

RESEARCH ARTICLE OPEN ACCESS

The Practice and Origins of Performance Measurement: Assessing the Sunnyvale Model, 1984–2023

Joyce Kasing Tagal | Derek Ouyang | Daniel E. Ho

Stanford Law School, Stanford, California, USA

Correspondence: Daniel E. Ho (dho@law.stanford.edu)

Received: 8 July 2025 | **Revised:** 20 July 2026 | **Accepted:** 12 August 2026

Keywords: gaming | local government | performance measurement

ABSTRACT

Few regulatory reforms are as celebrated or emulated as the performance measurement system of Sunnyvale, California, which inspired the Government Performance and Results Act of 1993 and countless other national and local efforts. This study presents the first longitudinal, cross-programmatic analysis of a performance measurement system, introducing a digital archival approach to analyze 152 performance indicators across 35,000 pages of documents from 1984 to 2023, supplemented with interviews of key stakeholders. Even if Sunnyvale's system initially functioned as an exemplar of performance-based decision-making, it deteriorated with weakened commitment and, more importantly, sustained significant implementation costs while its benefits reduced over time. The findings illustrate three core trade-offs for performance measurement systems: measuring what matters (the “quantity-quality” tradeoff); ensuring continuity while allowing for adaptation (the “continuity-flexibility” tradeoff); and ensuring accountability without having implementation detract from productivity (the “accountability-productivity” tradeoff).

1 | Introduction

Few regulatory reform proposals have been so widely acclaimed or adopted as performance measurement. Its central insight is simple and powerful: “what gets measured gets managed”.¹ The 1990s bestselling book *Reinventing Government* (Osborne and Gaebler 1992) introduced the idea of performance measurement—a concept originating in the private sector—to streamline government. At the core of *Reinventing Government* was the performance measurement system employed by the City of Sunnyvale, California, which tracked “thousands of measures” on the “quantity, quality, and cost of every service” delivered (Osborne and Gaebler 1992).

By the early 1990s, Sunnyvale was hailed by the media as a “model of efficiency” and an “ideal” to follow (Richter 1993;

Thurm 1993). The Office of Management and Budget (OMB) called Sunnyvale the “single best example of a comprehensive approach to performance measurement that we have found in the United States.” (Report of the Committee on Government Affairs 1993). Osborne and Gaebler argued that the Sunnyvale system generated great productivity and efficiency. The “city's average cost per unit of service went down 20 percent” between 1985 and 1990, even though it “used 35 to 45 percent fewer employees to deliver most services” compared to other cities of similar size and type (Osborne and Gaebler 1992, 145).

Upon assuming office, President Clinton and Vice President Gore, inspired by Oswald and Gaebler, announced the National Partnership for Reinventing Government, a series of reforms to reignite an economically stagnated America. The centrality of Sunnyvale in the namesake bestseller was not lost on them. In

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *Public Administration Review* published by Wiley Periodicals LLC on behalf of American Society for Public Administration.

Evidence for Practice

- This study provides the first longitudinal, cross-programmatic assessment of a performance measurement system, introducing a digital archival approach to analyze 40 years of municipal budget documents from Sunnyvale, California: the prototype for the Government Performance and Results Act of 1993.
- Sunnyvale's system, initially an exemplar of performance-based decision-making, deteriorated with weakened commitment and, more importantly, sustained significant implementation costs to bureaucrats while its benefits reduced over time.
- Three trade-offs complicate the use of performance measurement systems: quantity-quality, incentivizing progress towards the right outcomes versus merely the most observable ones; continuity-flexibility, ensuring stability versus allowing evolution of measures; and accountability-productivity, time spent on reporting versus implementation.

1993, Clinton and Gore sang the praises of a gathered crowd of city employees on the steps of Sunnyvale's community center:

The satisfaction of the citizenry with the quality of government services has risen steadily as the cost of providing those services has declined steadily. You were able to do that by measuring performance and concentrating on results instead of just looking at the inputs. You got rid of the top-down, bureaucratic, centralized, hierarchical approach, and instead you worked with your employees, empowered them with flexibility, encouraged them to innovate, and then held them strictly accountable for the results that they produce every year [...] It is high time that the federal government did exactly the same thing.

(Clinton 1993)

The crux of their efforts was the Government Performance and Results Act of 1993 (GPRA), a bipartisan act of Congress modeled after Sunnyvale's system, which required federal agencies to develop five-year strategic plans and report performance towards annual targets to Congress.

The legacy of Sunnyvale still reverberates. In 2003, OMB wrote: "Sunnyvale, California has become internationally recognized for performance budgeting" (Office of Management and Budget 2003, 4).² Later reforms built upon GPRA. The W. Bush administration introduced the Program Assessment Rating Tool (PART), a rating system to measure federal program performance, complementing GPRA (Office of Management and Budget 2004). In 2010, President Obama signed the GPRA Modernization Act requiring agencies to publish their performance plans online (Rep. Cuellar 2010). During the Biden administration, the public could peruse participating federal agencies' GPRA-mandated performance measures on [Perfo rmance.gov](https://perfo rmance.gov).³ OMB guidance to agencies and federal grant applicants regularly includes recommendations for performance

tracking and reporting.⁴ Efforts in President Donald Trump's second administration by the "Department of Government Efficiency" were described by Elon Musk as "similar to Clinton/Gore Dem policies of the 1990s", sparking considerable debate about the degree to which DOGE has adhered to or distorted the *Reinventing Government* playbook—all the while recapturing public attention on its legacy (Nix 2025; Bruggeman and Eilbert 2025; Weinstein 2025; Menchaca 2025; Hubbard 2025; Bilmes et al. 2025; Hawickhorst 2025; Tonks 2025).

Performance measurement has spread beyond the U.S. federal government. Lu et al. found that "the past 20 years represent one of the most remarkable eras for [municipal] performance budgeting initiatives in the United States" (Lu et al. 2015, 426). By 1999, 38% of municipalities in America with populations of more than 25,000 used some form of performance measurement, while 23% had adopted a citywide performance measurement system (Poister and Streib 1999). Performance-based budgeting is legislatively mandated in 40 states, up from five in the 1980s (Lu and Willoughby 2012). Globally, growth in national-level performance indicators has been "nearly exponential" since the late 2000s, spurred by international actors, such as the World Bank and Transparency International, to reverse "failing" development reforms and government structures (Kelley and Simmons 2020, 3).

Organizational theorists have long studied the role of performance measurement systems such as Sunnyvale's, noting how a curated recording of history surfaces the inherent difficulty of learning from past outcomes for the purpose of reforming future actions, while conferring long-term legitimacy to organizational routines through its stability across individual tenures (Levitt and March 1988). It is therefore all the more surprising that, to our knowledge, for the model Sunnyvale system which lent its early legitimacy to so many emulators, no follow-up study from an organizational change perspective exists in the literature. Put simply, the following question has not been asked or answered: Did Sunnyvale continue to live up to its reputation as a paragon of performance measurement? It is with this context that we conduct a longitudinal analysis of 40 years of Sunnyvale municipal budget documents to understand how performance measurement was actually practiced in the city.

This article makes two contributions to the study and practice of performance measurement. First, in contrast to conventional wisdom, we show that Sunnyvale's system deteriorated due to loss of leadership and managerial buy-in. We document inconsistent reporting, mutable indicators, and accounts of gaming from department heads and line staff. Over time, the system shifted from a decision-making tool into a symbolic system that required substantial resources to maintain, but was largely ignored by decision makers. This study challenges prevailing narratives about the efficacy of performance measurement as a mechanism for bureaucratic reform. Sunnyvale is not a success story to be emulated, but a cautionary tale that highlights the fragility and limitations of government performance measurement.

Second, we introduce a digital archival approach to study the micropractice of performance measurement by analyzing

thousands of pages of archival municipal budget documents longitudinally across 40 years. We tracked 152 performance indicators, their associated target values (representing the target goal), and reported achieved values (reporting actual achievement) over the time period, noting changes in wording or when the indicators disappeared. This approach allows for an assessment of the actual implementation, trends, and changes of Sunnyvale's performance measurement system over the 40-year period. We supplement this archival analysis with interviews with experts directly familiar with Sunnyvale's system. As more municipalities digitize their budget and performance data, this study demonstrates a novel methodology for the study of public administration and regulation.

The remainder of this article is structured as follows. Section 2 examines theoretical frameworks underpinning performance measurement adoption in government, and the limited empirical evidence for linkages between performance measurement and improved government efficiency. Section 3 describes our research methodology, which includes an archival document analysis and interviews with Sunnyvale employees. Section 4 details our findings. Section 5 discusses implications for the field of performance measurement, particularly three trade-offs that long-running performance measurement systems face: quantity-quality, incentivizing progress towards the right outcomes versus merely the most quantifiable ones; continuity-flexibility, ensuring stability versus allowing evolution of measures; and accountability-productivity, time spent on reporting versus implementation. Section 6 concludes.

2 | Prior Work

This section examines prior work on performance measurement, and is structured as follows. In Section 2.1, we examine two theoretical frames for understanding performance measurement in government, principal-agent theory and organizational theory, and then discuss three management trade-offs described within performance measurement scholarship. In Section 2.2, we examine empirical evidence of the link between performance measurement and government efficacy, highlighting gaps in prior empirical methodology. In Section 2.3, we summarize governmental and regulatory documentation on GPRA and Sunnyvale's performance measurement system.

2.1 | Performance Measurement Theory and Its Challenges

Performance measurement is classically understood through a principal-agent framework (Moe 1984) in which political principals (e.g., elected officials), instead of micromanaging the actions of agents (e.g., bureaucrats), can more efficiently achieve their aims through outcome monitoring (e.g., tracking progress towards targets), and high-powered incentives (e.g., performance bonuses) tied to specific, desired outcomes (Holmstrom and Milgrom 1991). While principal-agent theory provides a foundation for why performance measures are initially adopted, organizational theory adds a framework for how measures and goals can evolve or endure over time within a bureaucracy, for

example, to provide organizational legitimacy through ongoing routines (Levitt and March 1988; Meyer and Rowan 1977; Scott 2013). In particular, Argyris and Schön (1978) distinguish between single-loop learning, simply adjusting effort levels to meet set goals as is understood through principal-agent theory, and double-loop learning, where an organization may revise and change goals and metrics as it processes new information.

However, some scholars argue that existing theories of performance measurement and organizational learning leave out additional complexities that reflect the nuances of real-world decision-making within complex bureaucracies (Dixit 2002; Tirole 1986; Moe 1984). This study highlights three such complexities in the form of trade-offs. First, performance managers face a “quantity-quality” trade-off: balancing agent incentives towards the right goals (which are usually harder to measure) against incentives towards the most observable and quantifiable outcomes (D. E. Ho and Sherman 2017; Holmstrom and Milgrom 1991; Kerr 1975). Poorly designed incentives induce gaming (Bevan and Hood 2006), as has been documented in myriad public sector contexts from education (Cullen and Reback 2006; Figlio and Getzler 2006) to healthcare (Bevan and Hood 2006) to public spending (Hood and Piotrowska 2021). Holmstrom and Milgrom (1991) argue that, in complex bureaucracies, the optimal incentive scheme could actually be low-powered (i.e., a salary-based job instead of compensation based on performance bonuses) to encourage balanced effort across tasks, and that, by holding incentives equal, agents also tend to allocate their efforts towards tasks with lower marginal costs. Additionally, scholars of stewardship theory suggest that entrusting agents with increased levels of authority and autonomy incentivizes more intrinsic effort towards desired goals (Moynihan 2008).

Second, performance managers must balance the “continuity-flexibility” trade-off: balancing measurement continuity for longitudinal assessment (Heckman et al. 1997) against the constant evolution of the measures themselves to reflect changed conditions and objectives (Heinrich 2002). Whether performance managers are able to achieve this long-term balance is laborious to study, as most empirical studies of performance measurement only span several years due to the constant “churn” of performance indicators within performance regimes (Pollitt 2011). Hood and Dixon argue that this churn is caused by governments changing administrative records, which make “before-and-after comparisons” challenging, going on to argue that research into such “data breaks” can provide insight into a government's “changing priorities, preoccupations, and ways of working over a generation” (2015, 47–48). Our study attempts to fill this gap by recreating performance trendlines for 40 years—including such data breaks—and providing related insights from interviews and other documents to highlight the city's changing priorities.

Third, such a system must optimize the “accountability-productivity” trade-off: time spent on costly agent reporting versus time spent on other tasks (Jensen and Meckling 1976; Townsend 1979). While elected officials tend to gain the political benefits of government reforms, the costs of tracking performance associated with such reforms are borne mostly by bureaucrats (Moynihan 2008).

2.2 | Empirical Assessment of Performance Measurement

Despite the widespread use of performance measurement to improve government, empirical studies on its efficacy have been limited and offer mixed results.

First, some scholars use budget documents to draw conclusions about government officials' usage of performance information in decision-making. Gilmour (2005) finds that, while PART scores were significantly correlated with OMB budget decisions, the sub-scores measuring program *intent* had a larger impact than those measuring program *results*. Gilmour and Lewis (2006, 742) show that PART scores for "traditionally Democratic" departments had a stronger influence on budget allocation. Ho (2011) uses budget data from the city of Indianapolis' IndyStat program, finding a positive relationship between the performance measures a city department reports to IndyStat and budget changes in the following year. Such empirical studies, although showing a positive correlation between information reporting to budget changes (as a proxy for information usage), do not investigate the process and costs of the information reporting—and implicitly, the cost–benefit analysis of the reporting process. Similarly, widely employed guidelines, such as Bloomberg Philanthropies' What Works Cities Certification,⁵ often focus on the benefits of adopting performance measurement, without expressing the operational costs to the agency or city of developing and maintaining such a system. This study surfaces many of those underappreciated costs through both interview insights and the substance of the digital archival analysis.

A second research approach has been to survey government officials. Melkers and Willoughby survey city and county administrators and budgeters from nearly 300 governments: their "[f]indings indicate the use of performance measurement by local departments is pervasive, although survey respondents are less enthusiastic about measurement effectiveness" (Melkers and Willoughby 2005, 180). Sanger (2008, 81) interviews employees from six cities employing performance measurement, finding that "[p]erformance measurement is growing in states and local governments, but more often without the engagement of citizens and with unrealized use for management." Moynihan and Lavertu (2012, 598), analyzing survey responses from federal employees about their use of GPRA and PART scores, argue that government-wide performance reforms have been most effective in "encouraging passive forms of performance information use" where employees simply report performance information, without using the information to guide decision-making. This study uses interview insights but combines those with digital archival analysis to showcase the actual operationalization of performance measurement.

Third, other studies have employed state-level variation to descriptively characterize differences in adoption and study effectiveness of performance measurement. Lu et al. (2011), for instance, descriptively characterize performance budgeting laws in all 50 states, finding that states with such laws are graded higher by the Government Performance Project. Others employ a comparative case study method, describing the features of a variety of performance management programs in an attempt to determine which features are most effective (Holzer

and Yang 2004). Such observational designs, however, are limited in drawing causal inferences. While at least one study has investigated the effect of organization-level performance measures on individual employee behaviors through a regression discontinuity design (Kim 2025), we are not aware of any RCTs focused on the effects of performance information usage on future government performance, likely given the challenges of implementation. Furthermore, synthesis of performance measurement literature notes that much of the existing literature has focused on whether an organization possesses performance measurement "trappings", that is, the elements commonly associated with a performance regime, rather than on the use and usefulness of these elements (Ammons 2022). This article documents microdecisions made by Sunnyvale stakeholders to show how common elements of a performance system were slowly discontinued over time.

Fourth, empirical studies conventionally cover only short (2–3-year) periods. This study examines differences over time, and across programs and management, to show the difficulties of maintaining a performance measurement system over decades. In one recent example, Rivenbark and Vignieri (2025) look at operating budget changes from 3 U.S. municipalities, comparing FY1994–95 operating budget to FY2021–22 for three departments within the municipalities: fire services, solid waste, and human resources. They conclude that the three municipalities have made progress by shifting from "output" indicators to "outcome" indicators, but continue to use an ad hoc approach for developing the best performance metrics (Rivenbark and Vignieri 2025). Although the authors extract valuable insights by comparing two separate budget years with a gap of 30 years, our study compares budget information for a full 40 years within one city; we demonstrate that such a deep dive uniquely surfaces long-term changes and challenges in performance measurement.

2.3 | Legal and Institutional Context

The Sunnyvale playbook was developed by Tom Lewcock, Sunnyvale city manager from 1979 through 1998, who was described by former employees as the "mad genius" behind the system. Under Lewcock, Sunnyvale, which operates under a council-manager system, developed long-term strategic plans, including projecting the city's operating budget and revenues up to 10 years in the future. To be included in an upcoming budget or strategic plan, important issues were first collected by city employees and then ranked by city council members. The system also matched every program activity within a city department to a budget line item in its performance measurement system.

Program activities, in turn, had specific, stated objectives and annual targets. Sunnyvale's budget items and related performance metrics were published in the city's annual operating budget and discussed at length within city council meetings. As mentioned in *Reinventing Government*, because of this bottom-up reporting of performance metrics tied to budget items, "Sunnyvale's city council can predict quite accurately the results it will achieve...[the city] gives its decision makers the information they need to make intelligent decisions" (1992, 144). Ultimately, the Sunnyvale system provided its city council with

greater transparency into the inner workings of the city. For city employees, the incentive was twofold: increased managerial flexibility and financial incentives provided to Sunnyvale's managerial staff for meeting their targets, "up to 10 percent of their salary" (Osborne and Gaebler 1992, 145).

GPRA was designed by John Mercer, a former Sunnyvale mayor and long-time council member, dubbed the "father of GPRA" (Radin 2006, 119), with key elements taken directly from the Sunnyvale playbook. Under GPRA, federal agencies are required to (1) develop five-year strategic plans with long-term goals covering each of their major functions, (2) prepare annual performance goals for each fiscal year specifying their annual goals and performance indicators towards those goals, and (3) report their annual performance towards these goals (Sen. Roth Jr. 1993). In return for agencies abiding by these elements, GPRA promised greater managerial flexibility by waiving certain administrative controls as proposed by pilot agencies that met their performance goals. As OMB analysts noted, flexibility was an essential incentive for managers to change their behavior and to accept greater accountability under the new act (Groszyk 1995).

Some scholars were more skeptical of GPRA, and by proxy, the Sunnyvale system. Mashaw argued that GPRA focused on improving "management" through setting objective metrics while ignoring "the interpretive enterprise" of "governance" (Mashaw 1995, 412). Radin observed that GPRA's "use of administrative rhetoric has caused it to collide with institutional, functional, and policy/political constraints that are a part of the American decision-making system," resulting in "cynicism and a compliant attitude within the federal government" (Radin 2000, 133; Radin 2006, 126). She argued that a "government-wide" approach for reform led to a promotion of "a set of generic activities and requirements... that do not really touch the core of the nation's decision-making processes" (Radin 2000, 112). Notably, by 1997, none of the 61 waiver proposals submitted by pilot agencies seeking greater managerial flexibility were accepted by OMB under the GPRA pilot waiver process (U.S. Government Accountability Office 1997). Others pointed out the complexities of holding federal agencies accountable to performance standards for work that is increasingly being done by third party contractors (Frederickson and Frederickson 2006). This study contributes to the literature on the efficacy of GPRA by examining the actual practice of performance measurement within its prototype, Sunnyvale.

3 | Methods

We employ a mixed-methods approach for studying Sunnyvale's performance measures over time. Section 3.1 details our quantitative document research, and Section 3.2 describes our methodology for interviews with employees and elected officials. This research design provided several advantages over relying solely on either quantitative or qualitative methods alone. First, our archival research enables an objective measure of the micro-usage of performance measures over time, but does not provide contextual insights into implementation challenges and strategic motivations of those tasked with performance measurement, which can be complementarily surfaced through semi-structured

interviews. Second, the substantive findings of the quantitative analysis provided rich material that we were able to directly use as inputs for informing the direction and depth of our qualitative analysis, such as having interviewees directly react to the archival results. Third, the integration of two analyses allowed us to cross-validate key takeaways, in that empirical evidence from historical records corroborated the recollections of our expert interviewees, and vice versa. This convergent mixed-methods design (Creswell and Plano Clark 2018) was chosen to achieve methodological triangulation, whereby the qualitative and quantitative strands mutually corroborate and contextualize one another (Denzin 1989; Jick 1979).

3.1 | Archival Research

The indicators, targets, and achieved values for every Sunnyvale department are reported in their annual operating budget documents. At the beginning of this project, archived budget documents were paper-based and only available in Sunnyvale's archives, requiring in-person visits. By 2024, these documents had been digitized and made publicly available on Sunnyvale's city website, accelerating the analysis. For each year of analysis (1984–2023), the reports for selected departments were read, and values and targets for that year's performance indicators were documented. To identify how one performance indicator changed from fiscal year (FY) to the next, each individual indicator was tracked over time, including any associated goals and achieved values reported. In summary, budget documents totaling over 35,000 pages were reviewed.

Coding guidelines were created at the beginning of the review process, and then updated collaboratively for situations falling outside the existing guidelines. Each department area was coded by one reviewer who reviewed each year's annual budget document and coded each indicator in a spreadsheet following said guidelines. All coding was then reviewed by a second reviewer. This procedure—an a priori codebook refined iteratively, with independent double-coding to establish intercoder reliability—follows established standards for systematic content analysis (Krippendorff 2019; Saldaña 2016). Tracked indicators and further details on coding methodology are available in the Appendix S1.

An example of a page of reported indicators for Public Works' Street Maintenance program (Program 112) for FY 1995 is shown in Figure 1. Each indicator has an associated code, units of measurement, and three columns for the indicator's reported or target values for subsequent years from FY 1994 through FY 1996. For example, for the indicator "112K-021: Number of complaints concerning dirty walkways/roadside areas received," the department had received four complaints for the previous year (FY 1994), while the targeted number of complaints for the following 2 years (FY 1995–96) was 7. The indicators are grouped into activity programs (e.g., road safety, tree maintenance, hiring and recruitment), then assigned to particular city departments such as Public Works.

Performance indicators that were reported, updated, or added for each year were manually documented for (1) street and road quality in Public Works (68 indicators) and (2) employee recruitment in Human Resources (73 indicators). We were limited to selecting

1IND 04/24/95		CITY OF SUNNYVALE		PAGE 3	
PROGRAM 112		PERFORMANCE INDICATORS			
PERF IND #	PERFORMANCE INDICATOR STATEMENT		93 - 94 ACTUAL	94 - 95 PLAN	95 - 96 PLAN
PROGRAM 112 STREET MAINTENANCE					
112A-02	MAXIMUM PERCENT OF STREET AREA RECONSTRUCTED PRIOR TO COMPLETE FAILURE (CONDITION RATING OF 65).	UNITS %	100.0%	100.0%	100.0%
112A-03	PERCENT OF RECONSTRUCTION COMPLETED BASED ON PLANNED UNITS FROM BI-ANNUAL CONDITION SURVEY.	UNITS %	100.0%	100.0%	100.0%
112B-02	MINIMUM PERCENT OF STREET AREA PLANNED FOR OVERLAY WHICH RECEIVES AN ASPHALTYIC CONCRETE OVERLAY PRIOR TO ITS DECLINE TO BELOW A VERY POOR CONDITION (CONDITION RATING OF 70).	UNITS %	100.0%	100.0%	100.0%
112B-03	PERCENT OF OVERLAY COMPLETED BASED ON PLANNED UNITS FROM BI-ANNUAL CONDITION SURVEY.	UNITS %	0.0%	100.0%	100.0%
112C-02	MINIMUM PERCENT OF STREET AREA PLANNED FOR CHIP SEAL WHICH RECEIVES A CHIP SEAL WITHIN TWO YEARS OF ITS DECLINE TO A POOR CONDITION (CONDITION RATING OF 75).	UNITS %	---	100.0%	100.0%
112C-03	MAXIMUM PERCENT OF STREET AREA REQUIRING PERMANENT PATCHING PRIOR TO APPLYING THE CHIP SEAL.	UNITS %	---	10.0%	10.0%
112C-04	PERCENT OF CHIP SEAL COMPLETED BASED ON PLANNED UNITS FROM BI-ANNUAL CONDITION SURVEY.	UNITS %	---	100.0%	100.0%
112D-02	PERCENT OF SLURRY SEAL COMPLETED BASED ON PLANNED UNITS FROM BI-ANNUAL CONDITION SURVEY.	UNITS %	105.0%	100.0%	100.0%
112E-01	MINIMUM PERCENT OF STREETS WHICH RECEIVE CORRECTIVE PAVEMENT MAINTENANCE WITHIN 2 YEARS WHEN A CRACKING SCORE OF 8 OR MORE IS ACCUMULATED ON A CONDITION RATING SURVEY.	UNITS %	85.0%	85.0%	85.0%
112E-02	PERCENT OF TIME THAT A RESPONSE IS MADE TO A HAZARDOUS PAVING FAILURE WITHIN 8 HOURS OF REPORT.	UNITS %	100.0%	100.0%	100.0%
112E-03	NUMBER AND PERCENT OF CLAIMS RESULTING FROM STREET PAVEMENT FAILURES AND HAZARDS AS A PERCENTAGE OF THE PREVIOUS THREE (3) YEAR AVERAGE NUMBER OF CLAIMS.	UNITS %	---	4 100.0%	4 100.0%
112K-01	NUMBER OF COMPLAINTS CONCERNING DIRTY WALKWAYS/ROADSIDE AREAS RECEIVED.	UNITS %	4 ---	7 ---	7 ---
112K-02	PERCENT OF WALKWAYS/ROADSIDE AREAS THAT, COMPARED TO PHOTOGRAPHIC STANDARDS, ARE RATED AT A LEVEL OF "SATISFACTORY" OR BETTER.	UNITS %	90.0%	80.0%	80.0%
112O-01	PERCENT OF STREET AREAS COMPARED TO PHOTOGRAPHIC STANDARDS THAT ARE RATED AT A LEVEL OF "SATISFACTORY" OR BETTER.	UNITS %	90.0%	90.0%	90.0%

FIGURE 1 | An example of a performance indicator statement within Sunnyvale's operating budget document, FY1995. The selected statement shows indicators for Public Works' Program 112: Street maintenance. The column on the left lists 14 street maintenance performance indicators, while the columns on the right show actual achieved performance for FY1993 and target performance values for FY1994 and FY1995. Dashed lines indicate missing or non-relevant values.

two departments due to the extensive effort required for archival documentation, and without the ability to base our particular selections on contrasting outcomes of interest *ex ante*, we instead selected Public Works and Human Resources because they focus on distinct domains (physical infrastructure vs. human workforce) and constituencies (resident-facing vs. employee-facing) while being relatively standard departments across municipalities. Information regarding (3) performance indicators that were explicitly provided as examples by Osborne and Gaebler in *Reinventing Government* were collected into a third group of analysis (11 indicators) so as to particularly scrutinize the indicators that were publicly showcased as most emblematic of Sunnyvale's performance measurement acumen. Selected examples of performance indicators are listed in Table 1 below, and all tracked indicators are documented in the Appendix S1.

The information from the archival documents were coded as follows. First, an indicator was considered "continuous" from year to year if no significant changes were made to the indicator in the following year, even if its indicator code, target value, or indicator wording differed but the substance of the indicator remained unchanged. Conversely, an indicator was coded as "ended" if the indicator was marked "deleted" and/or no substantively similar indicator was introduced in the following year. Detailed examples of coding decisions are available in the Appendix S1.

Second, each indicator's reported value and target value for each year was collected. Third, the directionality of an indicator was coded, that is, whether the objective is to increase or lower the value. For instance, for indicator 112O-01, the goal is to increase the percentage of street areas that are rated as "satisfactory" or above, hence the directionality was coded as positive. Such coding allows characterization of whether changes to targets constitute "raising" or "lowering" of the performance bar. For example, if in FY 1995, indicator 112O-01 had a target value for FY 1996 of 90% streets rated as satisfactory, but in FY 1996 the target value was later revised to 80%, and the directionality of this indicator is positive, then this revision constitutes a relaxation of ambitions. Some indicators were coded as neutral directionality, that is, an indicator without a specified goal (e.g., "Miles of arterial streets," or "Number of regular new hires").

3.2 | Qualitative Interviews

The findings from this quantitative analysis were then contextualized through expert interviews (Bogner et al. 2009) with 11 individuals, all of whom were selected because of their roles as current and former Sunnyvale employees or elected officials and their direct involvement in the creation, reporting, and usage of performance measurement in Sunnyvale. Participants were contacted from January 2020 through May

TABLE 1 | Examples of performance indicators tracked between 1984 and 2023, by department. Indicators labeled as “Reinventing Government” were highlighted by Osborne and Gaebler (1992).

Department/Group	Indicator	Start year (FY)	End year (FY)
Human resources	Number and percent of individuals who rate training received as being “satisfactory” or higher.	1985	2024
	Number and percent of female managers in city workforce in relationship to percentage of women in community population (49.4%)	1993	2003
	Number of employees who received a flu shot.	2014	2024
Public works	Percent of street areas compared to photographic standards that are rated at a level of “satisfactory” or better.	1984	2005
	115: Preventive maintenance is performed as scheduled 95% of the time	2000	2005
	P/WL: Miles of collector and residential streets in the City.	2011	2024
Reinventing government	Percent of participants that rank the recreational swimming at good or above.	1989	1995
	Percent of downtown parking district trees which are replaced for various reasons within 2 months of notification determined by divisional plant material replacement forms.	1990	2005

2024, using publicly available emails, and interviews were conducted in 2020 and in 2024. Snowball sampling (Parker et al. 2020) was utilized to identify more potential expert interviewees after the initial seed. A total of 17 individuals were contacted, 6 of whom, including 4 current employees or elected officials as of writing, either did not respond or declined to interview. Interviews lasted between one and two hours and were conducted over Zoom. A list of the anonymized interviewees and their positions in the Sunnyvale administration is available in the Appendix S1.

Qualitative literacy criteria (Small and Calarco 2022) were used to develop our semi-structured interview guide, which is provided in the Appendix S1. The interviews were guided by three main areas of inquiry: (1) the process of tracking, monitoring, and reporting performance information within the system; (2) to what extent the performance indicators were used in day-to-day decision-making; and (3) whether there were other non-publicly available performance information sources that were used by managers or councilmembers to make decisions. Additionally, the initial results of the quantitative analyses were presented to the 2024 interviewees for their reactions and for any context they could provide. Interviews were documented through contemporaneous handwritten notes, which were later analyzed collectively to identify and refine cross-cutting themes, coded by the same analyst into the finalized themes, and lastly exemplified through curated excerpts. The research was approved by the Institutional Review Board.

3.3 | Statistical Analysis

We use simple χ^2 tests to test substantive hypotheses. First, to test whether reporting varied across city managers, we employ a two-tailed χ^2 test of homogeneity of reporting rates across city managers. Second, to test whether targets may have been

strategically revised based on the prior-year performance relative to the target, we also employ a two-tailed χ^2 test. Specifically, we test for whether the direction of target revision in a given year (coded as no change, increase, or decrease) was independent of whether the target had been met, exceeded, or missed in the preceding year.

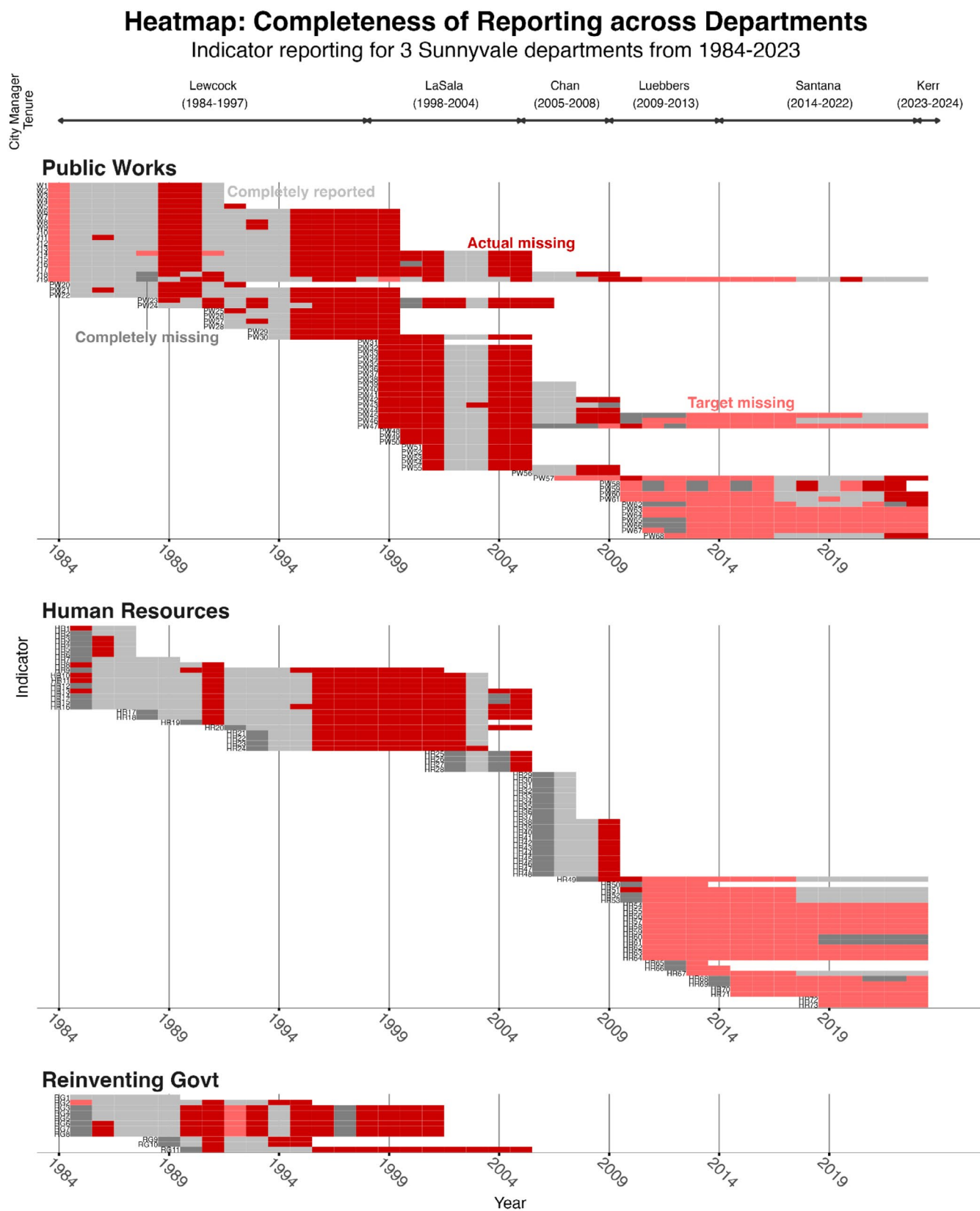
4 | Results

Sunnyvale's performance indicators changed frequently over time. Moreover, targets and performance values were inconsistently tracked and reported. Detailed analysis of indicators also shows substantial changes in target levels over time.

4.1 | Indicator Phases and Disappearing Weights

A detailed analysis of indicators showed certain “phases” that highlight the shifting priorities within the city. In 1984–1999, indicators were based on specific, minute descriptions of tasks, what some interviewees later called “widget-counting.” In 2000 through 2011, indicators were focused on customer satisfaction, while from 2011 to present, indicators split into “Performance” and “Workload.” Performance indicators tracked department productivity and performance (e.g., percent of calls responded to within a certain amount of time), while workload indicators tracked static metrics that indicate the amount of work completed by the department (e.g., miles of sidewalk in the city).

From 1994 to 2006, Sunnyvale also included weights (1 to 5) for each of the reported performance indicators to show its relative priority to city council and to calculate an overall weighted performance index for each program (City of Sunnyvale, 1996). However, in contrast to rhetoric around



Source: Sunnyvale budget documents, 1984 - 2023

FIGURE 2 | Heatmap of Sunnyvale's performance indicators as reported between 1984 and 2023. Each section shows performance indicators ($n = 1637$) for one of three groups tracked. A row represents a performance indicator, and a horizontal row cell represents 1 year in which a performance indicator was reported. Red years indicate that all values were fully reported, incompletely recorded years are noted in light red (target value missing) or dark red (achieved value missing) or dark gray (completely missing). The figure shows that target and achieved values were only reported 33% of the time.

the system, none of the budget documents reviewed reported a composite performance index. In 2007, indicator weights were substituted with “priority” legends for each indicator, ranging from M (Mandatory), to C (Council Highest Priority), I (Important), and D (Desirable). Priority legends were last reported in 2009, and no longer appeared in any documents reviewed from 2010.

4.2 | Impermanence of Indicators and Incompleteness of Reporting

On average, a performance indicator was tracked for 11 years across the 40-year period before being replaced by new indicators reflecting changed priorities. Within each department, an indicator was recorded for an average of 12 years in the Public Works, 9 years in Human Resources, and 13 years in the *Reinventing Government* set. Between 2005 and 2010, the average duration of an indicator dropped to 3 years. In some cases, indicator tenure was so short that no achieved values were ever reported.

Second, even when indicators were reported in documentation, targets or achieved values were not reported. Each of the heatmaps in Figure 2 represents a set of performance indicators tracked for Public Works, Human Resources, and those highlighted in *Reinventing Government* between 1984 and 2023, respectively. Each row represents the reporting timeline of a single performance indicator, and the cell colors represent reporting status for a given year. An indicator is marked “completely reported” in gray for a year if the indicator, the target, and value were reported that year. If an indicator was present for that year but incompletely reported, those instances are recorded in either light red (target missing) or dark red (reported achieved value missing). If an indicator went completely missing in between reporting instances, those instances are recorded in dark gray. This figure illustrates the piecing together of such “data breaks” discussed by Hood and Dixon (2015), lending insight into the city’s changing priorities over time.

The 152 performance indicators tracked had a total of 1637 reporting instances across the 40 years. Within those reporting instances, 67% of the instances were incompletely reported, with 385 instances missing a target value and 596 reporting instances never having an achieved value reported, while 112 instances had neither a target nor achieved value reported.

Reporting practices under each city manager were compared. Under each city manager’s tenure, there were significant differences in missingness for reported indicators. Tom Lewcock had the highest completeness of reported values (53.3%), whereas under Gary Luebbers’ tenure, 136 out of 137 indicators reported had at least one missing value. We tested whether completeness of reporting was equivalent across city managers. A χ^2 test of homogeneity indicated that the proportion of missing reporting differed significantly across city managers ($\chi^2(5) = 234.63$, $p < 0.001$).

Our interview insights also provide evidence that the individual city manager greatly affected the tenure of the indicators. For example, interviewees noted that when Amy Chan, a protégé of Tom

Lewcock’s and long-time finance director, took over as Sunnyvale city manager in 2005, she revamped all performance indicators to be focused on customer and satisfaction ratings. These revamped indicators, however, only lasted several years in practice. When Gary Luebbers became city manager in 2009, they were discontinued, and substituted with new indicators in 2010.

Even within Lewcock’s tenure as city manager, the performance indicators underwent a complete change between 1998 and 1999 (as evidenced in our heatmap) with the discontinuation of a majority of previously-reported indicators and introduction of new indicators. According to interviews, this shift was due to Lewcock’s attempts to move towards more “outcome-oriented” indicators, based on feedback from city employees that the previous indicators were too granular and process-oriented. These observations underscore the importance of city leadership to maintain use of a performance measurement system, and also the fragility of such a system’s longevity over time. As each leader attempts to solve for the political necessities of their era, these decisions deteriorate any through-line measures by which a government’s long-term performance can be evaluated.

4.3 | Inconsistency of Targets

Targets were revised frequently over the period of analysis. If the target was met in a certain year, the target remained the same in the second year 86% of the time. When a target was not met, Sunnyvale lowered the target in the second year 36% of the time. When Sunnyvale outperformed the target, the bar was lowered only 12% of the time. This association is statistically significant ($\chi^2(4, N = [488]) = 70.03$, $p < 0.0001$), refuting the notion that target value revision was independent of the likelihood of achieving the target value.

As an example of a target revision in Public Works, in 2009, 4 targets were changed—2 targets lowered, 2 targets raised—after achieved values were reported and taken into consideration. The department relaxed from 70% to 68% the target for “Customers satisfied with the quality of roadsides and medians” after a customer satisfaction survey showed that the current percentage of satisfied residents was 68%. Meanwhile, a target cost per irrigation price was increased from \$41 to \$55 to match the reported achieved value from the previous year. In other instances, the department “raised the bar” by strengthening a target after performance had exceeded the target: Public Works increased the target number of calls responded to within a 3 h window (target of 50 in 2007, increased to target of 150 in 2009) after their performance in 2007 had surpassed the target (achieved value of 175 calls in 2007).

4.4 | Influence of Leadership

Interviewees noted that employee usage of the performance indicators, and the system as a whole, was dependent on city manager prompting and city council’s continued requests for transparency. Over time, both of these influences lessened. City manager support was highly dependent on the individual city manager. Tom Lewcock, city manager at the time of *Reinventing Government* and chief architect of the Sunnyvale system, had

developed strong buy-in across Sunnyvale employees for the system. During Lewcock's tenure as city manager (1979–1998), city employees used the indicators for tracking their work tasks and budget expenditure. Lewcock ensured that crucial performance measurement elements—including bonuses for managerial staff and regular performance indicator discussions with city council—were in place to incentivize system usage.

In contrast, interviewees noted that Lewcock's successor, Robert LaSala (1998–2004), did not consider the Sunnyvale system a priority. One councilmember noted: “LaSala didn't like to be burdened with numbers; he liked to make decisions and have them carried out.” LaSala's lack of transparency in the city budget and other managerial quirks upset Sunnyvale's city council and employees, resulting in low employee morale and staff turnover (Wilkinson and Harper 1999). In 2003, he was ousted through a city council vote (Van Sant 2008). Arguably, LaSala's refusal to invest in the performance measurement system—and the resulting reduction in transparency to city council—contributed directly to his firing. “LaSala let all of Tom Lewcock's work go down the drain”, said one former councilmember, who was present during LaSala's firing process. The councilmember noted that LaSala's budget items were reported in aggregate, not in detail, so that city council was unable to know the actual amount of money spent on each budget item.

4.5 | Loss of Performance Incentives

Interviewees pointed out that incentives that had buoyed the system in the early days had slowly disappeared. First, pay-for-performance for managerial staff, which was based on meeting performance targets, was removed completely in 2005 due to the economic downturn. In 2002, the city manager review process was simplified, and any direct linkages to performance indicators were removed. Second, the performance bonus for the city manager was also discontinued. By 2008, Amy Chan's bonus was decided based on a successful real estate negotiation and her “savvy with the city budget” (Kraatz 2007), with no mention of performance indicators. Without incentives, the Sunnyvale system had lost the explicit linkage between the city's day-to-day tasks and its long-term planning and performance indicators.

4.6 | Gaming

Interviewees also provided context for the gaming of performance measures. In the late 2000s, department heads took advantage of the lack of city manager direction and managerial attention to take a more *laissez faire* approach to indicator reporting. One former finance director noted that they began to notice gaming behavior during LaSala's tenure as city manager. Department heads and employees would edit their financial reporting, which was linked to their performance indicators, in order to access more budget allocations. If they exceeded their target budget on a particular activity, they would charge future spending to a different performance area, similar to “creative categorization” behavior (Hood and Piotrowska 2021). The finance team would try to catch and flag these behaviors: “[We] would go through the budgets with program managers, and highlight for the city manager anything that was in dispute... We

were trying to be as robust as possible.” However, this was difficult without the strong backing of department heads. As the finance director mentioned, “If [the department head] were a good manager, [they] could look at the [Computer-Aided Dispatch] or the utility billing manager and figure out if [their team] were charging the right things...if [the managers had managed their team correctly], gaming would be difficult.”

Gaming was not limited to employees. One interviewee noted that in the early days of the Lewcock system, changing targets and indicators had required extensive discussions between department heads, the city manager, and city council. In later years, as leadership lost interest in the system, department heads found it easier to convince the city manager to drop or change indicators or targets at their discretion. Changing a target or indicator became a trivial matter. These observations corroborate the target change analysis, showing that employees had sufficient leeway to change their targets based on prior performance.

4.7 | Disproportionate Costs

Osborne and Gaebler praise Sunnyvale for using thousands of measures, which were tracked by ground-level staff, and collated and managed by managers (Osborne and Gaebler 1992, 145). Former Sunnyvale councilmembers describe the *Reinventing Government*-era Sunnyvale system at the time with great positivity, emphasizing that the system allowed councilmembers to determine if the city was delivering what it was meant to, calling the system a “wonderful mechanism”. Employees, however, described the system as “extremely time consuming.” The interviewees frequently mentioned the “substantial” costs of implementing performance measurement. One interviewee referred to the system as “overkill,” another as “way too complicated.” As one former employee put it: “The system became more and more cumbersome, more integrated ... A lot of it was widget counting, numbers of this, miles of that.” Even after Lewcock's “right-sizing” of the system in the late 1990s to more outcome-oriented indicators, tracking the performance indicators was still a tedious process. Performance measurement reporting continues to take up a substantial part of the Sunnyvale budget document, with hundreds of pages of the annual operating budget dedicated to publishing the city's metrics.

4.8 | Symbolic Use of Performance Indicators

Such costs of reporting as described in the previous section may be warranted if stakeholders use the produced performance information for decision-making. But city councilmembers interviewed more recently noted that the information is no longer useful in their decision-making processes. Reading through the multitude of performance indicators was described as “making council's eyes glaze over.” From the mid-2000s onwards, city councilmembers rarely looked at the performance indicators, as they were buried in large budget documents. One councilmember noted that they “trusted the city manager and didn't even bother to know if staff carried out” the measures.

Councilmembers noted that they used signals other than performance indicators to make decisions—for example, talking to

constituents that are impacted by city services, and if necessary, assigning particular issues as “study issues” for the upcoming budget. As one councilmember stated, citizens evaluate the government not by numerical measures but by governmental services that they as individuals interact with. The change in city councilmembers’ reluctance to rely on performance information corroborates organizational critiques that performance information may be less actionable for decision-makers than its advocates believe (Moynihan 2008). Consistently, the archival research finds that while indicator metrics continued to be reported from 2009 onwards, targets and achieved values were frequently missing from budget documents.

Employees also corroborated that the system was functionally “dismantled” and eventually “unraveled.” One department head noted that, as the performance indicators continued to be imperfectly reported, the measures lost their “managerial value” and were rarely discussed by city employees or in city council. Certain department heads stopped reporting or tracking their measures completely. The former finance director also noted that employee time stopped being tracked in budgeting hours—which had been directly linked to performance indicators—removing another linkage between city activity and the performance measurement system.

5 | Discussion

This study presents several implications for research and practice. First, despite being the most widely-replicated and praised performance measurement system in America, Sunnyvale’s system in practice struggled with performance measurement over the last 40 years. The stark disconnect between the perceived success of Sunnyvale, both then and now (Richter 1993; Fudge 2013; Kettl 2023), and the actual unraveling of its system, which we were able to uncover through extensive archival analysis, underscores the critical role of historical research in reexamining and testing the validity of longstanding legacies in public administration. In much the same way as organizational theory posits that double-loop learning is invaluable within organizations to shape both effective performance and continual refinement of performance measures (Argyris and Schön 1978), researchers and practitioners can take an even broader view and *measure the performance* of performance measurement systems as a whole. Such a “third loop” of learning might reveal important blind spots in our understanding of complex bureaucracies and chart the path forward for further scholarship. In particular, this study illustrates new opportunities for scholars of public administration and regulation in the increasing availability of digital archival records. Municipal documents (e.g., budget documents, public hearings, councilmember meetings) have become increasingly digitized, enabling assessments over longer periods of time and at more fine-grained programmatic levels than previously possible.

Second, leadership appears to play an important role in how such systems are actually practiced. In the early days of the Sunnyvale system, the city manager and city council carried an outsized influence in defining the targets and the incentive structure for bureaucrats, while city councilmembers in

particular enjoyed the transparency provided by the system. Over time, councilmembers became less eager to examine—and make decisions based on—the minutiae of the countless metrics. Later city managers also became less invested in maintaining the system in its original format, leading to inconsistent reporting and metrics without targets. Without strong leadership buy-in and usage, Sunnyvale’s system became a shell of its former self. Yet, possibly due to the system’s prior importance to the city council—see, for example, the firing of a city manager due to underinvestment in the system—the system persists in its diminished form, corroborating Levitt and March’s conception of a legitimizing routine (1988).

Similarly, media attention can also bolster incentives to pay attention to metrics and contribute to legitimization (Levitt and March 1988; Modell 2022). The Clinton-Gore fanfare drove waves of employees to Sunnyvale, and the system retains its reputation and respect within the city today, despite its lesser role in actual decision-making. Such media reporting—tying a city’s identity in the eyes of its constituents and beyond to performance measurement—can produce a lock-in effect, making it difficult for current managers to formally abandon an albeit ignored system, thus pressuring them to absorb the costs of superficial implementation. Recent research also shows that performance information presented through neutral third parties (such as the media), as opposed to coming directly from a government, is more influential and deemed more credible by the public (Meng and Li 2025), bolstering the likelihood that the Sunnyvale system’s positive depiction through the media created ongoing and important credibility for the city.

Finally, this study provides rich illustrations of three trade-offs that long-running performance measurement systems face. The first trade-off, quantity-quality, weighs incentivizing progress towards the right outcomes versus merely the most observable ones (Kerr 1975). Throughout the period of analysis, city leaders vacillated between different forms of performance indicators to spur employees towards certain outcomes, most starkly illustrated by Lewcock’s about-turn at the end of his tenure as city manager from “widget-counting” to “outcome-oriented” indicators, as he learned that the early indicators were causing employees to be overly focused on process. Our findings in Sunnyvale corroborate longstanding arguments by organizational theorists about the limitations of a classical conception of rational decision theory, favoring instead a sober skepticism that performance indicators share any consistent decision relevance (Feldman and March 1981). Herein lies an additional caution towards research efforts that over-rely on a quantitative approach to archival document analysis, where merely counting the prevalence of measures may miss qualitative shifts, like Lewcock’s, which we surfaced through our mixed-methods approach. Advances in natural language processing will increasingly augment the ability of quantitative methods to capture semantic meaning (or the lack thereof, in the case of widget-counting) in such public sector reporting contexts (Surani et al. 2025).

The second trade-off, continuity-flexibility, weighs ensuring stability versus allowing evolution of measures (Heckman et al. 1996; Heinrich 2002; Hood and Dixon 2015). Sunnyvale’s performance system has stayed in place for over 40 years—by many accounts an indicator of the regime’s stability—but as

our observations show, the system has continued despite lack of usage, inconsistent reporting, and constantly changing metrics. At first glance, this observation is mystifying. One might argue, like Argyris and Schön (1978), that such shifts in indicators were driven by intentional and strategic leadership decisions—a reflection of organizational learning. As one employee responded, to our questions about inconsistency: “There’s not anything subversive going on, these [changes] just reflect a shift in priorities.”

Yet our observations of indicators and measures disappearing without fanfare or discussion, inconsistent reporting, and accounts of gaming suggest that policies and practice may have instead become entirely “decoupled” (Meyer and Rowan 1977), that is, a performance measurement regime may persist to conform to external media and public expectations, but in practice does not actually bind the internal focus and flexibility of managers (Scott 2013). Such a decoupling, of course, implies the existence of practical barriers to connecting the dots, as we ourselves encountered in conducting our extensive archival research. Ultimately, we posit that Sunnyvale has not avoided the continuity-flexibility tradeoff, *per se*, but has found itself caught in a double-bind: the city must maintain the outward semblance of its performance regime for external legitimacy, while individual leaders have, over time, taken flexibility to such an extreme—by repeatedly changing the metrics and the signals of metric importance back and forth—that any present or future attempt to genuinely measure and manage performance may have lost its spring.

The third trade-off, accountability-productivity, weighs time spent on reporting versus implementation. While much research has focused on the benefits of performance measurement, this research highlights the need to understand the costs of performance measurement, notably to bureaucrats themselves. Stakeholders repeatedly emphasized that time spent on reporting and tracking was onerous, taking time away from actual program implementation and service delivery. Many expressed strong distaste for tracking and reporting, even when agreeing with the premise. Once incentives were lost and the benefits of the system rendered unclear, employees shifted their efforts to other tasks while still maintaining a modicum of data collection and reporting. Notably, the very model to counter high administrative costs of public oversight to the principal *generates its own* high reporting costs to the agency. These lessons are particularly important in light of contemporary reflections about how administrative processes can undermine state capacity (Bagley 2019; Lindsey 2021; Pahlka 2023; Klein and Thompson 2025). The findings imply that the global proliferation of performance measures since the 1990s has incurred unanticipated administrative costs that exacerbate state capacity challenges, an increasing area of concern in the present day.

6 | Conclusion

This study presents a novel digital archival approach that presents the first longitudinal, cross-programmatic analysis of a performance measurement system, namely its original exemplar in the heart of Silicon Valley, California. Over time, this model system actually deviated considerably from the very playbook it provided for national and local efforts across the globe, offering

up insights illustrating the inherent challenges and trade-offs within performance regimes. Using a digital archival method, we recreate Sunnyvale’s micro-practice of performance measurement across 40 years, while providing holistic insights into the evolution of the system’s leadership and protocol changes through interviews with key employees and elected officials.

While these findings provide rich insights into the practice of performance measurement in Sunnyvale—a central prototype in the field of performance measurement—this study has several limitations. First, the archival research is limited to public documents. City employees may have made decisions using performance information in non-publicly available documents, reports, and deliberations, but no evidence of such internal documents was found. Even if such documentation existed, the lack of reporting on annual budget documents raises serious concerns for transparency and accountability towards citizens.

Second, due to the complexity and time costs of the archival research, the selection of performance indicators and departments was limited. To address this, a variety of indicator types were selected, including some in Public Works that have concrete, real-world targets such as road maintenance, and others in Human Resources that have more qualitative targets, such as employee satisfaction.

Third, although a variety of perspectives from current and present Sunnyvale employees and councilmembers were included, some officials—such as the current city manager or councilmembers—did not respond to interview requests for this study. Our expert interviews also often required interviewees to recall events and work-related details far in the past (i.e., more than three decades), heightening the risks of memory distortions or biases. Indeed, in many instances, the interviewees recalled a greater completeness of performance indicator reporting in public records than we ultimately found through our archival research. Future work could expand the interview pool with more Sunnyvale experts and further triangulate their experiences with quantitative findings.

Fourth, while much evidence suggests a role for leadership, attributing the success of performance measurement to specific leaders is inferentially challenging. This remains a general challenge in public administration literature, and this study’s interview evidence paints a richer picture for the mechanisms of how leadership was seen to matter.

This depth of analysis enables a critical revision of conventional wisdom on performance measurement as a straightforward, cost-effective solution to bureaucratic reform, illuminating multiple difficult trade-offs practitioners must carefully consider if their systems are to stand the test of time. Ultimately, this study offers a fundamental note of caution: while inspiring generations of regulatory reform, Sunnyvale may also be a model for administrative process gone astray.

Acknowledgments

We are grateful to Claudia Moses and Karen Cho for their invaluable research support, Sam Sherman for kicking off this project, and

Lindsey Gailmaid for her helpful comments on this paper. We also thank our interview participants for their participation and time in this project.

Conflicts of Interest

The authors declare no conflicts of interest.

Endnotes

¹ This quote is usually attributed to Peter Drucker, but at least one source suggests that quote was apocryphal (Zak 2013). Nonetheless, the quote has long been one of the most widely known encapsulations of performance measurement.

² The basic concept of a performance measurement system is called by many names across the literature. Moynihan describes “performance management” as the “effort to link simple performance measurement with strategic planning” (2008, 31). Heckman et al. (1997) refer to a “performance standards system”, while Lu et al. (2011) call it “performance budgeting.” This article uses “performance measurement” to describe the system, while recognizing that the phrase can refer to the specific activity of measuring one area of performance.

³ <https://www.performance.gov/blog/reimagining-us-performance-reporting/>.

⁴ An example guide for federal grant applicants: <https://results4america.org/2024-white-house-omb-uniform-guidance/>.

⁵ <https://whatworkscities.bloomberg.org/assessment/>.

References

- Ammons, D. N. 2022. “Performance Management Research for State and Local Government: Where It Should Go Next - David N. Ammons, 2022.” *State and Local Government Review* 54: 195–201. <https://doi.org/10.1177/0160323X221084751>.
- Argyris, C., and D. A. Schön. 1978. *Organizational Learning: A Theory of Action Perspective*. Addison-Wesley.
- Bagley, N. 2019. “The Procedure Fetish.” *Michigan Law Review* 118.3: 345. <https://doi.org/10.36644/mlr.118.3.procedure>.
- Bevan, G., and C. Hood. 2006. “What’s Measured Is What Matters: Targets and Gaming in the English Public Health Care System.” *Public Administration* 84, no. 3: 517–538. <https://doi.org/10.1111/j.1467-9299.2006.00600.x>.
- Bilmes, L. J., J. Donahue, S. Goldsmith, et al. 2025. *Analyzing DOGE Actions One Month Into Trump's Second Administration*. Harvard Kennedy School. <https://www.hks.harvard.edu/faculty-research/policy-topics/democracy-governance/analyzing-doge-actions-one-month-trumps-second>.
- Bogner, A., B. Littig, and W. Menz. 2009. *Interviewing Experts*. Palgrave Macmillan UK. <https://doi.org/10.1057/9780230244276>.
- Bruggeman, J., and C. Eilbert. 2025. “What Previous Government Reform Efforts Tell Us About DOGE.” *TIME*. <https://time.com/7260722/attempts-at-government-reform-highlight-challenges-for-doge/>.
- Clinton, W. J. 1993. “Audio Recording of President Clinton's Remarks to the North Valley Job Training Partnership in Sunnyvale, California 9/10/1993.”
- Creswell, J. W., and V. L. Plano Clark. 2018. *Designing and Conducting Mixed Methods Research*. 3rd ed. Sage.
- Cullen, J. B., and R. Reback. 2006. “Tinkering Toward Accolades: School Gaming Under a Performance Accountability System.” In *Advances in Applied Microeconomics*, vol. 14, 1–34. Emerald (MCB UP). [https://doi.org/10.1016/S0278-0984\(06\)14001-8](https://doi.org/10.1016/S0278-0984(06)14001-8).
- Denzin, N. K. 1989. *The Research Act: A Theoretical Introduction to Sociological Methods*. Prentice Hall.
- Dixit, A. 2002. “Incentives and Organizations in the Public Sector: An Interpretative Review.” *Journal of Human Resources* 37, no. 4: 696. <https://doi.org/10.2307/3069614>.
- Feldman, M. S., and J. G. March. 1981. “Information in Organizations as Signal and Symbol.” *Administrative Science Quarterly* 26, no. 2: 171. <https://doi.org/10.2307/2392467>.
- Figlio, D. N., and L. S. Getzler. 2006. “Accountability, Ability and Disability: Gaming the System?” *Advances in Applied Microeconomics* 14: 35–49. https://books.google.com/books?hl=en&lr=&id=F_4ktk3-_noC&oi=fnd&pg=PA35&dq=Accountability,+Ability+and+Disability:+Gaming+the+System%3F&ots=-sN4s-BKjn&sig=G3LMMsp3ppr9EIQK2dwsfbF48JM.
- Frederickson, D. G., and H. G. Frederickson, eds. 2006. *Measuring the Performance of the Hollow State*. Public Management and Change Series. Georgetown University Press.
- Fudge, M. K. 2013. *Performance Budgeting Within State and Local Governments: The Disconnect Between Public Manager Efforts and Legislative Action*. American Society for Public Administration. PA Times. <https://patimes.org/performance-budgeting-state-local-governments-disconnect-public-manager-efforts-legislative-action/>.
- Gilmour, J. B. 2005. “Assessing Performance Budgeting at OMB: The Influence of Politics, Performance, and Program Size.” *Journal of Public Administration Research and Theory* 16, no. 2: 169–186. <https://doi.org/10.1093/jopart/muj002>.
- Gilmour, J. B., and D. E. Lewis. 2006. “Does Performance Budgeting Work? An Examination of the Office of Management and Budget's PART Scores.” *Public Administration Review* 66, no. 5: 742–752. <https://doi.org/10.1111/j.1540-6210.2006.00639.x/full>.
- Groszyk, W. 1995. *Implementation of the Government Performance and Results Act of 1993*. Office of Management and Budget. <https://govinfo.library.unt.edu/npr/library/omb/gpra.html>.
- Hawickhorst, K. 2025. “To Defeat the Bureaucracy, Embrace Bureaucracy.” *The Foundation for American Innovation*. <https://www.thefai.org/posts/to-defeat-the-bureaucracy-embrace-bureaucracy>.
- Heckman, J., C. Heinrich, and J. Smith. 1997. “Assessing the Performance of Performance Standards in Public Bureaucracies.” *American Economic Review* 87, no. 2: 389–395. <http://www.jstor.org/stable/2950951>.
- Heckman, J. J., J. A. Smith, and C. Taber. 1996. *What Do Bureaucrats Do? The Effects of Performance Standards and Bureaucratic Preferences on Acceptance into the JTPA Program*. Working Paper 5535. National Bureau of Economic Research. <https://doi.org/10.3386/w5535>.
- Heinrich, C. J. 2002. “Outcomes-Based Performance Management in the Public Sector: Implications for Government Accountability and Effectiveness.” *Public Administration Review* 62, no. 6: 712–725. <https://doi.org/10.1111/1540-6210.00253>.
- Ho, A. T.-K. 2011. “PBB in American Local Governments: It's More Than a Management Tool.” *Public Administration Review* 71, no. 3: 391–401. <https://doi.org/10.1111/j.1540-6210.2011.02359.x>.
- Ho, D. E., and S. Sherman. 2017. “Managing Street-Level Arbitrariness: The Evidence Base for Public Sector Quality Improvement.” *Annual Review of Law and Social Science* 13, no. 1: 251–272. <https://doi.org/10.1146/annurev-lawsocsci-110316-113608>.
- Holmstrom, B., and P. Milgrom. 1991. “Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design.” *Journal of Law, Economics, and Organization* 7, no. special_issue: 24–52. https://doi.org/10.1093/jleo/7.special_issue.24.
- Holzer, M., and K. Yang. 2004. “Performance Measurement and Improvement: An Assessment of the State of the Art.” *International*

- Review of Administrative Sciences* 70, no. 1: 15–31. <https://doi.org/10.1177/0020852304041228>.
- Hood, C., and R. Dixon. 2015. "Performance Data Breaks: Breaking the Mould and Burying the Evidence." In *A Government That Worked Better and Cost Less?: Evaluating Three Decades of Reform and Change in UK Central Government*, edited by C. Hood and R. Dixon, 44–64. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199687022.003.0003>.
- Hood, C., and B. Piotrowska. 2021. "Goodhart's Law and the Gaming of UK Public Spending Numbers." *Public Performance & Management Review* 44, no. 2: 250–271. <https://doi.org/10.1080/15309576.2020.1749092>.
- Hubbard, K. 2025. "How Clinton's 'Reinventing Government' Compares to DOGE's Approach: 'We Cut Fat and They Cut Muscle'." CBS News. February 20. <https://www.cbsnews.com/news/clinton-reinventing-government-doges-comparison/>.
- Jensen, M. C., and W. H. Meckling. 1976. "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure." *Journal of Financial Economics* 3, no. 4: 305–360. [https://doi.org/10.1016/0304-405X\(76\)90026-X](https://doi.org/10.1016/0304-405X(76)90026-X).
- Jick, T. D. 1979. "Mixing Qualitative and Quantitative Methods: Triangulation in Action." *Administrative Science Quarterly* 24, no. 4: 602–611. <https://doi.org/10.2307/2392366>.
- Kelley, J. G., and B. A. Simmons. 2020. "Introduction: The Power of Global Performance Indicators." In *The Power of Global Performance Indicators*, edited by J. G. Kelley and B. A. Simmons, 1st ed., 1–22. Cambridge University Press. <https://doi.org/10.1017/9781108763493.002>.
- Kerr, S. 1975. "On the Folly of Rewarding A, While Hoping for B." *Academy of Management Journal* 18, no. 4: 769–783. <http://amj.aom.org/content/18/4/769.short>.
- Kettl, D. F. 2023. "'Long, Long Overdue': An Oral History of the Government Performance and Results Act." *Government Executive*. <https://www.govexec.com/management/2023/08/long-long-overdue-oral-history-government-performance-and-results-act/389469/>.
- Kim, T. K. 2025. "Impact of Low-Performance Signals on Employee Fraud in Public Organizations: Evidence From a Pay-For-Performance Context." *Public Administration Review* 85, no. 3: 681–697. <https://doi.org/10.1111/puar.13873>.
- Klein, E., and D. Thompson. 2025. *Abundance*. Avid Reader Press.
- Kraatz, C. 2007. "Amy Chan Is Stepping Down as Sunnyvale City Manager." *The Mercury News*, December 21. <https://www.mercurynews.com/2007/12/21/amy-chan-is-stepping-down-as-sunnyvale-city-manager/>.
- Krippendorff, K. 2019. *Content Analysis: An Introduction to Its Methodology*. 4th ed. SAGE.
- Levitt, B., and J. G. March. 1988. "Organizational Learning." *Annual Review of Sociology* 1: 319–340. <http://www.jstor.org/stable/2083321>.
- Lindsey, B. 2021. *State Capacity: What Is It, How We Lost It, and How to Get It Back*. Niskanen Center.
- Lu, Y., Z. Mohr, and A. T.-K. Ho. 2015. "Taking Stock: Assessing and Improving Performance Budgeting Theory and Practice." *Public Performance & Management Review* 38, no. 3: 426–458. <https://doi.org/10.1080/15309576.2015.1006470>.
- Lu, Y., and K. Willoughby. 2012. "Performance Budgeting in the States: An Assessment."
- Lu, Y., K. Willoughby, and S. Arnett. 2011. "Performance Budgeting in the American States: What's Law Got to Do With It?" *State and Local Government Review* 43, no. 2: 79–94. <https://doi.org/10.1177/0160323X11407523>.
- Mashaw, J. L. 1995. "Reinventing Government and Regulatory Reform: Studies in the Neglect and Abuse of Administrative Law Conference for the Second Century of the University of Pittsburgh School of Law - Panel: The Structure of Government Accountability." *University of Pittsburgh Law Review* 57, no. 2: 405–442. <https://heinonline.org/HOL/P?h=hein.journals/upitt57&i=415>.
- Melkers, J., and K. Willoughby. 2005. "Models of Performance-Measurement Use in Local Governments: Understanding Budgeting, Communication, and Lasting Effects." *Public Administration Review* 65, no. 2: 180–190. <https://doi.org/10.1111/j.1540-6210.2005.00443.x>.
- Menchaca, G. 2025. "From Gore to DOGE: The Bipartisan History of Failed Workforce Reform." *Niskanen Center*. <https://www.niskanencenter.org/from-gore-to-doge-the-bipartisan-history-of-failed-workforce-reform/>.
- Meng, X., and C. Li. 2025. "Does Public Sector Performance Information Impact Stakeholders? Evidence From a Meta-Analysis." *Public Administration Review* 85, no. 4: 1086–1097. <https://doi.org/10.1111/puar.13924>.
- Meyer, J. W., and B. Rowan. 1977. "Institutionalized Organizations: Formal Structure as Myth and Ceremony." *American Journal of Sociology* 83, no. 2: 340–363. <https://doi.org/10.1086/226550>.
- Modell, S. 2022. "New Developments in Institutional Research on Performance Measurement and Management in the Public Sector." *Journal of Public Budgeting, Accounting & Financial Management* 34, no. 3: 353–369. <https://doi.org/10.1108/JPBAFM-04-2021-0070>.
- Moe, T. M. 1984. "The New Economics of Organization." *American Journal of Political Science* 28, no. 4: 739. <https://doi.org/10.2307/2110997>.
- Moynihan, D. P. 2008. *The Dynamics of Performance Management: Constructing Information and Reform*. Public Management and Change Series. Georgetown University Press.
- Moynihan, D. P., and S. Lavertu. 2012. "Does Involvement in Performance Management Routines Encourage Performance Information Use? Evaluating GPRA and PART." *Public Administration Review* 72, no. 4: 592–602. <https://doi.org/10.1111/j.1540-6210.2011.02539.x/full>.
- Nix, N. 2025. "Musk Says DOGE's Work Recalls Clinton-Era Cuts. Al Gore Disagrees." *The Washington Post*, April 26. <https://www.washingtonpost.com/us-policy/2025/04/26/doge-musk-clinton-cuts/>.
- Office of Management and Budget. 2003. *Budget of the U.S. Government, Fiscal Year 2003: Analytical Perspectives*. Government Printing Office. <https://www.govinfo.gov/content/pkg/BUDGET-2003-PER/pdf/BUDGET-2003-PER.pdf>.
- Office of Management and Budget. 2004. "Rating the Performance of Federal Programs." <https://web.archive.org/web/20080718205419/http://www.gpoaccess.gov/usbudget/fy04/pdf/budget/performance.pdf>.
- Osborne, D., and T. Gaebler. 1992. *Reinventing Government: How the Entrepreneurial Spirit Is Transforming the Public Sector*. Addison-Wesley.
- Pahlka, J. 2023. *Recoding America: Why Government Is Failing in the Digital Age and How We Can Do Better*. Henry Holt and Company.
- Parker, C., S. Scott, and A. Geddes. 2020. "Snowball Sampling." In *SAGE Research Methods Foundations*. SAGE Publications Ltd. <https://doi.org/10.4135/9781526421036831710>.
- Poister, T. H., and G. Streib. 1999. "Performance Measurement in Municipal Government: Assessing the State of the Practice." *Public Administration Review* 59, no. 4: 325. <https://doi.org/10.2307/3110115>.
- Pollitt, C. 2011. "Performance Blight and the Tyranny of Light?" In *Accountable Governance: Problems and Promises*, edited by H. G. Frederickson and M. J. Dubnick. M.E. Sharpe.
- Radin, B. A. 2000. "The Government Performance and Results Act and the Tradition of Federal Management Reform: Square Pegs in Round

Holes?" *Journal of Public Administration Research and Theory* 10, no. 1: 111–135. <https://doi.org/10.1093/oxfordjournals.jpart.a024260>.

Radin, B. A. 2006. *Challenging the Performance Movement: Accountability, Complexity, and Democratic Values*. Georgetown University Press.

Rep. Cuellar, Henry [D-TX-28]. 2010. "GPRA Modernization Act of 2010." <https://www.congress.gov/bill/111th-congress/house-bill/2142>.

Report of the Committee on Government Affairs. 1993. "Senate Committee on Government Affairs GPRA Report." <https://obamawhitehouse.archives.gov/omb/mgmt-gpra/gprptm>.

Richter, P. 1993. "Clinton, Gore Hail Sunnyvale's City Efficiency - Los Angeles Times." *Los Angeles Times*, September 11. <https://www.latimes.com/archives/la-xpm-1993-09-11-mn-34070-story.html>.

Rivenbark, W. C., and V. Vignieri. 2025. "The Shift From Performance Measurement to Performance Management in Local Government: Do Performance Measurement Systems Support the Drivers of Performance Data Use?" *International Journal of Public Sector Management* 38, no. 3: 368–385. <https://doi.org/10.1108/IJPSM-12-2023-0356>.

Saldaña, J. 2016. *The Coding Manual for Qualitative Researchers*. 3rd ed. SAGE.

Sanger, M. B. 2008. "From Measurement to Management: Breaking Through the Barriers to State and Local Performance." *Public Administration Review* 68, no. s1: S70–S85. <https://doi.org/10.1111/j.1540-6210.2008.00980.x>.

Scott, W. R. 2013. *Institutions and Organizations: Ideas, Interests, and Identities*. SAGE Publications.

Sen. Roth Jr., William V. [R-DE]. 1993. "S.20 - 103rd Congress (1993–1994): Government Performance and Results Act of 1993." <https://www.congress.gov/bill/103rd-congress/senate-bill/20>.

Small, M. L., and J. M. C. Calarco. 2022. *Qualitative Literacy: A Guide to Evaluating Ethnographic and Interview Research*. 1st ed. University of California Press. <https://doi.org/10.2307/j.ctv2vr9c4x>.

Surani, F., L. A. Gailmard, A. Casasola, V. Magesh, E. J. Robitschek, and D. E. Ho. 2025. "What Is the Law? A System for Statutory Research (STARA) With Large Language Models." In *20th International Conference on Artificial Intelligence and Law*. ACM Digital Library. <https://doi.org/10.1145/3769126.3769219>.

Thurm, S. 1993. "President in Sunnyvale: 'We Want to be Like You'." *San Jose Mercury News*, September 11.

Tirole, J. 1986. "Hierarchies and Bureaucracies: On the Role of Collusion in Organizations." *Journal of Law, Economics, and Organization* 2, no. 2: 181–214. <https://doi.org/10.1093/oxfordjournals.jleo.a036907>.

Tonks, H. M. J. 2025. "Reform Won't Stick Without Real Goals." *Hypertext*, July 25. <https://hypertext.niskanencenter.org/p/reform-wont-stick-without-real-goals>.

Townsend, R. M. 1979. "Optimal Contracts and Competitive Markets With Costly State Verification." *Journal of Economic Theory* 21, no. 2: 265–293. [https://doi.org/10.1016/0022-0531\(79\)90031-0](https://doi.org/10.1016/0022-0531(79)90031-0).

U.S. Government Accountability Office. 1997. *GPRA: Managerial Accountability and Flexibility Pilot Did Not Work as Intended: GGD-97-36*. BiblioGov.

Van Sant, W. 2008. "Hazard of Job: Political Conflict." *Tampa Bay Times*, August 8. <https://www.tampabay.com/archive/2008/08/08/hazard-of-job-political-conflict/>.

Weinstein, P. 2025. "Seven Lessons for Elon Musk And DOGE." *Forbes*, February 3. <https://www.forbes.com/sites/paulweinstein/2025/02/03/seven-lessons-for-elon-musk-and-doge/>.

Wilkinson, K., and W. Harper. 1999. "Metroactive News & Issues | Sunnyvale Staffing Problems." <https://www.metro siliconvalley.com/papers/metro/08.19.99/sunnyvale-9933.html>.

Zak, P. 2013. "Measurement Myopia." <https://drucker.institute/thedx/measurement-myopia/>.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Data S1:** Supplementary Appendix.