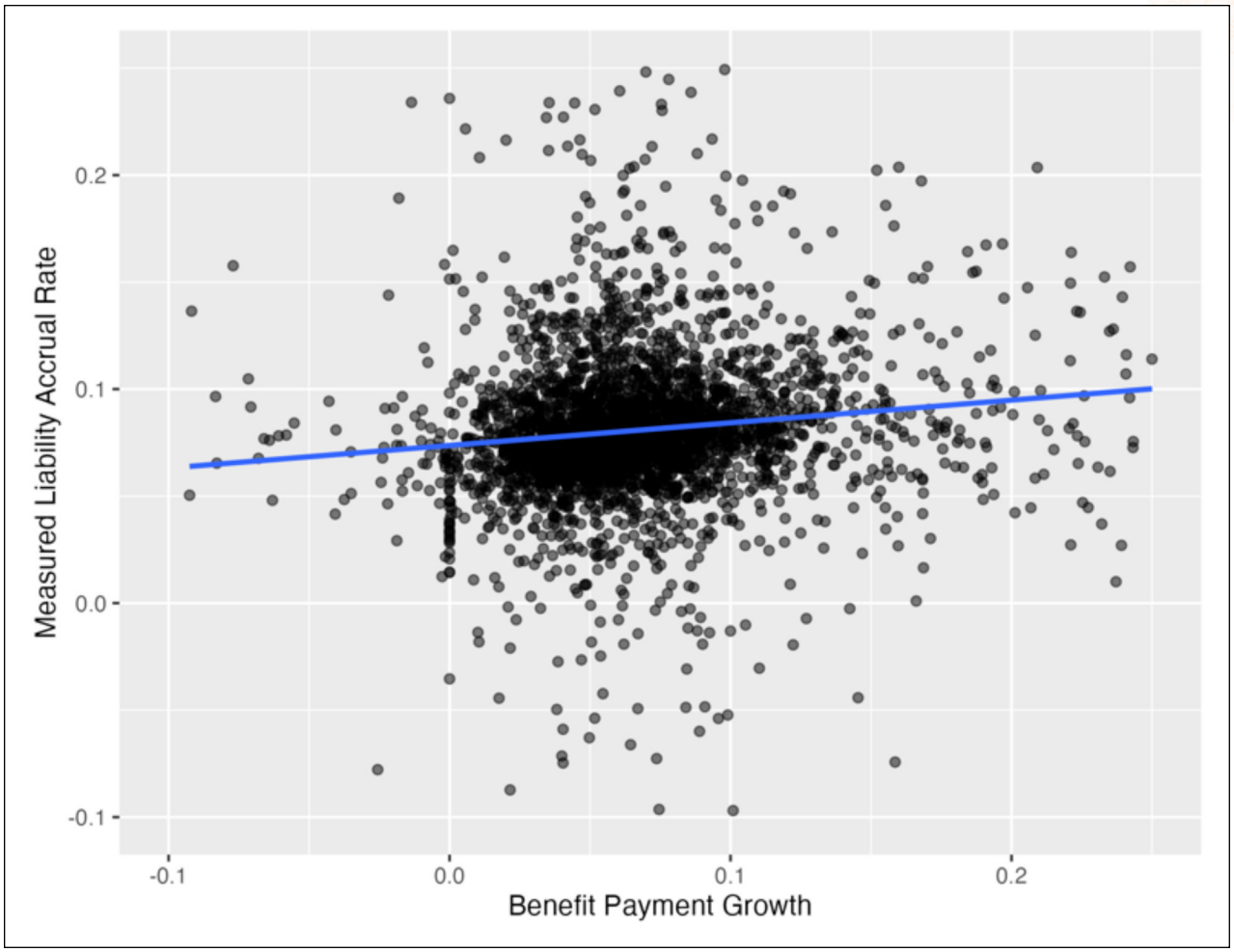
Analysis of the PPD data shows that the annual rates of benefit growth and payroll growth provide a slightly more reliable estimator of liability accrual, though still far from perfect. Here, for example, is δ against the annual rate of payroll growth.



Payroll growth versus measured liability accrual rate. The fit of the line is still good, with better than a 99 percent confidence rate in the trend, but it also is not a very good predictor of the accrual rate, even if it is slightly better than in the figure on the previous page..

The correlation is slightly better than it was against the assumed rate, but the variance is still substantial, even if it is smaller than before. For payroll growth of 5 percent, there is a 60 percent chance that liability accrual is between 5.7 percent and 10.1 percent, a slightly smaller range.

Annual growth of benefit payments provides a still better fit to the liability accrual, though again the variance is substantial. The figure on the next page shows measured δ against the annual growth in benefit payments.



Benefit payment growth versus measured liability accrual. Benefit payment growth is a slightly better predictor of liability accrual than payroll growth, but only slightly.

The variance here is again slightly smaller than for payroll growth, and for benefit growth of 5 percent, there is a 60 percent chance of the range of liability accrual being between 6 percent and 9.3 percent. The correlation between liability accrual and annual benefit growth is also higher than it is with the assumed rate of return or payroll growth.

The evidence is that the liability accrual rate, δ, from first equation on page 58 is determined by a fairly complex combination of different factors relating to the management of a pension plan: payroll and benefit growth, but also changes in benefit policy, the employment contract, and workforce demographics. Using the assumed rate of return as a proxy for this value may provide an adequate approximation for plans with high funding ratios, but cannot be assumed to be accurate at lower levels.

In addition, there is significant evidence for systematic variation in the fit of these models from one plan to another. The statistical term for this is heteroskedasticity, where the important characteristics of the randomness of the whole – the mean, variance, and distribution character – is not reproduced in the subsets. In other words,

the randomness of the Oklahoma PERS liability accrual is different from the randomness of the same variable for the Rhode Island ERS or any other system. If one separates the data by plan, as much as 15 percent to 25 percent of the variance in the resulting models can be attributed to the plan itself. This implies that plan and sponsor policy choices are a major determinant of the accrual rate. In other words, the most accurate value to use for liability accrual might best be defined in the context of an individual plan's performance and history, which also implies that policy choices by that plan's management can have a significant effect on the accrual rate of the liability.

One can, however, use benefit growth and payroll growth as independent variables in a regression against the measured liability accrual rate. The examples shown in the UAL Stabilization Payment chapter on page 20 are plans in which this estimated rate, smoothed over five years, was close to the assumed rate of return.

APPENDIX C: DERIVING RISK WEIGHTS

Assuming that investment returns are normally distributed, we assigned various well-known indexes to different asset classes, and for each one calculated the mean one-year return, starting at any point, and its standard deviation. The data ran up to December 2021. We repeated the procedure for an annualized long-term mean return and its standard deviation. The available time series for each index varied considerably. Where the data was adequate, we used 30 years as the long-term period, and where we have less data, we used 25 or 20 years.

Asset class	Index	Start date	Term	Short-term σ	Long-term σ
US Equities	Russell 3000	1/1979	25	17.0%	2.1%
Large Cap	S&P 500	1/1969	30	16.8%	1.54%
Mid Cap	S&P 400	1/1985	20	17.7%	2.35%
Small Cap	Russell 2000	1/1979	25	21.1%	1.87%
International	MSCI ACWI ex. US IMI (net)	6/1994	20	19.6%	1.07%
International Developing	MSCI EAFE (net)	1/1970	30	21.4%	3.25%
International Emerging	MSCI Emerging Markets (net)	1/1988	20	27.0%	4.1%
International Small Cap	MSCI ACWI ex. US Small Cap (net)	6/1994	20	23.3%	1.71%
Fixed Core	Bloomberg US Aggregate	1/1976	25	6.72%	1.91%
Fixed High Quality	Bloomberg US Corp High Yield	7/1983	20	11.5%	1.2%
Fixed Investment Grade	Bloomberg US Corp Invest Grade	1/1973	30	8.62%	1.17%
Public Real Estate	FTSE NAREIT Equity REITs	1/1972	30	18.9%	3.14%
Private Real Estate*	NCREIF Property	3/1978	25	7.37%	0.87%
Hedge Funds	HFRI Fund of Funds Composite	1/1990	20	8.51%	1.71%
Private Equity*	CA US Private Equity Index	6/1986	20	14.4%	1.47%

** Data available quarterly instead of monthly.*

Cash flow is defined as contributions received net of benefits, as a fraction of assets. For most plans, this ranges from below zero to somewhat more than 10 percent. We used zero and 10 percent as bounds, and for a cash flow C, defined an interpolation factor f as:

$$f = \max(0.0, \min(0.1, C)) \times 10.$$

This provides a factor that ranges from 0 to 1, according to the cash flow. You can think of this number as the percentage of assets that will be valued with the short-term discount, with the remainder being valued using the long-term discount.

For some given asset class, given a short-term discount of σ_{short} and a long-term discount of σ_{long}, the risk weight is then:

$$W = (1 - f) \times \sigma_{short} + f \times \sigma_{long}.$$

For an asset of value A, the risk-weighted asset value becomes:

$$A_{rw} = A \times (1 - W).$$

For private real estate, hedge funds, and private equity, we use 99 percent as the short-term risk weight (σ_{short}) instead of the short-term standard deviation, because these investments often have significant short-term liquidity constraints. Note that only the plans with the greatest negative cash flow will feel the full short-term discount. Most plans will find a value partway between the short- and long-term discount values, so a plan with a cash flow of -5 percent of assets would discount its private equity investments by about 50 percent, for example. The goal, again, is a quick and easy simulation of a worst-case scenario, so while this seems dramatic, it is a plausible outcome of a catastrophic market downturn.

National Conference on
Public Employee Retirement Systems
1201 New York Avenue, NW, Suite 850
Washington, DC 20005
202-601-2445 • info@ncpers.org • www.ncpers.org