ACQUISITION RESEARCH PROGRAM SPONSORED REPORT SERIES

Establishing Standardized Metrics for the Acquisition of Services Pathway

December 2025

1st Lt. Daniel J. Cary, USAF
Capt Michael E. Alexander, USAF
Capt Ashley M. Presley, USAF

Thesis Advisors: Lt. Col. Jamie Porchia, Assistant Professor
Dr. Robert F. Mortlock, Professor

Department of Defense Management

Naval Postgraduate School

Approved for public release; distribution is unlimited.

Prepared for the Naval Postgraduate School, Monterey, CA 93943

ACQUISITION RESEARCH PROGRAM
DEPARTMENT OF DEFENSE MANAGEMENT
NAVAL POSTGRADUATE SCHOOL

ACQUISITION RESEARCH PROGRAM
DEPARTMENT OF DEFENSE MANAGEMENT
NAVAL POSTGRADUATE SCHOOL

ABSTRACT

The Department of Defense (DoD) relies heavily on contracted services but lacks a standardized, outcome-based framework to assess program health. Current practices emphasize process compliance and obligation rates rather than mission impact, limiting oversight and strategic visibility. This research develops a standardized performance measurement framework for the Acquisition of Services Pathway to improve accountability and decision-making. Using a qualitative design, policy documents were analyzed and 24 interviews conducted with senior DoD and industry stakeholders to identify and validate high-value performance indicators. Stakeholders reconceptualize "service health" as measurable contributions to operational readiness rather than administrative activity. The study validates a "Core-Plus-Tail" governance model and identifies seven foundational metrics covering cost, schedule, quality, customer satisfaction, and compliance, enabling enterprise reporting while allowing category-specific tailoring. The Enterprise Performance Visibility (EPV) model and a Flexible Balanced Scorecard (FBS) are introduced to operationalize this model. Recommendations include updating DoDI 5000.74 to mandate the core set, automating data collection, and piloting a three-tier rollup structure to shift the acquisition culture from compliance to outcomes.

THIS PAGE INTENTIONALLY LEFT BLANK

ABOUT THE AUTHORS

Capt Michael Alexander is a Contracting Officer in the United States Air Force. He joined the Oklahoma Air National Guard in 2014. While serving as a traditional Guardsman, he pursued a Bachelor of Arts in Accounting at the University of Oklahoma. During his time at the university, he had the opportunity to join the Air Force ROTC program, which led to his commission in 2019 and entry into Active Duty. Captain Alexander began his active duty career at Seymour Johnson AFB, NC, and was later PCS'd to Hanscom AFB, MA, before arriving at the Navy Postgraduate School. In his free time, he enjoys traveling, volunteering at church, playing golf, and spending time with his family. Captain Alexander and his family are looking forward to his follow-on assignment to Robins AFB, GA.

1st Lt Daniel "Dan" Cary is a Contracting Officer in the United States Air Force. He was commissioned in 2022 through Officer Training School after being selected for the Senior Leader Enlisted Commissioning Program–Officer and holds a Bachelor of Science in Business Administration from Southern New Hampshire University. After enlisting in 2007, he first served as an A-10 Dedicated Crew Chief at Davis-Monthan AFB, Arizona, before cross-training into contracting and serving at Grand Forks AFB, North Dakota; Ramstein AB, Germany; and Wright-Patterson AFB, Ohio. His previous assignments include roles as Contracting Specialist, Contracting Officer, Team Lead, Manager of Contingency Operations, and Section Chief and Deputy Division Chief for the USAF F-16 portfolio, where he led a team managing a multibillion-dollar modernization and service life extension program. He has earned multiple Meritorious Service Medals and other awards for leadership and contracting excellence. In his free time, he enjoys traveling, reading, woodworking, volunteering, cooking, and spending time with his family.

Capt Ashley Presley is a United States Air Force Contracting Officer. She was commissioned in 2020 through Reserve Officer Training Corps, Detachment 088 after being selected for the Airmen Scholarship and Commissioning Program and holds a Bachelor of Science in Aviation Business Administration from Embry Riddle University.

She first enlisted in 2013 and served as a Security Forces member in the Contingency Response Wing at Travis AFB, California and returned to serve the contracting squadron before attending Naval Postgraduate School. In her free time, she enjoys volunteering, and spending time with her family outdoor and indoor.

ACKNOWLEDGMENTS

First and foremost, we give thanks to God, recognizing that we are nothing without Him and that this work was only possible through His guidance, strength, and provision.

We would like to thank our thesis advisor, Jamie Porchia, for her guidance and steady encouragement through this research. Her thoughtful feedback and willingness to challenge our thinking were instrumental in strengthening both the analysis and the practical relevance of this work.

Finally, we are grateful to our families for their patience, understanding, and unwavering support during the many hours devoted to this effort.

THIS PAGE INTENTIONALLY LEFT BLANK

ACQUISITION RESEARCH PROGRAM
SPONSORED REPORT SERIES

Establishing Standardized Metrics for the Acquisition of Services Pathway

December 2025

1ˢᵗ Lt. Daniel J. Cary, USAF
Capt Michael E. Alexander, USAF
Capt Ashley M. Presley, USAF

Thesis Advisors: Lt. Col. Jamie Porchia, Assistant Professor
Dr. Robert F. Mortlock, Professor

Department of Defense Management

Naval Postgraduate School

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

AAF	Adaptive Acquisition Framework
ADVANA	Advancing Analytics
AERs	Annual Execution Reviews
AoS	Acquisition of Services
AVDF	Acquisition Visibility Data Framework
BSC	Balanced Scorecard
DAF	Department of the Air Force
DAFI	Department of the Air Force Instruction
DBB	Defense Business Board
DFARS	Defense Federal Acquisition Regulation Supplement
DIB	Defense Industrial Base
DoD	Department of Defense
DoDI	Department of Defense Instruction
DOE	Department of Energy
DSB	Defense Science Board
EMAs	Expectation Management Agreements
EPV	Enterprise Performance Visibility
EVM	Earned Value Management
FAR	Federal Acquisition Regulation
FBS	Flexible Balanced Scorecard
FPDS	Federal Procurement Data System
FY	Fiscal Year
GAO	Government Accountability Office
IDA	Institute for Defense Analyses
IOM	Institute of Medicine
IPE	Integrated Performance Ecosystem
ISG	Information Services Group
ISO/IEC	International Organization for Standardization & International Electrotechnical Commission
IT	Information Technology
ITIL	Information Technology Infrastructure Library

KPI	Key Performance Indicators
NDIS	National Defense Industrial Strategy
NHS	National Health Service
NPS	Naval Postgraduate School
OBC	Outcome-Based Contracts
OECD	Organization for Economic Co-operation and Development
OLA	operation-level agreement
OUSD(A&S)	Office of Under Secretary of Defense for Acquisition and Sustainment
OUSD(AT&L)	Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics
PALT	procurement administrative lead time
PBSA	Performance-Based Services Acquisition
PEO	Program Executive Office
PWS	Performance Work Statement
QASP	Quality Assurance Surveillance Plan
RAND	Research and Development Corporation
SAM	System for Award Management
SLA	service-level agreements
SOO	Statement of Objectives
SRRB	Services Requirements Review Board
TIPS-EEP	Three Integrated Pillars of Success-Efficiency Effectiveness and Process
USD(A&S)	Under Secretary of Defense for Acquisition and Sustainment

I.   INTRODUCTION

This chapter introduces the study's motivation, scope, key research questions, and report structure. It highlights the importance of improving performance measurement in Department of Defense (DoD) service acquisitions and sets the foundation for the analysis and framework development in later chapters.

A.   CONTEXT OF THE ADAPTIVE ACQUISITION FRAMEWORK AND THE SERVICES PATHWAY

The DoD relies on a diverse set of acquisition approaches to deliver capabilities effectively. Recognizing that a uniform model can hinder agility, the DoD implemented the Adaptive Acquisition Framework (AAF), governed by DoD Instruction (DoDI) 5000.02, *Operation of the Adaptive Acquisition Framework* (Office of the Under Secretary of Defense for Acquisition & Sustainment [OUSD(A&S)], 2020). The AAF allows program managers to choose from six tailored pathways based on the nature and urgency of the acquisition. Chapter II provides a detailed explanation of these pathways, with particular focus on the Acquisition of Services (AoS) Pathway.

This study is sponsored by the OUSD(A&S), which oversees the DoD's AAF. The sponsorship emerged from the 120-day AAF Pathway Review directed by the Under Secretary of Defense for Acquisition and Sustainment (USD[A&S], 2024a) to evaluate training, policies, and performance measurement across the acquisition pathways (OUSD[A&S], 2024b). During that review, the Acquisition of Services (AoS) team identified the absence of outcome-based performance measures as a major barrier to assessing program health and mission contribution. The OUSD(A&S) (2024b) initiated this research to develop standardized, outcome-oriented metrics that strengthen accountability, enable cross-program comparison, and inform policy refinement within the AoS Pathway.

The AoS Pathway provides the governance structure for the procurement and management of contracted services that span professional, technical, logistics, and sustainment support activities. These services represent a substantial portion of the DoD's annual contract obligations, accounting for nearly half of total spending in recent years

(U.S Government Accountability Office [GAO], 2023b). Consistent with DoDI 5000.74, *Defense Acquisition of Services*, the pathway is designed to ensure service acquisitions are planned and executed to achieve measurable results and mission outcomes (OUSD[A&S], 2021). This research supports the DoD's effort to strengthen performance accountability and establish standardized metrics that connect service acquisitions to readiness and mission impact (P. Smith, personal communication, December 13, 2024).

B. PRACTICAL PROBLEM STATEMENT

The core problem guiding this research is the absence of a consistent, department-wide framework for assessing the health and effectiveness of service acquisitions within the AoS Pathway. Sponsor feedback during early consultation with the OUSD(A&S) revealed inconsistency and uncertainty across the acquisition workforce regarding which performance metrics are most meaningful, how they should be applied, and how they influence accountability across service categories. This gap limits the DoD's ability to evaluate mission contribution and was identified by the OUSD(A&S) as a critical area for reform during the 120-Day Pathway Review.

The DoD relies heavily on contracted services to support its mission, with fiscal year (FY) 2019 service contracts comprising nearly 50% of all DoD contract obligations (GAO, 2021a). Over the period from 2017–2020, annual service contract obligations ranged between approximately $184 billion and $226 billion, consistently representing about half of total contract spending (GAO, 2023b). Although services comprise a substantial portion of the DoD acquisition portfolio, the department primarily uses process-based metrics to evaluate service contract performance. Metrics such as procurement administrative lead time (PALT), competition rates, and contract award timelines offer useful insight into efficiency but provide little information on service outcomes or mission effectiveness (GAO, 2021b; McKernan et al., 2022). This overreliance on procedural data obscures the actual value and performance of services rendered, leading to a phenomenon akin to "metric mania," which Casey et al. (2008, p. 92) define as measurement eclipsing achievement.

Without a standardized framework for outcome measurement, oversight bodies and acquisition leaders cannot determine whether contracted services are delivering

intended results or contributing optimally to mission readiness. According to the GAO (2021b), outcome-oriented measures emphasize results like cost savings, timeliness, quality, and customer satisfaction rather than process compliance. DoDI 5000.74 calls for the use of performance metrics but offers general guidance without establishing uniform standards (OUSD[A&S], 2021). Consequently, the current system is fragmented and lacks the tools necessary to support cross-program comparisons, trend analysis, and strategic oversight (OUSD[A&S], 2021; OUSD[A&S], 2024b). This shortfall presents broader consequences, such as weakened alignment with strategic goals and reduced capacity for continuous learning (GAO, 2013). External reviews, including GAO and Research and Development Corporation (RAND) studies, consistently emphasize the need for outcome-based metrics to strengthen the AoS Pathway's impact on readiness and mission outcomes, echoing the 120-Day Pathway Review tasking (OUSD[A&S], 2024a; GAO, 2021b; McKernan et al., 2022).

C. KNOWLEDGE GAP

Although oversight bodies and researchers have repeatedly identified the lack of outcome-based measures, existing studies have not produced a unified, empirically grounded framework that links service-acquisition performance to mission readiness. GAO (2013, 2021b) reports highlight incomplete or inconsistent data, while RAND analyses emphasize the difficulty of connecting acquisition activities to measurable outcomes (McKernan et al., 2022). Prior research focuses on describing deficiencies rather than designing a standardized model that the DoD could institutionalize across service categories. This study addresses that gap through the development and evaluation of a framework of outcome-based performance metrics tailored to the AoS Pathway, integrating insights from policy, literature, and practitioner perspectives.

D. RESEARCH QUESTIONS

This study centers on the need for standardized, outcome-based performance metrics within the DoD's AoS Pathway. The primary question guiding this research is:

> 1. What outcome-based performance measures should be used to assess the health of services?

The secondary questions support this core inquiry by exploring how these metrics can be developed and applied effectively:

2. How can standardized outcome focused performance metrics improve assessment, oversight, and strategic decision-making across DoD service programs?
3. What performance indicators used in service contracts can be adapted for DoD use?
4. What challenges exist for implementing standardized metrics, and how can these be addressed across different service categories?

Together, these questions are meant to inform a comprehensive framework for evaluating and improving service acquisition outcomes in the DoD.

E. OBJECTIVES

The overarching goal of this research is to improve the effectiveness, transparency, and accountability of service acquisitions within the DoD by developing a standardized, outcome-based performance measurement framework tailored to the AoS Pathway. Four supporting objectives guide this effort:

- Catalog and analyze outcome-based performance metrics currently used across DoD service acquisitions, based on literature, policy guidance, existing data systems, and stakeholder interviews.
- Compare and contrast DoD service acquisition metrics with those used in mature private-sector service organizations.
- Develop a performance measurement framework based on logic models and literature, categorizing metrics by mission impact, service quality, and cost-effectiveness.
- Evaluate the feasibility of implementing the proposed framework across different DoD organizations and service categories.

F. SUMMARY AND OVERVIEW OF SUBSEQUENT CHAPTERS

This chapter introduced the motivation, scope, and objectives of the study. It presented a description of how the DoD's AAF and AoS Pathway govern the acquisition of contracted services and identified the absence of standardized, outcome-based performance metrics as a central problem affecting accountability, comparability, and mission alignment. The discussion established the foundation for this research by framing the issue as both a policy and management challenge that limits the DoD's ability to evaluate the effectiveness of service acquisitions.

Chapter II provides essential background on the AAF and the AoS Pathway. It presents the policy environment, governance structure, and oversight mechanisms that shape service-acquisition practices, highlighting where existing processes and data systems constrain outcome-oriented measurement.

Chapter III presents a review of the relevant academic, policy, and practitioner literature to define what is known and what remains uncertain about performance measurement in defense services. It integrates theoretical models, oversight analyses, and comparative private-sector practices to identify enduring gaps that justify the development of a standardized, outcome-based framework.

Chapter IV presents the study's methodology, data, analysis, and findings. It outlines the qualitative research design, including interviews and document/Key Performance Indicator (KPI) analysis, and details how the integrated evidence informed the development of a tiered performance measurement framework that connects contract-level metrics to enterprise outcomes.

Chapter V concludes the study with practical recommendations and implications for policy, workforce development, and data governance. It translates the proposed framework into actionable steps for improving performance measurement within the AoS Pathway and outlines future research opportunities to advance outcome-based accountability across the DoD.

THIS PAGE INTENTIONALLY LEFT BLANK

II.   BACKGROUND

As part of the AAF's 120-Day Review, the USD (A&S; 2024a) tasked pathway leads with identifying appropriate performance metrics for each acquisition pathway. However, for the AoS Pathway, most available metrics remain process-oriented and do not inform whether contracted outcomes meet mission needs (GAO, 2013; 2021b). Independent assessments by both the GAO (2021b) and the RAND Corporation (McKernan et al., 2022) have emphasized that improving service acquisition outcomes requires not just oversight but also meaningful performance management grounded in outcome-based data.

Private-sector organizations commonly rely on a structured set of KPIs, such as cost control, timeliness, quality of service delivery, and customer satisfaction, to strengthen accountability and support informed decision-making (Icertis, 2025; Singh, 2024). These metrics are used not only to evaluate contract outcomes but also to identify areas for improvement and ensure alignment with strategic objectives. Although certain military departments have begun piloting similar approaches, these efforts remain uneven and are not yet institutionalized through formal policy or adopted consistently across the DoD (McKernan et al., 2022).

Without reliable, outcome-based metrics, program managers and contracting officials lack the tools to assess service performance with consistency or draw meaningful comparisons across contracts. The absence of standardized data structures further limits the DoD's ability to synthesize results across programs, making it difficult to institutionalize best practices or ensure accountability at scale (GAO, 2023b; Anton et al., 2022).

The remainder of this chapter provides an exploration of the foundational elements of the AoS Pathway and the broader acquisition framework. It outlines key policy documents, system structure, stakeholder roles, and oversight challenges to establish the basis for the proposed research in subsequent chapters.

A. ADAPTIVE ACQUISITION FRAMEWORK

The AAF was implemented to provide a flexible, responsive acquisition system tailored to the unique needs of different capability types. It represents a departure from previous, more rigid frameworks by allowing acquisition professionals to select from six tailored pathways: Major Capability Acquisition, Middle Tier of Acquisition, Urgent Capability Acquisition, Software Acquisition, Business Systems, and AoS (OUSD[A&S], 2020; see Figure 1)
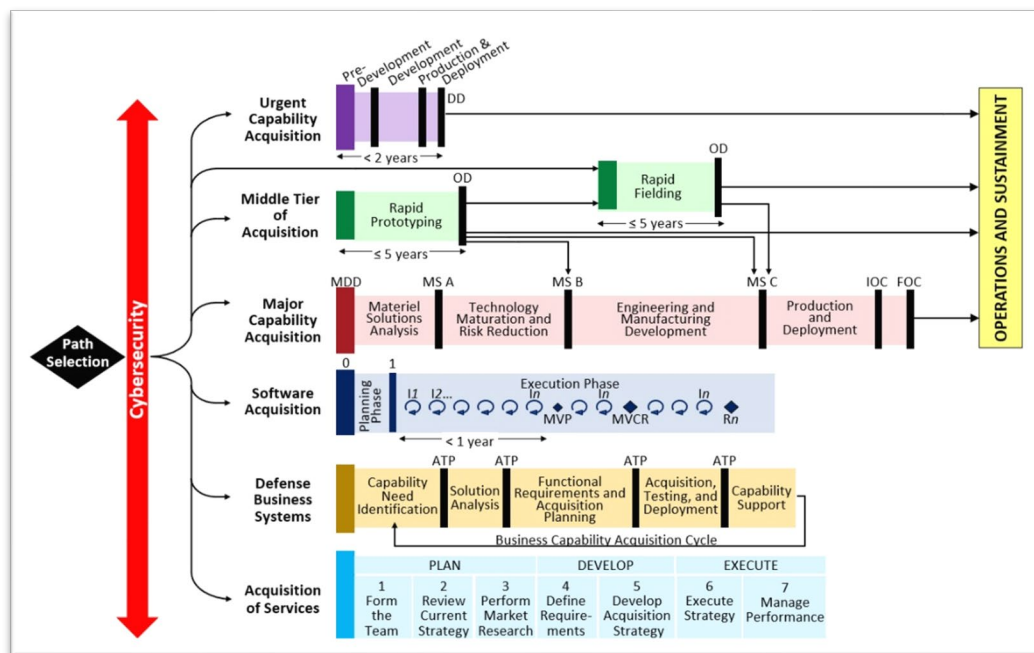
Figure 1.   Adaptive Acquisition Framework. Source: OUSD (A&S; 2020).

DoDI 5000.74 outlines the AoS pathway as responsible for managing a diverse portfolio of contracted services, such as base support, professional services, information technology, and logistics. Given the scale of service acquisition within the DoD, the AoS pathway holds significant potential for performance improvement and cost efficiency. It lacks consistent entry documentation, which complicates tracking usage and assessing program performance. RAND reported that analysis of AoS Pathway projects "relies solely on labor-intensive data collection to assess who is using the pathway," and found limited post-award performance information outside of PALT (McKernan et al., 2022, p. 7). These structural issues impede the DoD's ability to aggregate outcome-based insights for enterprise-level performance reviews (see Figure 2).
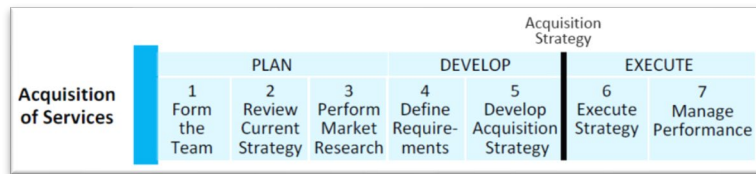
Figure 2.   Acquisition of Services Pathway within the Adaptive Acquisition
Framework. Source: Defense Acquisition University (n.d.).

B.   CHARACTERISTICS OF THE AOS PATHWAY

According to the DoD's Guidebook for the Acquisition of Services, service acquisitions typically follow a structured life cycle that includes requirements definition, acquisition planning, source selection, contract administration, and performance assessment (Office of the Under Secretary of Defense for Acquisition, Technology and Logistics [OUSD(AT&L)], 2012a, pp. 5–7). These activities rely heavily on documentation such as Performance Work Statements (PWS) and Quality Assurance Surveillance Plans (QASPs), with a strong focus on post-award monitoring.

Key players include the requiring activity, contracting officers, Contracting Officer's Representatives (CORs), and various oversight entities. Successful acquisition depends on collaboration among these stakeholders, yet current practices often lack shared performance metrics, leading to fragmentation in oversight and accountability. Some military departments have begun experimenting with outcome-oriented performance-measurement approaches; for example, the Air Force has piloted dashboards to track contract-to-capability time, and the Army initiated outcome-oriented metric pilots in 2020 (GAO, 2021a). Despite these efforts, they remain inconsistent and are not embedded in policy or practice across the enterprise (OUSD[A&S], 2024b).

Services span multiple mission categories, including operational support, logistics, professional advisory, and sustainment, and may be defined by either performance-based or level-of-effort requirements. This inherent diversity complicates the development of uniform metrics applicable across all service types, a challenge consistently noted by acquisition professionals (McFall & La, 2011). Both scholars and policymakers have underscored the need for bucketing or categorizing services to enable meaningful, yet adaptable, metric application. For example, the DoD's Taxonomy for the

Acquisition of Services and Supplies & Equipment formally groups services into categories such as knowledge-based, logistics management, equipment-related, and electronics and communications, providing a foundation for tailoring performance indicators to service characteristics (OUSD[AT&L], 2012b).

While pathways like Major Capability Acquisition employ frameworks such as Earned Value Management and milestone testing to assess progress, the AoS Pathway does not have comparably rigorous tools for tracking effectiveness. This disparity often leads to overreliance on process indicators like PALT, defined by the GAO (2021a) as the average time between solicitation release and contract award. While useful for tracking efficiency, PALT provides limited insight into service quality or mission impact, creating a system in which "measurement eclipses achievement" (Casey et al., 2008, p. 92). This overreliance on process metrics has been criticized in various external reviews (GAO, 2013; McKernan et al., 2022)

C.   POLICY AND REGULATORY CONTEXT

Federal Acquisition Regulation (FAR) Part 37 and Defense Federal Acquisition Regulation Supplement (DFARS) Part 237 establish general principles for performance-based acquisition and surveillance. However, these regulations do not prescribe specific outcome-oriented metrics, leaving their adoption to agency discretion. The *Guidebook for the Acquisition of Services* encourages flexibility and mission focus in defining metrics but lacks enforceable standards or data definitions, contributing to variation across commands and services (OUSD[AT&L], 2012a).

Section 4.5 of DoDI 5000.74 promotes performance management "to the maximum extent practicable" (OUSD[A&S], 2021, p. 22) but does not identify concrete metrics or a unified framework, resulting in inconsistent implementation across the acquisition enterprise. Notably, DoDI 5000.74 is currently under rewrite, with an expected FY 2026 publication, signaling forthcoming policy changes that may impact metric mandates. The Acquisition Visibility Data Framework is a DoD-wide framework designed to standardize data collection and business rules across acquisition categories; however, RAND found that its applicability to services is limited because its data fields

are inconsistently populated and do not capture service outcomes (McKernan et al., 2022, pp. 30–31).

Recent policy memoranda, such as the FY 2024 Acquisition of Services Review Panel Wrap-up, have emphasized the importance of measurable outcomes in all AAF pathways (USD[A&S], 2024b). However, implementation across the AoS Pathway has lagged due to legacy practices and limited enforcement mechanisms.

D.   EXTERNAL CRITIQUES AND RECOMMENDATIONS

Reviews of the DoD's service acquisition practices have repeatedly identified problems with setting performance goals and measuring results. In 2006, the GAO reported that the DoD lacked a tailored approach to improving service acquisition outcomes and instead relied on broad, process-focused metrics. A later GAO (2013) review reached a similar conclusion, noting that the DoD still had not established clear goals or metrics to track progress. These findings showed that the DoD lacked clear and standardized measures to evaluate whether service contracts were achieving intended outcomes and contributing to mission success.

More recent evaluations confirm that these problems remain. The GAO (2021a) found that the DoD's report to Congress on service acquisition management improvements left out key planning elements, including outcome-based measures. The GAO's (2023b) High-Risk List has also continued to flag DoD contract management as a long-standing problem area, citing weaknesses in acquisition data and performance measurement as risks that limit effective oversight. These assessments suggest that despite years of oversight, the DoD has made limited progress in developing effective performance measures for service acquisitions.

The RAND Corporation (McKernan et al., 2022) and the Defense Business Board (DBB; 2022) both have noted that the DoD relies too much on process-driven measures when assessing acquisition performance. RAND found that AAF pathways, including the AoS Pathway, mainly use activity-based indicators that track compliance or administrative actions rather than mission outcomes. The DBB (2022) similarly recommended that the DoD adopt next-generation business health metrics, modeled on

private-sector practices, to better capture efficiency, effectiveness, and mission contribution. The GAO (2021b) supported this perspective, observing that while federal procurement leaders depend heavily on process measures, leading private-sector companies use outcome-focused metrics such as cost savings, timeliness, quality, and customer satisfaction. Together, these studies highlight that the DoD still lacks standardized outcome-based metrics for evaluating service acquisitions.

The difference between process- and outcome-oriented measures is central to understanding this gap. Process measures track compliance and administrative efficiency, such as procurement lead time, competition rates, or small business participation (GAO, 2021b). While useful for monitoring acquisition procedures, these measures give little insight into whether contracted services are achieving intended results. Outcome measures, in contrast, assess whether services support larger organizational goals, such as improving mission readiness, reducing costs, delivering on time, producing higher-quality results, or satisfying end users (McKernan et al., 2022; GAO, 2021b). Without standardized outcome measures, the DoD cannot determine how well its service acquisitions contribute to mission outcomes.

Collectively, these critiques show that the DoD continues to rely on process-focused measures, even after years of oversight attention and expert recommendations. RAND (McKernan et al., 2022), DBB (2022), and GAO (2021b) have each stressed the lack of standardized outcome-based metrics and called for more meaningful approaches to evaluating efficiency, effectiveness, and mission contribution. These observations not only highlight the persistence of the problem but also reinforce the need for further research to address the gap in outcome-based performance measures for DoD service acquisitions.

E.   IMPACTS OF A LACK OF STANDARDIZED METRICS

At the strategic level, the lack of standardized, outcome-based metrics prevents the DoD from forming a complete picture of service acquisition performance across the enterprise. The GAO (2013) first reported that the DoD had not established measurable goals or consistent metrics for service acquisitions, making it difficult to assess progress or determine effectiveness. Subsequent reviews reaffirmed that the DoD continues to lack

reliable data to evaluate results across portfolios (GAO, 2021a; 2025). RAND also found that limited post-award data and inconsistent indicators make it difficult to connect service performance to mission outcomes (McKernan et al., 2022).

At the program level, contracting officials and program managers lack reliable benchmarks for evaluating contract performance. Individual organizations often create local metrics to fill this void, but these efforts vary widely in scope and rigor, which constrains the ability to share lessons learned or apply best practices across commands (GAO, 2013; 2021a; McKernan et al., 2022; OUSD[A&S], 2021). The GAO (2021a) reported that this inconsistency undermines comparability and prevents aggregation of results for higher-level decision-making. As a result, programs operate with limited visibility into relative performance, making it difficult to manage resources efficiently or demonstrate progress within their portfolios.

At the oversight level, incomplete or inconsistent data restrict senior leaders and Congress from holding the DoD accountable for service-acquisition outcomes. The GAO (2013) identified that the absence of standardized metrics limited the DoD's ability to assess performance and demonstrate results to Congress. Later reports confirmed that unreliable data and a lack of consistent performance information continue to hinder transparency and meaningful oversight (GAO, 2021a). The most recent High-Risk Series again noted that the DoD continues to face challenges in validating service-acquisition requirements, forecasting needs, and aggregating data to identify efficiencies, which limits visibility into contract performance and keeps contract management on the High-Risk List (GAO, 2025, pp. 84–89). These conditions weaken accountability mechanisms and delay policy reforms needed to improve service-acquisition outcomes.

At the mission level, the continued use of process-based measures prevents the DoD from linking service acquisition outcomes to mission readiness (GAO, 2021b; OUSD[A&S], 2021). Metrics such as procurement lead time or competition rates capture internal efficiency but provide little insight into whether contracted services contribute to operational effectiveness or readiness outcomes (GAO, 2013; McKernan et al., 2022). The GAO (2021b) and the OUSD (A&S; 2021) indicate that outcome-oriented data remain inconsistent across service programs. RAND analysis revealed that the limited use

of post-award performance information makes it difficult to determine how services affect mission success and value delivered (McKernan et al., 2022). Without standardized, outcome-based metrics, the DoD cannot fully assess how effectively services contribute to readiness objectives or make evidence-based adjustments to improve future performance.

Overall, these challenges highlight the DoD's continued need for standardized, outcome-based performance metrics that provide a clear picture of service-acquisition effectiveness. Without a common framework, strategic visibility, program management, and mission readiness remain fragmented, and decision-makers lack reliable evidence to guide improvement efforts. The following literature review presents an examination of prior research, policy analyses, and oversight findings that further define this gap and inform the development of a standardized, outcome-based metric framework for the AoS Pathway.

III. LITERATURE REVIEW

The purpose of this chapter is to synthesize academic, policy, and practitioner literature that informs how performance is measured and managed within the DoD's AoS Pathway. The review presents an examination of what is known and what remains uncertain about performance measurement in the DoD context to clarify why additional research is warranted. It integrates theoretical, policy, and applied perspectives to show how performance measurement has evolved, where current approaches fall short, and how outcome-based methods could strengthen accountability and decision-making.

The chapter is organized around three primary thematic domains of literature:

- Performance Measurement and KPIs: Foundational theory and public-sector scholarship defining KPIs, distinguishing outputs from outcomes, and addressing the risks of overemphasizing process compliance
- Federal and DoD Policy and Oversight: Statutes, guidance, and oversight studies—particularly from the GAO, RAND, and the DBB that assess the effectiveness of service-acquisition performance management and identify recurring measurement gaps
- Transferable and Comparative Practices: Private-sector and cross-agency approaches such as service-level agreements, outcome-based and relational contracting, and balanced scorecard (BSC) systems that illustrate how outcome-oriented metrics are implemented and what practices may be adapted for DoD application

Each domain contributes to understanding the evolution of performance measurement and the persistent barriers to standardizing outcome-based metrics across DoD service portfolios. These thematic areas also reflect the practical needs identified by the project sponsor, ensuring that the literature review remains grounded in operational relevance while advancing theoretical insight. Findings are synthesized thematically across sections rather than summarized article by article, with emphasis on leading versus lagging measures, process versus outcome orientation, and mechanisms for cascading and standardizing KPIs across the enterprise.

A. PERFORMANCE MEASUREMENT AND KPIS

This section provides a review of the foundations of performance measurement and the use of KPIs in both theory and practice. It begins with an establishment of

theoretical principles that explain how metrics translate strategy into results and continues with an examination of how these concepts are applied within the DoD's service-acquisition environment. These discussions clarify how public organizations measure performance, where current practices diverge from intent, and why outcome-oriented frameworks are needed.

1. Theoretical Foundations

This section includes a survey of foundational theory to show how performance measurement, when properly designed, translates strategy into behavior and results in the public sector. The discussion begins from an outcomes-first stance (Casey et al., 2008; Hatry & Bryant, 1998) and then lays out the architecture of balanced systems that integrate leading and lagging indicators (Kaplan, 2010). Two canonical frameworks, the BSC and the performance prism, demonstrate how objectives, measures, and governance can be aligned across stakeholders (Kaplan & Norton, 1992; Neely et al., 2002). Because measures can mislead if they are unreliable or invalid, indicators must be grounded in basic measurement theory (Messick, 1994 Nunnally & Bernstein, 1994) and in a causal chain from structure-process-outcome (Donabedian, 1988). Public organizations require this discipline because, unlike profit-driven firms, they must connect metrics to public value (Bouckaert & Halligan, 2008; Symes, 1999).

According to Casey et al. (2008), effective performance measurement systems in the public sector focus on outcomes rather than on the volume of metrics. Metrics should inform decision-making and accountability, serving as tools to evaluate progress toward strategic goals rather than becoming ends in themselves. Scholars distinguish between outputs, which capture the direct products of contractor activities (e.g., services delivered, reports completed), and outcomes, which reflect the results or value realized by the buying organization (e.g., mission readiness; Hatry & Bryant, 2006). Public organizations lacking a profit bottom line often struggle to measure real outcomes; as a result, they may emphasize task counts rather than results linked to mission success (Casey et al., 2008).

A sound KPI system balances leading indicators that predict future results with lagging indicators that confirm past performance (Kaplan, 2010). It also avoids overreliance on process measures, such as cycle times or contract actions, which can

obscure whether services actually achieve intended outcomes. The GAO (2021b, p. 15) observed that federal agencies rely heavily on process measures even as leading firms emphasize outcomes like cost savings, timeliness, quality, and user satisfaction. Poorly designed performance measurement systems can lead to "metric mania," an excessive focus on metrics where measurements multiply without strategic connection, diverting attention from achieving mission or strategic results (Casey et al., 2008, p. 92). To counter this, design should begin with mission outcomes and cascade downward.

The BSC provides a foundational framework for linking performance metrics to strategy. Kaplan and Norton (1992) originally proposed four perspectives—financial, customer, internal process, and learning and growth—to avoid single-measure myopia and provide a multi-dimensional view of organizational health. Later adaptations reframed the BSC as a strategy-management system by coupling objectives, measures, targets, and initiatives in a continuous feedback loop of execution and review, often using strategy maps to illustrate cause-and-effect (Kaplan, 2010). Public-sector versions place mission effectiveness at the top outcome and reframe the financial dimension as stewardship or cost-effectiveness. The customer perspective represents end users, citizens, or sponsors, preserving balance while aligning with public value (Kaplan, 2010).

Measurement theory extends beyond what to count into how well one is counting. Sound measures demonstrate reliability (stable results under consistent conditions) and validity (capturing the construct of interest), with construct validity linking indicators to the underlying outcomes the enterprise cares about (Messick, 1994; Nunnally & Bernstein, 1994). In practice, this means defining each KPI's purpose, construct, data source, method, and expected direction of movement, then testing for stability and bias over time. For service acquisitions, content validity improves when measures triangulate buyer, provider, and end-user perspectives (Neely et al., 2002). Without this discipline, organizations risk precise metrics that are precisely wrong—reliable but invalid proxies that steer behavior away from mission outcomes (Kaplan & Norton, 1996).

Leading and lagging indicators should be anchored in a causal logic of results. Donabedian's (1988) classic structure-process-outcome model clarifies how resources and activities (i.e., structure, process) are expected to yield results (i.e., outcomes;

Donabedian, 1988). Leading indicators typically track structures and processes that theory and evidence link to later outcomes (e.g., qualified cyber personnel on contract; mean time to resolve priority incidents), while lagging indicators confirm realized results (e.g., mission availability, user satisfaction, cost to serve; Institute of Medicine Committee on Quality of Health Care in America, 2001; Kaplan, 2010). Designing the set as a minimal, theory-backed chain helps avoid the proliferation of disconnected measures and keeps review conversations focused on the few signals that matter for decision-making (Kaplan & Norton, 1996; Neely et al., 2002).

Organizational performance theory explains why public metrics often look different from private-sector dashboards. Public agencies operate under multiple principals (Congress, oversight bodies, citizens) with diffused objectives, soft budget constraints, and nonmarket outputs, which complicate incentive alignment and measurement (Hatry & Bryant, 1998). Unlike firms, government organizations cannot rely on profit to summarize value; they must balance efficiency with equity, compliance, and legitimacy across stakeholders (Bouckaert & Halligan, 2008). This makes portfolio balance across cost, timeliness, quality, user outcomes, and stewardship not just desirable but necessary to maintain credibility and control (Kaplan & Norton, 1996; Organization for Economic Co-operation and Development [OECD], 2025).

Public value and governance models reframe the terms "customer" and "value" for government work. Moore (1995) argues agencies create public value when they produce outcomes that are substantively valuable, operationally feasible, and politically legitimate. In services contexts, modern service-management views reinforce value co-creation: outcomes emerge from the interaction of provider capabilities and user processes, which implies measures must capture both sides of the relationship (Axelos, 2019; Neely et al., 2002). Practically, this means complementing output and compliance measures with indicators of user outcomes, experience, and adoption, as well as stewardship metrics that maintain trust and legitimacy (OECD, 2025; Kaplan, 2010).

Stakeholder-centric frameworks help formalize this balance. The Performance Prism starts with stakeholder satisfaction and contribution (what stakeholders want from the organization and what the organization needs from stakeholders), then aligns

strategies, processes, and capabilities—a useful scaffold for service contracts where contractor, government sponsor, and end user all shape results (Neely et al., 2002). Periodic "metricide" (retiring measures that no longer serve decisions) keeps the system lean and prevents drift toward metric accumulation (Neely et al., 2002, p. 3).

Finally, governance must anticipate strategic behavioral responses and the potential for gaming. Bevan and Hood (2006) show that target regimes are vulnerable to gaming such as "hitting the target and missing the point" (p. 9), and they cite Goodhart's law that states "any observed statistical regularity will tend to collapse once pressure is placed on it for control purposes" (p. 9), to illustrate how measures can distort behavior once they become instruments of control. Similarly, audit-culture critiques warn that excessive reliance on indicators can crowd out professional judgment when measures become detached from their original purpose (Strathern, 1997). Drawing on multiple scholars' findings, effective countermeasures include (a) mixed baskets of indicators (outcomes with a few process/assurance checks), (b) transparent definitions and stable targets, (c) random audits and data-quality ownership, and (d) routine strategy reviews that validate causal links and prune weak proxies (Brubaker et al., 2018; Kaplan, 2010; Neely et al., 2002). In short, the measurement system is part of the control system—it shapes incentives, behavior, and ultimately mission outcomes.

Overall, these foundations yield a practical design logic: Start with mission outcomes, build a minimal theory-backed chain of leading and lagging indicators, and govern the system to sustain validity, parsimony, and learning (Kaplan & Norton, 1996; Neely et al., 2002). Public managers must also anticipate behavioral response—avoiding "hitting the target, missing the point" (Bevan & Hood, 2006, p. 9) through mixed indicator baskets, stable definitions, and periodic pruning (metricide) of weak proxies (Neely et al., 2002; Strathern, 1997). Finally, by framing value as co-created between provider capabilities and user adoption, and by aligning stewardship and legitimacy alongside efficiency, measurement becomes a control system that supports accountability and public value (Axelos, 2019; OECD, 2025; Moore, 1995). This theoretical base sets the stage for the next section's move from concepts to policy and oversight practice (e.g., GAO findings, DoD governance expectations) and, ultimately, to the contract-level

heuristics the researchers apply in the Acquisition of Services context (OUSD[A&S], 2021; GAO, 2021b).

2. Public Sector and DoD Service Acquisitions

Building on the theoretical foundations from the previous section, the application of performance measurement within the DoD services acquisition enterprise reveals a complex interplay of policy design, organizational behavior, and implementation practice. Over 2 decades of reform have sought to embed performance-based principles in how the DoD plans, executes, and evaluates service contracts. This section presents an examination of how those efforts have evolved, beginning with the conceptual foundation of performance-based services acquisition (PBSA), progressing through the institutionalization of oversight mechanisms under DoDI 5000.74 and component-level policies, and concluding with assessments from the GAO, RAND, and the Institute for Defense Analyses (IDA) that reveal recurring gaps in data integration, accountability, and outcome alignment. Together, these perspectives show that while the DoD has clearly articulated its intent to manage services through measurable results, the persistence of process-heavy metrics and decentralized governance continues to hinder consistent, outcome-oriented implementation.

The DoD relies heavily on contracted services; according to the GAO (2023a), contractor-performed work represents roughly half of the DoD's total contract obligations, underscoring that performance measurement is a core acquisition management concern. Within this context, Lucyshyn and Rigilano (n.d.) define PBSA as an approach that structures contracts around measurable results rather than prescribing the processes for achieving them. This method grants contractors flexibility in how to deliver outcomes and shifts performance management from process compliance to results. DoDI 5000.74 institutionalizes PBSA principles within the AAF, emphasizing data-driven oversight and specifying data elements needed to measure progress and performance (OUSD[A&S], 2021, sec. 2.3). The instruction assigns governance responsibilities from USD(A&S) policy development to component-level data sharing to ensure that performance reviews and decision authorities are informed by measurable outcomes rather than procedural compliance. Yet, as later GAO (2021a, 2021b, 2023a,

2023b) reports confirm, implementation remains uneven, with data integration and accountability varying widely across components.

Organizational dynamics also affect how public organizations measure performance. This tendency to accumulate indicators that are easy to count rather than those that demonstrate mission value, a practice previously identified as metric mania (Casey et al., 2008), is a known challenge. To counter this, Casey et al. (2008) advocate focusing on a small set of "whole goal" (p. 97) indicators directly linked to strategic outcomes, while minimizing or eliminating metrics that measure activities or processes without clear ties to results. At the contracting-office level, traditional throughput measures, such as number of contract actions or days to award, can misrepresent true effectiveness. O'Sullivan (2003) finds that volume-based metrics such as cycle-time goals may undermine customer satisfaction and decision quality. O'Sullivan (2003) proposes a BSC model for evaluating field contracting offices that integrate efficiency, customer results, internal operations, and employee development to provide a more comprehensive assessment of performance.

DoD policy directs performance-management expectations but allows flexibility in execution. DoDI 5000.74 directs the use of tailored performance metrics and periodic reviews for major service contracts (OUSD[A&S], 2021). Component policies reinforce this guidance; DAF Instruction 63–138 (DAF, 2024) directs multifunctional teams to plan early and use performance metrics that inform requirements approval and oversight. Secretary of the Navy Instruction 4200.37A (Department of the Navy, 2019) formalizes the Procurement Performance Management Assessment Program Council to standardize self-assessments and annual reviews; and Army Regulation 70-13 (Department of the Army, 2010) emphasizes performance-based contracting, surveillance planning, and periodic reporting of contractor performance and manpower data. These reflect a clear policy intent to institutionalize oversight without uniform implementation.

Earlier GAO reviews reinforce that this challenge is structural rather than new. As early as 2006, the GAO recommended that the DoD adopt an "integrated, department-wide architecture" linking service-acquisition decisions to strategic goals and risk-based metrics but found limited progress in establishing reliable baseline data or enterprise

comparability (pp. 20–35). Later, GAO (2013) reiterated that the DoD still lacked specific goals and associated metrics to assess whether reforms were improving outcomes, citing fragmented data and unlinked financial systems as ongoing barriers to strategic oversight. These findings underscore that policy intent alone has not produced sustained measurement capability.

Despite component directives, the GAO has found persistent weaknesses. In 2013 it concluded that the DoD had not established department-wide goals or metrics to assess whether service acquisitions were improving. Subsequent GAO (GAO, 2021a, 2023a) revealed that persistent planning and data challenges continued to hinder the DoD's ability to assess progress and manage service acquisitions effectively. Across government, the GAO (2021b) observed that agencies overuse process measures like competition rates and lead times, whereas leading private-sector companies focus on outcomes such as cost savings, timeliness, quality, and customer satisfaction. RAND echoed these findings, concluding that the AoS Pathway and other AAF pathways rely largely on activity-based indicators rather than mission outcomes (McKernan et al., 2022). These findings suggest that DoD policy expectations have not yet translated to consistent, outcome-oriented implementation.

More recent observations highlight both progress and remaining gaps. The *FY 2024 Acquisition of Services Review Panel Wrap-Up* memorandum (USD[A&S], 2024b) notes increased use of automation tools and multifunctional teams within the AoS Pathway but also persistent disparities in post-award oversight and training among components. Likewise, the IDA reports that high-performing commercial firms maintain "smart-buyer" functions that centralize data, validate bids, and preserve institutional knowledge through knowledge management systems—capabilities still maturing within DoD services acquisition (Kelly et al., 2015, pp. 18–19). Together, these studies suggest that achieving outcome-oriented measurement requires not only policy consistency but also enterprise data governance, professional development, and feedback loops comparable to private-sector strategy management systems.

In summary, the literature depicts a DoD services-acquisition ecosystem that has steadily embraced performance-based principles in policy yet continues to fall short in

consistent implementation and data integration. The DoD's heavy reliance on contracted services magnifies the stakes of these shortcomings; without standardized, outcome-oriented metrics, oversight remains fragmented, and decision-making reactive. The persistent gap between intent and execution underscores the need for a more coherent, enterprise-level approach to performance measurement that links component policies, contracting behaviors, and strategic outcomes. These systemic challenges provide the foundation for the framework developed in Chapters IV and V, in which the researchers aim to design and operationalize an outcome-based performance measurement model aligned with the DoD's readiness and mission priorities.

B.   FEDERAL AND DOD POLICY AND OVERSIGHT

This section synthesizes 2 decades of oversight reports and analytical studies on DoD services acquisition, organized chronologically to show how findings evolved and where they converged. Across the GAO, Defense Science Board (DSB), DBB, RAND, IDA, and recent DoD reviews, a consistent pattern emerges: Metrics are plentiful but often process-oriented, data systems are fragmented, and links between contract performance and mission outcomes are weak. The evidence that follows highlights the most policy-relevant lessons to inform the KPI cascade developed in the next chapter.

1.   Chronological Review of Evidence

Table 1 summarizes the major studies published between 2000 and 2024, presenting each report's direct findings and the key takeaways. For example, a 2000 RAND study on strategic sourcing found that commercial buyers and providers achieved better results by jointly developing a few outcome-based performance indicators tied to cost, quality, and satisfaction (Baldwin et al., 2000). Together, these works reveal the analytical foundation for outcome-based performance measurement and inform the KPI cascade developed later.

Table 1.　Chronological Evidence Review

Year	Source / Title	Key Finding	Key Takeaway	Report ID / Citation
2000	RAND, Strategic Sourcing (Installations/FM)	Commercial buyers and providers co-designed small set of outcome KPIs (cost, quality, satisfaction) to drive value over compliance.	Shared accountability for limited outcome metric set strengthens performance	DB-287-AF (Baldwin et al., 2000)
2002	RAND, Implementing Best PSM Practices	Linked purchasing outcomes to enterprise goals through measurable standards and incentives w/KPIs refined through pilots and feedback.	Keep KPI set lean, adaptable, and linked to strategy to avoid metric sprawl	DB-334-AF (Moore et al., 2002)
2004	RAND, Installation Support Contracts	2–7 KPIs (quality + cost) tracked per service; outcome-based metrics more useful; two-way comms sustained alignment.	Limit KPIs, emphasize outcome measures	MR-1812-AF (Baldwin & Hunter, 2004)
2006	GAO, Acquisition Workforce/Data	Unreliable workforce data undercuts oversight.	Data integrity is a precondition for any KPI framework	GAO-06-987R (GAO, 2006)
2011	DSB, Improvements to Services Contracting	Advocated treating services separately from weapons; added outcome measures (quality, timeliness, continuity, productivity).	Adopt taxonomy & outcome measures; reframe KPIs to effectiveness	ADA550491 (DSB, 2011)
2011	Finkenstadt & Peterson, Air Force (AF) Program Executive Office (PEO)/CM Benchmark	Annual Execution Reviews (AERs) and Expectation Management Agreements (EMAs) provided visibility on cost/schedule/performance.	These artifacts operationalize portfolio-level KPI governance	(Finkenstadt & Peterson, 2011)
2013	GAO, Services Portfolios	Lack of measurable goals and baselines; weak portfolio authority limit improvement; incomplete Federal Procurement Data System (FPDS) limits visibility.	FPDS constraints obscure spend patterns and outcomes	GAO-13-634 (GAO, 2013)
2015	GAO, Non-Major Acquisition Oversight	Missing APBs and data errors block trend analysis.	Without standardized baselines, KPI cascades won't scale.	GAO-15-188 (GAO, 2015b)
2015	IDA, Commercial Best Practices (KBS/ERS)	Private sector links contractor performance to organizational outcomes and uses post-award analytics/scorecards	Embed outcome statements in PWS	P-5257 (Kelly et al., 2015)
2016	CRS, Using Data to Improve Defense Acquisitions	DoD lags in analytics; fragmented/unreliable data impedes outcome-aligned decisions	Integration of data systems is necessary for forecasting and prioritization.	R44329 (Peters, 2016)
2019	RAND, DoD Use of Analytics & Data Mgmt	Stovepiped data and inconsistent definitions impede cross-program KPI comparability.	Strong data governance is essential for	RR-3136-OSD (Anton et al., 2019)

				enterprise metrics.	
2021	GAO, Outcome-Oriented Metrics (Agency Practices)	End-user co-development and feedback loops balance compliance vs. mission value.	Reframe KPI design toward stakeholder relevance and outcome validity.		GAO-21-491 (GAO, 2021b)
2021	GAO, Services Inventory & Services Requirements Review Board (SRRB)	SRRBs matured, but System for Award Management (SAM)/FPDS blind spots and weak budget links still block outcome measurement	Governance exists; data quality lags.		GAO-21-267R (GAO, 2021a)
2022	RAND (Anton), Early Predictive Indicators	Developed leading indicators (CPI-like risk signals).	Shift DoD KPI thinking from lagging to predictive metrics.		RRA-542-1 (Anton et al., 2022)
2022	RAND (McKernan), AAF Pathway Metrics	Logic models reveal AoS Pathway relies on manual, activity metrics and thin post award outcomes	Outcome KPIs and pathway-aligned measures are needed.		RRA-1349-1 (McKernan et al., 2022)
2022	DBB, Business Health Metrics	Proposed 20–30 enterprise indicators in Advancing Analytics (ADVANA) across cost/schedule/performance/people/ innovation using BSC logic.	Demonstrates feasibility of a DoD-wide BSC cascade.		DBB FY23-01 (DBB, 2022)
2022	DoD, National Defense Strategy	Called for data-driven readiness and measurable implementation oversight.	Top-down demand signal for outcome KPIs across pathways.		(DoD, 2022)
2023	GAO, DoD Service Contracts: Identify & Forecast	High-risk status persists; forecasting/ prioritization of services requires better enterprise metrics.	Emphasizes KPI cascades for portfolio planning.		GAO-23-106123 (GAO, 2023b)
2024	DoD, National Defense Industrial Strategy (NDIS)	Introduces proto-KPIs for resilience and capacity link contract to industrial-base outcomes.	Aligns service KPIs with enterprise-level national-security metrics.		(OUSD[A&S], 2024)

Note. Table created by author based on findings from the GAO (2006, 2013, 2015b, 2021a, 2021b, 2023a, 2023b), Baldwin et al. (2000), Moore et al. (2002), Baldwin & Hunter (2004), Anton et al. (2019), McKernan et al. (2022), DSB (2011), IDA (Kelly et al., 2015), DBB (2022), CRS (Peters, 2016), DoD (2022), and OUSD(A&S; 2024a)

This timeline of policy and oversight reports reveals a consistent, two-decade narrative of persistent challenges in measuring and managing service-contract performance. The following section synthesizes these findings into cross-cutting themes.

2. Cross-Cutting Lessons Learned

Convergent themes across 2 decades of oversight from the GAO (2006, 2013, 2015, 2021a, 2021b, 2023a, 2023b), Baldwin et al. (2000), Moore et al. (2002), Baldwin

and Hunter (2004), Anton et al. (2019), and various defense bodies (DSB, 2011; DBB, 2022; Kelly et al., 2015), as well as recent AAF pathway assessments (Anton et al., 2022; McKernan et al., 2022; OUSD[A&S], 2024b), reveal persistent barriers to effective services-acquisition governance and measurement. Collectively, these sources underscore the need for a paradigm shift toward integrated, outcome-oriented KPIs that connect contract-level execution to enterprise mission outcomes.

The cross-cutting lessons learned are as follows:

- **Outcome Orientation Missing**. Metrics remain activity-based, with weak linkage to outcomes. Despite calls for tailored measures such as quality and timeliness in services taxonomy (DSB, 2011), oversight prioritizes compliance over results, reflected in limited use of customer satisfaction and mission-effectiveness indicators (GAO, 2021b; Kelly et al., 2015).

- **Fragmented Data Systems**. SAM, FPDS, and financial data systems lack integration, and predictive analytics remain underused. AAF review confirms current metrics do not link pathways to capability outcomes, and non-major programs lack standardized baselines (GAO, 2015b). Persistent data reliability issues hinder trend analysis and forecasting, though initiatives like ADVANA offer potential for consolidation (GAO, 2015b, 2023b; DBB, 2022; USD[A&S], 2024b; OUSD[A&S], 2024b).

- **Compliance Culture versus Performance Culture**. Oversight mechanisms emphasize rule adherence and surveillance rather than mission value. Commercial benchmarks show that co-developed KPIs foster collaboration, yet DoD structures like SRRBs continue to generate process tracking without substantive outcome measures (Moore et al., 2002; Baldwin & Hunter 2004; GAO, 2013, 2021a).

- **Workforce and Leadership Gaps**. Insufficient analytic capacity and high turnover undermine accountability and institutional learning. Reports highlight the lack of personnel skilled in data governance and metrics development, compounded by inconsistent training across pathways (Finkenstadt & Peterson, 2011; GAO, 2006; OUSD[A&S], 2024b).

- **Strategic Alignment Absent**. Metrics rarely cascade from contract performance to enterprise objectives. While logic models and BSCs could align services with readiness, current frameworks fall short, as evidenced in AAF evaluations (McKernan et al., 2022; DBB, 2022; USD[A&S], 2024b).

- **Learning and Feedback Loops Deficient**. The DoD lacks a consistent mechanism for translating performance data into institutional learning. Predictive indicators and post-award analytics remain underutilized, limiting adaptive responses in services contracting (Anton et al., 2022; Kelly et al., 2015; Peters, 2016).

These lessons from DoD reports and studies affirm that while reforms such as SRRBs, category management, and ADVANA have advanced data capabilities, systemic gaps in outcome-focused metrics continue to impede the effectiveness of services acquisition. The AAF assessments specifically found that existing metrics were inadequate for pathway performance, with the AoS Pathway requiring enhancements like certified program managers, greater SRRB flexibility, and value-tied indicators (OUSD [A&S], 2024). Collectively, these insights justify the development of a comprehensive KPI framework to strengthen governance and measurement, setting the stage for Chapter IV's design and validation efforts. Having established how performance measurement has evolved within the DoD, the next section presents an examination of transferable and comparative practices.

C. TRANSFERABLE AND COMPARATIVE PRACTICES

Literature from the private sector and international governments provides useful contrasts for understanding how outcome-based and collaborative performance frameworks operate in environments that mirror the DoD's scale and complexity but differ in flexibility and incentive structures. These comparative perspectives highlight approaches like service-level agreements (SLAs), outcome-based contracting (OBC), and relational governance that demonstrate measurable results in managing service performance. Insights from these models inform subsequent analysis of frameworks for standardizing KPIs across DoD service portfolios.

1. Private Sector and International Insights

Private-sector and international practices provide valuable contrasts for DoD service acquisitions. In industry, commercial firms frequently employ SLAs and operational-level agreements (OLAs) to define expected outcomes and monitor performance. SLAs establish measurable deliverables such as system uptime, response times, or defect rates that link provider performance to customer expectations (Axelos, 2019). In contrast, OLAs are internal agreements among support teams that specify how interdependent processes must perform to sustain SLA compliance (Kirvan, 2023). Together, these mechanisms reinforce the Information Technology Infrastructure Library

(ITIL) principles of co-creating value through transparent performance expectations across organizational layers.

ITIL, a globally adopted set of best practices for Information Technology (IT) service management, aligns IT operations with business objectives and enables systematic measurement and improvement of service delivery through processes such as incident management and continuous feedback (Axelos, 2019; Iden & Eikebrokk, 2013). Related standards, including International Organization for Standardization/International Electrotechnical Commission (ISO/IEC) 20000-,1 operationalize these principles by requiring a Service Management System encompassing performance evaluation, internal audits, management reviews, and service reporting, often integrated with ISO 9001 to achieve unified quality and service management (Cooper, 2022). Evidence from the commercial aerospace sector also supports the ITIL and ISO principles of systematic performance measurement and continuous improvement. Guajardo et al. (2012) found that outcome-based contracts improved reliability (using mean time between unscheduled removals) by 25–40% compared with time-and-materials contracts.

High-performing organizations are increasingly incorporating user-experience and satisfaction metrics into sourcing contracts, recognizing that service value extends beyond meeting technical performance targets (Gacenga et al., 2010; Herbert & Colenso, 2018). These indicators complement traditional KPIs and SLAs by capturing end-user perceptions of service quality and linking performance more closely to business outcomes. For example, process maturity assessments integrate customer satisfaction with operational metrics such as service quality and SLA compliance (Iden & Eikebrokk, 2013). In IT service management, these approaches enhance efficiency measures by connecting service delivery to organizational outcomes, while studies continue to reveal gaps and emphasize the need for empirical metrics to fully realize their benefits (Gacenga et al., 2010).

OBC is another practice with potential relevance for the DoD. Commonly used in manufacturing, logistics, and IT services, OBC ties supplier compensation to achieved results, such as system availability or performance outcomes, rather than inputs like labor hours or materials. This structure fosters innovation, risk-sharing, and collaborative

problem-solving between buyers and providers (Hypko et al., 2010). Akkermans et al. (2019) describe how collaborative KPIs enable both buyer and supplier to align incentives around shared goals, evaluating not only supplier contributions but also buyer behaviors that support performance. This represents a shift from siloed, technical SLAs to end-to-end indicators that promote collaboration and mutual accountability. Such approaches emphasize co-production and partnership in contrast to transactional contracts that define performance narrowly in terms of cost and compliance. Relational contracting literature similarly highlights trust, flexibility, and joint governance as enablers of sustained performance (Poppo & Zenger, 2002). In logistics, for instance, performance measurement frameworks for third-party providers include numerous KPIs for quality and efficiency, though underutilization reveals persistent challenges in achieving outcome-oriented results (Domingues et al., 2015). Compared with performance-based logistics—a defense analog on availability and readiness through incentive structures— OBC in private sectors often involves greater uncertainty in costs and dependencies. However, both share underlying principles of risk incentives and effectiveness factors like governance and supply chain management (Hypko et al., 2010).

International governments have experimented with models similar to the private-sector outcome-based and relational contracting approaches regarding collaborative, outcome-focused frameworks that emphasize partnership over adversarial transactions. In the United Kingdom, the Ministry of Defense and the National Health Service have both adopted outcome-based and relational approaches to contracting. Results have been mixed. For example, while the United Kingdom Ministry of Defense Availability Transformation: Tornado Aircraft Contract and Air Defence Availability Project programs successfully delivered availability metrics (e.g., mission-ready flying hours, spares availability) and reportedly saved £1.3 billion (Ng & Ding, 2010), these models also faced significant implementation challenges, including attribution of outcomes, balancing of risk-sharing, and data quality assurance (Roehrich et al., 2014). Comparable adaptations appear in Australia and Germany, where performance-based logistics-like strategies in defense procurement emphasize outcome incentives for availability and capability. Case studies highlight buyer input and supply chain integration as critical factors for effectiveness (Alqahtani et al., 2023). These lessons indicate that outcome-

based and collaborative models are viable in the public sector but require robust governance and carefully designed performance frameworks.

For the DoD, these private-sector and international insights present both opportunities and cautions. Practices such as SLAs, customer-satisfaction surveys, and outcome-based partnerships illustrate ways to link services more directly to delivered value. At the same time, persistent challenges related to attribution, incentives, and governance underscore why direct transferability remains difficult. The literature therefore supports selectively adapting private-sector approaches to fit the DoD's scale, complexity, and accountability environment. These examples of transferable practices provide essential context for the broader structural frameworks used to standardize performance metrics, which are discussed next.

2.   Frameworks for KPI Standardization

Several streams of literature and DoD-specific guidance address how performance metrics should be structured and governed to provide reliable, actionable insights across organizational levels. A recurring theme is the need for KPIs that align with strategic objectives while supporting comparability, accountability, and continuous improvement. This section synthesizes key frameworks from the research, emphasizing cascading structures, balanced perspectives, and governance mechanisms. Although not exhaustive, the review focuses on frameworks most relevant to this study's objectives. These approaches mitigate risks like *metric gaming*—optimizing reports at the expense of outcomes—and help ensure that metrics remain meaningful within complex environments like DoD service acquisitions (Bevan & Hood, 2006).

a.   *Cascading KPI Frameworks: Aligning Levels from Enterprise to Contract*

A foundational approach is the tiered cascade framework, which distinguishes among strategic, operational, and tactical levels to ensure vertical coherence (Axelos, 2019). At the strategic (enterprise) level, metrics capture broad outcomes such as mission readiness, cost stewardship, or cyber resilience. Portfolio-level measures aggregate performance across service categories, such as knowledge-based versus logistics services, enabling leaders to compare domains and identify systemic trends. At the tactical

(contract) level, metrics focus on specific deliverables like timeliness, defect rates, or customer satisfaction, which roll up to inform higher-level indicators (Brubaker et al., 2018; Neely & West, 2022).

This cascade is explicitly reinforced in the ITIL Service Value System framework, which links organizational goals to critical success factors and then to KPIs (Axelos, 2019). For example, a strategic goal such as enhancing warfighter readiness (DoD, 2022) may cascade to operational critical success factors like service availability above 95% and to tactical KPIs such as mean time to repair. Governance processes, including measure design & selection, data capture, and information provision, are essential to maintain alignment and validity (Franco-Santos et al., 2007). Within DoD contexts, this approach addresses data and governance gaps identified by the GAO (2015b) and Peters (2016), where inconsistent baselines and data reliability hinder performance tracking for non-major programs.

To operationalize cascading in the AAF AoS Pathway, frameworks recommend limiting KPIs to five to 10 per level to maintain parsimony, clearly defining data owners—such as program managers for tactical metrics—and establishing review cadences such as quarterly portfolio assessments (Neely & West, 2022). Weak governance can lead to the proliferation of disconnected measures, creating risks of perverse incentives and metric gaming (Bevan & Hood, 2006) or the audit-culture critiques (Strathern, 1997) discussed in the theoretical foundations.

b.   BSC Frameworks: Integrating Multiple Perspectives

Building on the cascading approach, the BSC framework provides a multidimensional structure that prevents over-reliance on financial or output metrics (Kaplan & Norton, 1992). This model is particularly relevant for the AAF, which must balance speed and compliance with mission-level outcomes. The BSC integrates four perspectives: financial (e.g., cost savings, return on investment), customer (e.g., satisfaction scores), internal processes (e.g., cycle time, compliance rate), and learning & growth (e.g., workforce skills gaps) (Kaplan & Norton, 1992). This ensures a holistic view, where outcomes are balanced against leading indicators such as process efficiency (Brubaker et al., 2018).

In practice, the Department of Energy (2016) adapted BSC for procurement systems, using it to shift reviews from process-oriented to results-focused assessments. Similarly, the Institute of Medicine Committee on Quality of Health Care in America (2001) applied the BSC to workforce health management, tracking readiness compliance and patient satisfaction across perspectives. Hristov and Chirico (2019) extend this model into a sustainability balanced scorecard by incorporating environmental and social dimensions (e.g., resource consumption, skills gaps), which potentially relevant for DoD (2022) priorities like climate resilience outlined in the National Defense Strategy. Within DoD service acquisitions, the BSC could map contract-level KPIs (e.g., on-time performance or delivery) to portfolio outcomes—for example, readiness increases from the NDIS proto-KPIs like "increase in DIB capacity" (DoD, 2023).

Governance in the BSC framework emphasizes strategy mapping and periodic reviews to adapt measures over time, addressing common DoD data limitations such as unreliable FPDS data (Peters, 2016). The OUSD(A&S; 2024b) AAF review reinforces this point, noting that "existing metrics do not indicate policy effectiveness" and recommending measures for "capabilities delivered and value" (pp. 6–20).

c.   *The Contract Scorecard*

An adaptation of the BSC tailored specifically for outsourcing contracts is the Contract Scorecard framework proposed by Cullen (2016). Drawing on Kaplan and Norton's (1992) multidimensional approach, Cullen (2016) reframed the perspectives into four quadrants: Quality (e.g., precision, reliability, speed, effectiveness, satisfaction), Financial (e.g., cost reduction, benchmarking, total cost of ownership), Relationship (e.g., power dynamics, mutual reviews, trust drivers), and Strategic (e.g., innovation, business contribution, underlying practices like staff continuity). Together, these perspectives provide a holistic basis for evaluating contract performance. The framework emphasizes designing contracts *by design*—that is, intentionally structured to drive successful outsourcing—beginning with stakeholder assessment and cascading down to actionable KPIs that balance leading and lagging indicators. Unlike traditional process-heavy metrics, it prioritizes outcomes linked to stakeholder value and follows a seven-step process for KPI development: allocating responsibilities, assessing stakeholders, listing

KPIs, setting thresholds, generating calculations, prototyping reports, and testing for validity (Cullen, 2016).

For DoD service acquisitions, the Contract Scorecard provides transferable insights by addressing gaps in relational governance and strategic alignment noted in oversight reports (GAO, 2021b; McKernan et al., 2022). It mitigates risks like *metric gaming* (Bevan & Hood, 2006) through parsimonious, theory-based measures and promotes co-creation of value between buyer and provider, similar to collaborative KPI contracting (Akkermans et al., 2019). Although originally developed for the private sector, the framework could be adapted to incorporate DoD-specific elements, such as mission readiness in the *Strategic* quadrant or stewardship in the *Financial* quadrant to enhance the AAF's data-driven oversight (OUSD[A&S], 2021). This framework underscores the value of minimal, adaptable scorecards that evolve with contract life cycles, providing a bridge from tactical deliverables to enterprise outcomes.

d. Structure–Process–Outcome and Purpose-Driven Frameworks

Donabedian's (1988) structure-process-outcome framework, widely applied in performance measurement, categorizes metrics into structural enablers (e.g., governance policies), processes (e.g., acquisition workflows), and outcomes (e.g., health, financial or productivity results; Institute of Medicine Committee on Quality of Health Care in America, 2001). The model distinguishes between leading indicators, which emphasize proactive prevention, and lagging indicators such as defect rates, thereby promoting a continuous feedback loop for improvement. This approach aligns with Behn (2003), who identified eight purposes of measurement—including evaluation, motivation, and learning—and matched indicator types to each (e.g., outcomes for evaluation and real-time outputs for motivation. Within the DoD, this model could help standardize AoS Pathway KPIs by purpose, such as using customer satisfaction data (American Customer Satisfaction Index, 2024) for promotion or disaggregated data for learning (Peters, 2016).

e. Collaborative and Relational Frameworks for Interdependent Services

For outsourced services with significant buyer-supplier interdependencies, collaborative KPI contracting introduces bilateral metrics that evaluate the performance

of both parties such as supplier reliability and buyer demand discipline (Akkermans et al., 2019). This uses KPI trees and joint governance structures, such as a collaborative office, to map causal linkages and reduce moral hazard. It is analogous to Kirvan's (2023) description of OLAs, which states that OLAs KPIs such as mean time between failure and mean time to repair in internal service agreements. Applied to the DoD context, this framework can strengthen vendor relationships and improve compliance tracking as demonstrated by Fagundes (2016).

f.   DoD-Specific Assessment and Governance Frameworks

The NDIS assessment framework (DoD, 2023) provides proto-KPIs as illustrative outcomes and outputs such as, "decrease in cybersecurity incidents," and "reduction in skills gaps" (p. 24). These are accompanied by near-term tracking through intermediate measures and a call for dedicated data analysis. The GAO's (2005) *Framework for Assessing the Acquisition Function* emphasizes policy and process oversight across the acquisition life cycle. Building on this, Finkenstadt and Peterson (2011) illustrate how Air Force organizations apply the framework through tools like quad charts (covering scope, cost, and performance history) and AERs to evaluate progress. Both tie closely to AAF governance discussed by USD(A&S; 2024) and OUSD[A&S] (2024b), which prescribes metrics for pathway performance and data collection. Strategic alignment anchors these frameworks in the DoD's key guiding documents, which prioritize readiness, people, innovation, cyber resilience, and stewardship (DoD, 2022, 2023; OUSD[A&S], 2024a). Metrics could be mapped to these priorities, such as timeliness indicators to readiness objectives and workforce measures to the people dimension.

In summary, effective standardization draws on multiple complementary principles: cascading for alignment, the BSC for multidimensional balance, the Contract Scorecard for integrating relational and strategic dimensions, purpose-driven selection for relevance, and collaborative mechanisms for interdependencies. DoD-specific frameworks outlined in strategic documents offer tailored governance approaches, yet persistent gaps (data reliability) remain a central challenge (GAO, 2021b, 2023b). Implementing these frameworks require limiting KPI sets, clarifying definitions, and embedding review cycles to ensure decision usefulness and sustained strategic impact.

3.   Relevant Academic Studies and Research at the Naval Postgraduate School

The Naval Postgraduate School (NPS) has played a pivotal role in advancing research on defense acquisition, particularly in the areas of service contract performance and metrics. As a leading academic institution serving military and civilian acquisition professionals, NPS produces theses and studies that frequently draw on empirical data and practitioner insights to address real-world acquisition challenges. This section synthesizes key NPS research thematically, rather than chronologically, to build a conceptual understanding of how this body of work has evolved and where gaps persist in standardization, management practices, and outcome measurement.

a.   Foundational Studies on Quality, Workforce, and Efficiency (1989-2003)

Early NPS research established the foundational elements of contract performance. Barnd (1989) identified four primary factors influencing service quality—quality assurance, contractor capability, the statement of work, and contract type—providing an early framework for defining quality metrics. Subsequent studies examined the influence of policy and structure. Pope (1997) evaluated the effects of the Defense Acquisition Workforce Improvement Act on workforce quality, offering early insight into human capital as a performance enabler. O'Sullivan (2003) questioned the effectiveness of prevailing performance models, finding that systems used to evaluate field contracting offices were inadequate and recommending a balanced model to capture efficiency and effectiveness, moving beyond narrow output-based measures.

b.   Empirical Analysis of Management, Oversight, and Sourcing (2009-2011)

A subsequent wave of researchers empirically examined how management and oversight function in practice. Stevens et al. (2009) explored the management of specialized medical services, emphasizing the importance of market research in acquisition planning. Finkenstadt and Peterson (2011) conducted a benchmark study of the Air Force Program Executive Office for Combat and Mission Support, identifying governance tools like AERs and quad charts as effective mechanisms for standardized

oversight. During the same period, other studies analyzed specific drivers of performance, with Muir (2010) modeling the determinants of sourcing performance and McFall and La (2011) providing an empirical assessment of Army Mission Installation Contracting Command management practices, revealing variations in efficiency.

c.   Defining and Measuring Contract Success (2012)

A pivotal pair of studies in 2012 examined the fundamental question of how to define "success." Miller et al. (2012) applied stakeholder theory and found no standardized definition or measurement of success in DoD service contracts which is a critical gap this research also seeks to address. Similarly, Hagan et al. (2012) analyzed U.S. Navy contracts and determined that stakeholders prioritized outcome-related factors over process-oriented ones, reinforcing the need for a shift in measurement focus.

d.   Analysis of Performance Levers, Metrics, and Models (2013–2022)

More recent NPS research has focused on specific levers, metrics, and models that influence contract performance. Wilhite et al. (2013) identified management levers, like performance reporting, that drive contract success. Baker et al. (2016) provided quantitative analysis linking source selection strategies (*lowest price technically acceptable* versus *tradeoff*) to differing performance outcomes. Brubaker et al. (2018) directly critiqued existing performance metrics within DoD agencies, finding that they lacked balance and overemphasized outputs such as lead time; they proposed a hybrid framework (Three Integrated Pillars of Success-Efficiency Effectiveness and Process) to integrate personnel, platforms, and protocols with a balance of input, process, output, and outcome metrics. Other studies expanded the inquiry: Peters et al. (2019), explored maturity models; Hunter et al. (2019) analyzed the quantitative effects of service complexity on performance, and Lowry and Droeske (2022) clarified the definition and application of process metrics such as PALT.

e.   Synthesis of NPS Research

Collectively, this body of NPS research demonstrates a clear scholarly progression. Early studies identified foundational challenges in quality and efficiency (Barnd, 1989; O'Sullivan, 2003). Subsequent work moved toward empirical analyses of

management practices and the definition of success, consistently revealing stakeholder preferences for outcome-based measures (Hagan et al., 2012; Miller et al., 2012). More recent research has focused on identifying specific performance levers and models (Brubaker et al., 2018; Wilhite et al., 2013) yet confirms that a balanced, standardized, and outcome-oriented framework remains elusive. These studies provide a robust empirical foundation for this project and highlight the persistent, well-documented nature of the measurement gap this research seeks to address. When combined with the broader policy and theoretical literature, this focused body of work points directly to several enduring deficiencies, which are synthesized in the next section.

4. Identified Gaps

The literature on performance measurement in service acquisitions reveals significant progress in theoretical frameworks, policy guidance, and comparative practices, yet it also exposes enduring gaps that hinder the DoD's ability to effectively manage and evaluate its AoS Pathway. These shortcomings are particularly salient when viewed through the lens of the project's objectives: identifying current metrics, benchmarking them against private-sector practices, proposing a standardized framework, and assessing implementation feasibility. Recent oversight reports continue to cite chronic challenges in DoD acquisition reform, such as data fragmentation, weak cross-organizational integration, and slow adoption of adaptive pathways, while leaving specific deficiencies in services metrics unresolved. This section presents a synthesis of those gaps, an explanation of their implications, and an underscoring of why they necessitate targeted research such as this study.

First, despite clear policy directives in DoDI 5000.74 and the AAF, the DoD lacks a standardized set of outcome-based metrics tailored to service acquisitions. Oversight reports from the GAO (2013, 2021a, 2021b, 2023a) and RAND (McKernan et al., 2022) consistently document an overreliance on the process-oriented indicators, such as procurement lead times and competition rates, discussed in Sections B.2 and C.2, while outcome-focused measures remain fragmented and inconsistently applied across service categories. This is a critical gap because it prevents accurate assessment of program health and effectiveness. Existing metrics emphasize procedural efficiency over

substantive results such as service quality, cost-effectiveness, and mission impact. Without standardization, DoD components operate largely in silos, resulting in duplicated efforts, inefficient resource allocation, and an inability to demonstrate value from the roughly half of all contract obligations dedicated to services (GAO, 2023a). The persistence of this issue reflects decentralized governance and varying implementation maturity, as noted in cross-cutting lessons from earlier sections (e.g., fragmented data systems and compliance culture), and it requires research to develop a cohesive set of outcome-based metrics that align with the AAF's intent for data-driven oversight.

Second, a related gap concerns comparability and integration across organizational levels. The literature reveals no integrated, logic-based framework that cascades metrics from contract-level outputs, such as timeliness and quality, to portfolio and enterprise-level outcomes like mission readiness and strategic alignment. McKernan et al. (2022) and the AAF's 120-Day Review (OUSD[A&S], 2024b) emphasize that this absence limits leaders' ability to aggregate performance data, compare portfolios, or link service contracts to broader DoD objectives, a concern reiterated in recent assessments of AAF pathways. Data-quality challenges compound this problem, with definitional inconsistencies and unreliable systems (FPDS and SAM) distorting validity and perpetuating fragmented reporting (Brubaker et al., 2018; GAO, 2015a). Why is this a gap? It erodes enterprise visibility, hampers strategic decision-making, and contributes to the metric mania described by Casey et al. (2008), in which indicators proliferate without meaningful connections. Empirical studies from NPS, such as those by Hunter et al. (2019) and Brubaker et al. (2018), further illustrate this challenge, showing mixed results influenced by service complexity and relational factors but no scalable framework for comparability. Addressing this gap requires research to develop KPIs that ensure validity, parsimony, and alignment, filling a void that policy alone has not addressed.

Finally, a profound gap remains in explicitly linking service-acquisition performance to mission readiness and public value. While private-sector and international literature, such as outcome-based contracting (Guajardo et al., 2012) and relational frameworks (Akkermans et al., 2019), emphasize metrics like customer satisfaction, availability, and co-created value, defense-specific research has yet to produce a validated, DoD-adapted framework that connects service outcomes to operational

effectiveness. This shortfall is evident in chronological evidence (DSB, 2011; DBB, 2022; USD[A&S], 2024b) and NPS studies (Hagan et al., 2012; Miller et al., 2012), which identify stakeholder preferences for outcome metrics but reveal no standardized implementation. As a gap, it means the DoD cannot reliably demonstrate how services contribute to strategic priorities outlined in the *National Defense Strategy* (DoD, 2022) or the *National Defense Industrial Strategy* (OUSD[A&S], 2024a). This can lead to misaligned incentives and the kind of "hitting the target, missing the point" dynamic described by Bevan & Hood (2006), resulting in suboptimal mission impacts. The persistence of this gap reflects enduring challenges in data collection, outcome attribution, and the tendency to prioritize major weapons systems over services, as noted by the GAO (2021b). This research directly addresses the gap by benchmarking DoD practices against private-sector KPIs, assessing data challenges, and proposing feasible outcome-based metrics aimed at shifting the culture from process compliance to outcome accountability, thereby enabling better-informed decisions and enhanced mission effectiveness in service acquisitions.

In essence, these gaps reflect a systemic shortfall in the DoD's AoS Pathway, which is the absence of a standardized, outcome-oriented framework that integrates theory, policy, and practice for measurable mission impact. They persist because they perpetuate inefficiencies, diminish accountability, and hinder the AAF's objectives for adaptive, data-driven acquisition. Despite 2 decades of oversight and reform, no comprehensive solution has emerged, as the latest reports continue to document recurring challenges without pathway-specific resolutions. This reality justifies dedicated research, which seeks to fill that void by developing tailored metrics, evaluating their feasibility, and providing actionable recommendations to improve assessment, comparability, and alignment with DoD priorities. Addressing these needs through literature review, data collection, interviews, and comparative analysis advances both theoretical understanding and operational relevance, as detailed in the following chapters.

D.   SUMMARY

This chapter synthesized the literature on performance measurement and management within the DoD's AoS Pathway, integrating theoretical foundations, public-

sector scholarship, federal and DoD policy oversight, transferable private-sector and international practices, frameworks for KPI standardization, relevant NPS studies, and identified gaps. It traced the evolution of performance measurement from process-oriented compliance to outcome-based accountability, distinguishing outputs from outcomes, and highlighting risks like metric mania and Goodhart's law. Key frameworks such as the BSC, Performance Prism, Donabedian's (1988) strategy-process-outcome model, and cascading KPI structures, were examined for their contributions to aligning performance metrics across enterprise, portfolio, and contract levels. The review also analyzed DoD-specific policy instruments (e.g., DoDI 5000.74), chronological oversight from the GAO, RAND, and others, empirical NPS research on contract success factors, and private-sector models including SLAs, OBC, and relational governance.

Despite this extensive body of work, significant gaps persist. The DoD has yet to implement a standardized set of outcome-based metrics across service categories, continuing to rely on fragmented, process-heavy indicators that constrain comparability, strategic decision-making, and enterprise visibility. Data quality issues, definitional inconsistencies, and challenges in attributing outcomes to mission readiness remain unresolved at scale, as reflected in continuing GAO critiques of performance measurement in related domains like IT and weapon systems. These shortcomings underscore the urgency of targeted research that bridges theoretical, policy, empirical, and practical insights to develop a coherent, adaptable framework for outcome-oriented performance measurement. This review confirms that the core deficiencies—absence of standardized outcome metrics, poor data integration, and a weak link to mission value— are enduring and well-documented. These findings validate the necessity of this research, which aims to design the framework the literature proves is missing, thereby strengthening accountability, mission alignment, and overall effectiveness in DoD service acquisitions.

IV. METHODOLOGY, DATA, ANALYSIS & FINDINGS, PROPOSED FRAMEWORK

This chapter details the research methodology, analysis, and findings which present the proposed metrics of the study. It is organized into three main parts, following the logic of the research process:

1. **Methodology**: the "why" behind the research. This section outlines the overall research strategy and the theoretical grounding that guided the study's design.
2. **Analysis**: "how" the research was conducted. This section details the specific, methodical application of techniques used to collect, process, and evaluate the data.
3. **Findings**: "what" the analysis discovered. This section presents the outcomes of the analysis, including the key themes synthesized from stakeholder interviews and the set of validated KPIs that form the basis of this research.

A. METHODOLOGY: THE RESEARCH STRATEGY (THE WHY)

This study employs a qualitative, exploratory research design to develop a comprehensive framework for standardized, outcome-based performance metrics for the DoD's AoS Pathway. A qualitative approach was necessary because, as Miller et al. (2012) identified, there is no agreed-upon definition of success in DoD service contracts, meaning stakeholder values must be understood before measurable constructs can be defined. Grounded in established performance measurement theory, such as the BSC, and informed by the persistent gaps identified in the literature, this methodology enables in-depth examination of current practices, stakeholder perspectives, and policy constraints. Qualitative inquiry is valuable in this context because limited consensus or standardized practices exist across the DoD; it allows the research to probe how metrics are interpreted, implemented, and influenced by organizational culture and policy.

Through open-ended questioning and thematic analysis, the study uncovers the practical challenges and contextual factors that shape metric development and adoption, ensuring that the resulting framework is both empirically grounded and operationally relevant. A phased research design provided the structure needed to move from broad theoretical exploration to a refined set of practitioner-validated recommendations. This

sequential approach ensured that each step of the inquiry contributed to developing a relevant operationally grounded framework. The phases included:

- **Phase 1 (KPI collection)**: Conducting a comprehensive review of academic literature, policy guidance, oversight reports, and existing KPI repositories to gather a large, theoretically relevant pool of KPIs and metrics used in the field. This phase highlighted common themes, gaps, and opportunities for standardization.

- **Phase 2 (synthesis & refinement)**: Consolidating and refining the initial KPI list by identifying cross-category commonalities and removing redundancies, resulting in a coherent set of metrics aligned with strategic and operational priorities. Questions such as which metrics are most widely applicable across contexts, and which KPIs best capture meaningful outcomes guided this process.

- **Phase 3 (validation)**: Validating the refined KPI list through semi-structured interviews with senior DoD officials, industry leaders, and subject matter experts. This step ensured that the metrics were realistic, relevant, and aligned with stakeholder expectations and operational realities. It also provided an opportunity to gather feedback on potential implementation challenges.

- **Phase 4 (framework development)**: Integrating qualitative insights from interviews and KPI synthesis to develop the standardized, outcome-based metric framework, presented in Chapter V. The resulting framework aims to be flexible, practical, and aligned with strategic objectives.

This phased qualitative strategy supports the creation of a practical, adaptable, and standardized outcome-based metric framework that is tailored to the DoD service acquisition environment. By integrating stakeholder interviews with documentary evidence, the approach ensures that the framework is grounded in operational realities, policy guidance, and strategic priorities. It also enables continuous refinement stakeholder buy-in by capturing contextual factors and diverse viewpoints that quantitative methods might miss. Overall, this comprehensive approach helps bridge the gap between theory and practice, fostering a more effective performance measurement paradigm within the defense services context.

B.   ANALYSIS: THE METHODICAL APPLICATION (THE HOW)

This section details the specific methods used to execute the research strategy. It covers how participants were selected, how data was collected, and how both qualitative and quantitative data were systematically analyzed.

1. Interview Sampling and Participants

A total of 24 individuals were interviewed to gather diverse perspectives on current practices, challenges, and opportunities related to performance measurement within the DoD's services acquisition ecosystem. Participants were selected based on their extensive experience in managing, awarding, or developing requirements for services acquisitions. This purposive sampling aimed to ensure insights from a broad cross-section of the defense and related sectors (Bullard, 2024).

Participants represented a diverse range of organizations, including component headquarters such as the Air Force Sustainment Center, Department of the Army, DAF, Defense Advanced Research Projects Agency (DARPA), Defense Contract Management Agency (DCMA), Defense Logistics Agency (DLA), Defense Threat Reduction Agency (DTRA), Defense Health Agency (DHA), Naval Air Systems Command (NAVAIR), Naval Facilities Engineering Command (NAVFAC), and the Space Force. The sample also included representatives from Boeing, Jacobs Engineering, Northrop Grumman, Raytheon, and Lockheed Martin. Additionally, some participants work for municipalities and the National Park Service.

This diverse pool of expertise spans multiple levels of organization from installation and wing-level personnel to senior leadership, policy, and oversight offices, including the Office of the Secretary of Defense. The participants represented a range of functional areas, including contracting officers, program managers, functional leads, category managers, analysts, and senior leaders involved in services contracting and requirements development (see Table 2).

Table 2.   Interviewee Profile & Summary

Category	Representative Organizations	# of Interviews	Avg Yrs of Experience
DoD–AF	AFSC HQ, DAF Staff, Contracting Officers, PMs	11	20+
DoD–Army	Army Medical Command (Health Readiness Contracts)	1	20
DoD–Navy/Pentagon	NAVAIR, NAVFAC, MDA HQ Program Directors	3	20+

DoD–Agencies/Others	DCMA, DLA, DHA, VA, National Park Service, TRANSCOM, Contract Specialists	4	20+
Local Government	County Procurement Office (FL)	1	15+
Industry–Defense Primes	Raytheon, Northrop, Boeing, Jacobs Engineering	3	25+
Industry–Small/Other	Tribal Company, T Rowe Inc. PMs	2	20+
Total Interviews	-	24	~ 23 yrs

Note. Interviewees are grouped by organizational category rather than listed individually. Representative organizations illustrate the types of roles and offices included in each category. "Avg Yrs of Experience" reflects the approximate mean experience of all participants in that category. The table is intended to convey the diversity and depth of acquisition expertise included in the sample.

2. Interview Protocol and Data Collection

The semi-structured interview instrument was derived from the study's research purpose and informed by the theoretical framing and gaps identified during the Phase 1 literature and policy review. The guide was designed to elicit evidence that could validate and refine the initial set of KPIs while capturing variation in how stakeholders define and assess success. To support this goal, the interview protocol was organized into five sections, with question sets tailored to participant role and applied flexibly depending on the interviewee's expertise and time available.

Section 1 (Background) established participant context through a standardized opening question on role and services-acquisition experience, with follow-ups probing tenure in service acquisition and whether participants primarily worked with outcome-based or process-based metrics.

Section 2 (Agency-level questions) assessed whether a standardized performance-measurement process existed at the enterprise level; depending on the response, follow-ups examined either informal practices or—where standards existed—the required KPIs, selection rationale, reporting chain, accountability for reviewing performance data, and actions taken when services did not meet expectations.

Section 3 (Requirement owner questions) shifted to program execution and focused on critical success factors, current measurement practices and PWS/QASP data

requirements, alignment (or misalignment) with any agency-level standards, practical challenges in collecting and using data, and an unconstrained "ideal metrics" prompt that asked what they would measure if current constraints did not apply.

Section 4 (Industry representative questions), used only with contractor participants, captured the industry perspective on effective service-contract KPIs, the types of data typically collected to demonstrate performance, use of benchmarks, and private-sector performance-measurement approaches that could translate to DoD services acquisition.

Section 5 (Closing and additional insights) invited any missed considerations, allowed participants to raise issues not captured elsewhere in the guide, and asked whether they were willing to review preliminary findings or be re-contacted for clarification.

In addition to these question sets, most interviews incorporated a structured KPI adjudication exercise using the draft KPI list. After the conceptual questions, the interviewer shared the KPI list and walked through selected indicators, asking participants to comment on each KPI's clarity, usefulness, typical data sources, applicability across programs or services, and any conditions, caveats, or re-wording needed. Participants were also asked to identify or rank those KPIs they viewed as most critical for assessing services-program health. This exercise provided systematic, item-by-item validation and refinement of the candidate KPIs while anchoring abstract discussions of "good metrics" in concrete examples.

Each interview followed a repeatable funnel logic embedded in the guide. Sessions opened with rapport-building and a recording/consent notification, then moved from context-setting (Section 1) into progressively more analytic and future-oriented prompts (Sections 2–4), concluding with the KPI adjudication exercise and closing questions. The protocol intentionally transitioned from descriptive current-state questions ("what do you measure today and how?") to diagnostic constraint questions ("what limits good measurement?") and finally to prescriptive ideal-state questions ("what should be measured to reflect mission outcomes?"). This sequencing ensured cross-interview

comparability while preserving flexibility for follow-up probes and emergent themes relevant to the research questions.

A full copy of the semi-structured interview questionnaire is provided in the appendix, and the candidate KPI list used during the adjudication exercise is provided in Table 3. Interviews were predominantly conducted via Microsoft Teams, with one session held via FaceTime. Sessions lasted approximately 45 minutes to 1.5 hours and were conducted between August 6, 2025, and October 29, 2025. All interviews were audio-recorded and transcribed using Otter.ai to support systematic thematic analysis.

a. Use of interview Data

This study relied on semi-structured interviews as a primary qualitative data source. To preserve confidentiality in accordance with NPS IRB approval, all participants are anonymized and cited in the text as *Interviewee 1–24*, along with a brief descriptor of their role (e.g., "senior contracting officer," "requirements owner"). These interviews are treated as research data rather than as personal communications; therefore, they are described in the methodology and quoted directly in the findings but are not listed in the reference section.

3. Document and KPI Data Collection

The KPI Universe was constructed during Phase 1 through a structured review of the study's source documents (policy, oversight, literature, and contract-execution artifacts). The source set deliberately spanned multiple evidence types to capture both formal requirements and real-world execution signals. Specifically, KPIs were extracted from: (a) DoD and Service policy, instructions, playbooks, and performance-management guidance; (b) oversight and audit products (e.g., GAO, DSB, RAND, IDA, and NPS studies) that identify what is currently measured and where outcome gaps persist; (c) contracted execution artifacts such as PWS/SOO/QASP templates, SLAs, incentive constructs, surveillance plans, and CPARS narratives that reveal how performance is operationalized post-award; and (d) internal analytic products and dashboard repositories used for portfolio or enterprise reporting. Rather than adopting an existing published KPI repository, the research team extracted candidate measures and measurement concepts

directly from these source files and consolidated them into a single master list. Where the source materials implied a performance objective but did not specify a measurable indicator, the team used Claude Sonnet 4 within NIPRGPT to propose draft KPIs to close those gaps. All LLM-suggested metrics were treated as candidates, not ground truth, and were retained only when they aligned with the study's theoretical framing and could plausibly be measured in DoD services environments.

This process produced an intentionally expansive initial pool of approximately 338 KPIs, captured in the attached KPI universe supplemental. The pool was inclusive by design at the collection stage to preserve the full range of possible outcome and process measures across heterogeneous service categories. In Phase 2, the KPI universe was normalized for analysis by collapsing duplicates, reconciling synonymous measures, and standardizing definitions so that each KPI could be compared and coded consistently in later steps. The resulting dataset served as the baseline for systematic classification and for structured elicitation of stakeholder feedback during interviews.

4.   KPI Data Analysis and Refinement

The KPI universe described above was analyzed in two linked stages: (a) structured classification of each candidate metric, and (b) iterative refinement driven by stakeholder adjudication.

- **KPI Classification and Coding**. Each KPI in the 338-metric universe was coded using a common schema to enable cross-category comparison and to prepare the list for reduction. Coding dimensions included:

  - *Temporal orientation*: leading vs. lagging
  - *Result orientation*: process vs. outcome
  - *Category alignment*: cost, schedule, performance, customer satisfaction, regulatory compliance
  - *Balanced Scorecard perspective*: financial/stewardship, customer/ readiness, internal process, learning & growth/innovation
  - *Tier level*: contract, program/portfolio, enterprise

- **Stakeholder-Driven Refinement and Selection**. During Phase 3 interviews, participants were walked through the candidate KPI list and asked to assess each measure's usefulness, feasibility, caveats, and

applicability within real services programs. This adjudication functioned as the primary validation mechanism for the KPI universe: interview feedback identified which measures were realistically collectible, which were redundant or misleading, and where outcome-alignment was strongest or weakest. Metrics that consistently failed feasibility or relevance tests were removed, while measures supported by multiple stakeholders across roles were retained. Through this structured reduction, the KPI set was narrowed from 338 candidates to a validated cohort of 21 KPIs. Final selection proceeded in consultation with the project sponsor, resulting in 7 core KPIs positioned as foundational indicators for standardized services-acquisition performance management. The final set of 7 KPIs was supported by an estimated >90% convergence, as confirmed in the systematic thematic analysis. The researchers estimate the consensus level for the broader Top 21 cohort at approximately ~75-80% (Broad/Conditional Consensus) using the same analysis. This leads the team to believe these 21 KPIs still have viability at other tiers. The original 338 KPIs had a < 20% consensus and should only be used at the discretion of the contract level team.

5.   **AI-Assisted Systematic Thematic Analysis**

This study analyzed interview transcripts using the AI-assisted systematic thematic analysis process outlined by Naeem et al. (2025). Naeem et al.'s (2025) method adapts Braun and Clarke's (2022) six-step thematic approach into a research-question-driven workflow suitable for generative AI support. The six steps are as follows:

1.   Transcription, familiarization, and quotation selection
2.   Keyword selection guided by the "6 Rs" criteria
3.   Coding by grouping quotations and keywords into conceptually meaningful labels
4.   Theme development by organizing codes into higher-order categories aligned to the research questions and theoretical framing
5.   Conceptualization through interpretation of the relationships among themes and codes

6. Development of a conceptual model/framework to represent the final concepts and their linkages.

The defining feature of this systematic approach is that themes are generated to address the research gap and questions, rather than simply emerging from frequency-based commonalities in the data. The AI-generated summaries were treated as preliminary analytic products. Each AI output was rigorously checked against the original transcript by a researcher to confirm accuracy. Once verified, the insights from each interview were synthesized and grouped into higher-level thematic categories. All personally identifying information was removed from the transcript documents to protect confidentiality. This AI-assisted approach accelerated the initial synthesis of complex qualitative information while the research team performed final verification, curation, and interpretation of all themes.

In practice, the research team implemented Naeem et al.'s (2025) steps using Anthropic's Claude Sonnet 4 within NIPRGPT as an analytic assistant. Although Naeem et al. (2025) operationalized their six-step process with ChatGPT, the procedure itself is model-agnostic; therefore the team applied the same structured prompts with Claude Sonnet 4, which was selected based on its strong performance and higher ARC-prize leaderboard rating relative to the ChatGPT variant used in Naeem et al.'s (2025) study, as well as NIPRGPT's data-handling posture in which model interaction data are owned and controlled by the U.S. Government rather than OpenAI. This additional safeguard was important to further protect participant information, even though the interview transcripts used for analysis were anonymized prior to ingestion. Prior to analysis, the model was provided with the study's purpose, research questions, theoretical underpinnings, KPI list, and the full transcript corpus to ensure contextual grounding.

Following step 1 (familiarization and quotation selection), the model was prompted to extract high-salience quotations organized under research-question-specific extraction lenses, for example, medical contracting practitioners' descriptions of outcome-based personal services and telephone access standards (e.g., aiming for 95% of calls answered within two minutes to protect patient satisfaction and access to care), as well as senior leaders' concerns that traditional cost/schedule metrics fail to indicate whether services are actually delivering mission value. In step 2, the model then proposed

candidate keywords anchored in those excerpts—such as mission impact, customer satisfaction, service health, standardized metrics, enterprise visibility, and data quality—which were iteratively reviewed and consolidated by the research team into a vetted keyword set.

In step 3, the vetted quotations and keywords were used to generate concise abductive codes (≤3 words) that captured underlying ideas, including codes such as Mission-Driven Value, Customer-Centric Results, Fragmented Measurement Systems, Commercial Efficiency Models, and Metric Value Proposition. These codes were checked against the raw transcripts and refined by a researcher to ensure they faithfully represented stakeholder meaning. In step 4, the verified codes were clustered into candidate themes, such as Mission-Outcome Alignment Crisis (connecting mission-driven value, customer-centric results, and time-critical performance) and Enterprise Fragmentation and Visibility Paradox (linking fragmented measurement systems, data system fragmentation, and enterprise visibility gaps), each supported by multiple representative excerpts.

Step 5 focused on reviewing, defining, and naming the final themes: overlapping or overly broad themes were merged or split, and each theme (e.g., Mission-Outcome Alignment Crisis, Commercial-Military Measurement Translation Challenge, Measurement Implementation Ecosystem Barriers, and Integrated Performance Framework Evolution) was given a 2–3 sentence definition explaining what it revealed about DoD services performance measurement and why it mattered for standardized KPIs. Finally, in step 6, the refined themes were synthesized into conceptual frameworks by mapping theme-level insights to specific measurement domains—for example, translating Mission-Driven Value and Customer-Centric Results into mission readiness indicators, end-user outcome measures, and service health indicators, and linking fragmentation and implementation barriers to stewardship, process excellence, and innovation measures across the enterprise. A persistent human-in-the-loop verification at each step preserved analytic rigor and ensured that responsibility for interpretation remained with the research team, consistent with Naeem et al.'s (2025) transparency and accountability guidance.

C.   FINDINGS (THE WHAT)

This section presents the outcomes of the analysis. It begins with a comprehensive summary of the findings, followed by a detailed breakdown of the seven major themes derived from the systematic thematic analysis of stakeholder interviews, and concludes with the validated set of KPIs.

1.   Summary of Findings

This comprehensive analysis, synthesizing interviews with a diverse group of DoD and industry acquisition professionals, reveals a deep and shared understanding of the challenges in measuring service contract success. The findings are categorized into seven interconnected themes that explain the current state of services acquisition and the path forward. The analysis identifies a primary mission-anchored definition of "Services Health," where success is defined not by administrative compliance, but by the service's contribution to operational readiness. Stakeholders overwhelmingly endorsed a balanced, multi-perspective health scorecard, rejecting single-point metrics in favor of a holistic view that includes cost, schedule, performance, and customer satisfaction.

The data highlights service quality and reliability as the primary evidence of outcome health while cost-to-value stewardship emerged as the necessary replacement for simple spend rate tracking. A major insight was the identification of a leading indicator structure-process-outcome chain, where foundational elements like requirements clarity and workforce stability are seen as the strongest predictors of future success. Addressing the tension between standardization and customization, the findings point to a standardized Core-Plus-Tail governance model—a common core of metrics supplemented by program-specific indicators. Finally, the analysis underscores that cultural and data distortions, such as risk aversion and manual data burdens, pose the greatest threat to the validity of any new measurement system.

2.   Key Themes from Stakeholder Interviews

The following seven themes represent the most significant points of consensus derived from the qualitative interview analysis, structured according to the Naeem et al. (2025) thematic framework.

a. Theme 1: Mission-Anchored Definition of Service Health

Core Idea: Health is judged by whether the service moves the mission needle.

Participants consistently define healthy services as those with an explicit line-of-sight to mission and readiness consequences. Outcome measures are viewed as valid only when they demonstrate how service performance protects critical operations. As Interviewee 1, a senior Navy stakeholder noted, the ultimate definition of success is simple: "my definition of success is fleet impact."

Outcomes Over Outputs: There is a strong demand to shift focus from process metrics to mission outcome metrics. Interviewee 23, an industry program manager highlighted the emptiness of on-time delivery without value: "you could meet deliverables on time, but what you gave me didn't move the ball forward…we look at it as success if…the knowledge and content you're giving us is helping…achieve the vision."

Requirements as the Cornerstone: Interviewees argued that successful outcomes are predetermined by the quality of foundational acquisition activities. Interviewee 1, a Senior Navy stakeholder, summarized the consensus: "(a) was the requirement written well, (b) did you hire the contractor that bid…and (c) did you administer the contract?" Conversely, poorly written requirements make measurement impossible. As Interviewee 14, a government acquisition professional noted: "if you want to say, was the performance good…you have to walk it back to why—was it clearly defined?"

b. Theme 2: Balanced, Multi-Perspective Health Scorecard

Core Idea: Healthy services require balance across cost, schedule, performance, customer value, and compliance.

No single perspective should dominate; stakeholders explicitly describe balance as the hallmark of health. Participants pointed to a modified balanced scorecard approach as the ideal state. Interviewee 24, an academic expert and former Air Force contracting officer summarized this view: "a healthy system [is] where everything balances well together…cost/schedule/performance…against customer experience…against compliance…if one…goes up too high or too low…we're not healthy."

The Consensus Framework: A clear vision for an ideal framework emerged, measuring performance across four or five key domains. Interviewee 17, an enterprise-level services portfolio/category manager explained why this balance is critical to avoid blind spots: "if you just look at cost, schedule, performance, it forgets about the end user…having something that balances that… [is vital]."

Avoiding Distortion: Interviewees noted that over-focusing on one area creates risks. For example, a contract might be "green" on cost but failing to deliver end-user value. A balanced approach ensures that "health" is a composite of efficiency and effectiveness.

This finding empirically validates the theoretical requirement for multidimensional balance established by Kaplan and Norton (1992). It confirms that the single-measure myopia described by Brubaker et al. (2018), where agencies over-rely on lead time, is actively hindering current DoD service assessment.

c.   *Theme 3: Service Quality and Reliability as Outcome Evidence*

Core Idea: Health shows up in reliable, available, "right-first-time" service delivery.

Participants repeatedly point to objective quality outcomes, such as uptime, availability, and error-free delivery as the clearest health signals. Interviewee 23, an industry program manager described the clarity of these metrics: "very clear metrics…'resolve a help desk ticket in no more than 24 hours…incident in 2 hours…uptime 99%'…those kind of things focus everybody to performance."

Objective vs. Subjective: While private industry relies on relationship satisfaction, DoD stakeholders viewed general "customer satisfaction" as difficult to rely on. As Interviewee 4, a government services program/portfolio manager noted: "people love customer satisfaction. People hate trying to measure customer satisfaction…because it's too subjective."

Specific Metrics Endorsed:

- *Zero Critical Failures*: widely praised for mission-critical contexts. Interviewee 3, a defense medical / laboratory services stakeholder,

explained: "some of these are tissue samples…you get no failures on that. …If your metrics don't equal this number…this is a fail."

- *First Pass Yield*: viewed as a definitive indicator of quality, specifically for avoiding operational delays. Interviewee 24, an academic expert and former Air Force contracting officer noted: "deliverables accepted on first submission without requiring rework…aircraft maintenance…we want the work being done right the first time, because…that sortie doesn't launch."

- *Skepticism of Compliance Percentages*: many expressed aversion to broad "compliance" percentages because they mask specific failures. Interviewee 22, an IT services / help-desk program manager argued: "you can close every help ticket within 36 hours…you didn't resolve any of them, but you closed them all…you're not measuring the important thing."

d. *Theme 4: Cost-to Value Stewardship (Not Spend Alone)*

Core Idea: Cost health is about value and efficiency, not just burn rate.

Participants want spend-plan variance plus attribution, ensuring the government is "right-sizing" what is bought. A common refrain regarding financial tracking was a focus on efficiency. Interviewee 15, a government program manager stated: "burn rate tells me that they're spending money. It doesn't tell me they're spending money efficiently… the amount of money they spent tells you very little about actual progress."

Context-Dependent Metrics: Interviewee 8, a Defense Contract Management Agency contracting leader noted that financial metrics must match the contract type: "burn rate…useful…in almost every major contract… [but] you do have to question… is the juice worth the squeeze?"

Value versus Savings: "Savings" alone was viewed with skepticism if it meant cutting necessary output. True stewardship was defined as getting the "A+ work" only when A+ work is required, avoiding "over-performance waste." Interviewee 8, a Defense Contract Management Agency contracting leader noted regarding labor contracts: "I was paying…for butts in seats, and I was not getting value because the butt was not in the seat."

e. *Theme 5: Leading-Indicator "Structure-Process-Outcome Chain" (Donabedian, 1988) Enables Outcomes*

Core Idea: Upstream, structural and procedural leading indicators are the necessary predictors of downstream, lagging mission results.

This theme mirrors the logic of the Donabedian (1988) Model, a conceptual framework originally developed for evaluating quality in healthcare. The model posits that outcomes are not random; they are the direct result of structure which is the attributes of the setting and resources, and process—how the work is actually done. Applying this theory to the interview data reveals that practitioners view "Structure" and "Process" as the essential leading indicators required to predict contract success.

The Theoretical Shift: focusing upstream—outcomes do not appear by magic; they are the result of upstream health factors. By adopting a structure-process-outcome mindset, organizations can move away from purely retrospective reporting. Interviewee 1, a Senior Navy Stakeholder, emphasized the urgent need to shift focus from lagging results to the upstream components of the Donabedian (1988) chain: "We need the leading indicators…we're too hung up on measuring the lagging ones."

Structure: workforce stability—in the context of the Donabedian (1988) framework, structure refers to the capacity and stability of the resources performing the work. For professional services, human capital is the defining structural element.

Metric: Personnel Turnover & Fill Rate.

Significance: Without the "Structure" of a stable workforce, the "Process" cannot be executed effectively.

Interviewee 9, a County-level Procurement Manager and Former Air Force Contracting Officer, described workforce stability as a primary structural metric: "Number one was personnel turnover… [and] actual fill rate… because a lot of them were just 'butts-in-seats'… turnover and fill rate were very good [metrics]."

Process: Administrative Disciplines

Process represents the actual interaction between the provider and the government—specifically, how the contract is managed post-award. Active management serves as the mechanism that converts the "Structure" (the workforce) into the desired "Outcome."

Metric: Surveillance Plan Alignment.

Significance: Even with the right people (Structure), a failure in the management plan (Process) ensures failure in the mission.

Interviewee 19, a Department of Veterans Affairs Branch Chief identified the alignment of task and surveillance as the critical process factor: "Well defined task and a matching surveillance plan. If you don't have those two things, a service contract is likely to fail."

f.   Theme 6: Standardized Core-Plus-Tail Governance for Roll-Up

Core Idea: A standard enterprise core of outcome areas, with tailored KPIs by category/criticality.

This theme addresses the "central dilemma" identified in the interviews: the conflict between standardization and the bespoke nature of services. Interviewee 24, an academic expert and former Air Force contracting officer argued: "we buy so many different types of services. I don't see how you can get any standardized specific metrics… [except] very, very top-level metrics."

The "One-Size-Fits-All" Failure: Interviewees stressed that services range from "lawn mowing" to complex research and development. A rigid, standardized system would defeat the purpose of performance-based acquisition. Interviewee 23, an industry program manager noted: "Benchmarks…very different…facilities…housekeeping…transportation…category specific."

The Core-Plus-Tail Solution: To solve this, participants supported a "choose your own adventure" structure. The DoD establishes high-level outcome categories (the core), but programs select specific, relevant KPIs (the tail). Interviewee 17, an enterprise-level services portfolio/category manager described this architecture clearly: "you have to have it be more like areas, and somehow the individual KPIs roll up to those graded areas…everybody can individualize their own thing but still feed the larger machine."

g.   Theme 7: Cultural/Data Distortions Threaten Outcome Measurement

Core Idea: Even good KPIs fail if culture, incentives, and data realities distort them.

This theme captures the systemic barriers that threaten the validity of the measurement framework. A major concern is that current metrics are often driven by complaints rather than data. Interviewee 3, a government services program/portfolio manager, noted: "The main metric that is utilized is complaining. It's not based on quantifiable, measurable PWS metrics…very personal-based."

The Burden of Data: A powerful consensus exists that the "juice must be worth the squeeze." Manual data calls are seen as a major barrier. Interviewee 10, a senior industry contracts and operations leader stated: "I can measure anything... It just becomes an issue of, is it value added, or is it just a data junkie?... Are you actually using those metrics for something to make a decision on?"

Risk Aversion and Protests: The "fear of protest" drives a risk-averse culture. Interviewee 1, a Senior Navy stakeholder, noted that measuring protests themselves can be counterproductive: "Be really careful…just measuring quantity of protests…the risk is drive the risk-averse behavior even worse."

Leadership Top Cover: A cultural theme was the need for leadership to provide "top cover" for acquisition teams, allowing them the time to perform thorough upfront work rather than rushing to award.

3.   Rationale for Selected Performance Categories

The selection of cost, schedule, performance, customer satisfaction, and regulatory compliance as the five pillars of this framework was not arbitrary; it represents a deliberate synthesis of traditional acquisition doctrine and the specific health requirements identified by stakeholders. As noted by Kerzner (2017), while the traditional "iron triangle" of cost, schedule, and performance remains the bedrock of program management, the analysis revealed that cost, schedule, and performance alone offer an incomplete picture of service health. As noted in Theme 2, stakeholders consistently argued that a contract could be "green" on cost and schedule yet still fail to deliver mission value if the end-user experience was poor. Therefore, customer satisfaction was elevated to a primary category to serve as a counterbalance to efficiency

metrics, ensuring that the quality of the experience is weighed equally against the efficiency of the transaction as noted in Theme 2.

Furthermore, the addition of regulatory compliance as a distinct category addresses a unique constraint of the defense environment identified in Theme 5 (structure-process-outcome chain). Unlike the private sector, where agility often trumps procedure, DoD services are bound by non-negotiable statutory and security requirements. A contractor that delivers on time but fails to maintain security clearances or safety certifications presents an unacceptable risk to the government. By isolating compliance as its own category, the framework distinguishes between "mission performance" (i.e., doing the job well) and "legal adherence" (i.e., being qualified to do the job), preventing critical compliance failures from being buried under general performance ratings.

Finally, this five-category structure aligns the DoD's service acquisition approach with the BSC theory explored in Chapter III. Just as Kaplan and Norton (1992) argued that financial metrics must be balanced with internal process and learning perspectives, this framework ensures that no single dimension dominates the assessment. By measuring these five areas in concert, the framework prevents the "gaming" behavior identified in Theme 7, where vendors might sacrifice quality to meet a schedule target or cut corners on compliance to lower costs. This holistic structure forces a tension between the metrics, ensuring that "service health" is defined by equilibrium across all five domains rather than optimization of just one.

4.   Selected Key Performance Indicators

The analytical process of classifying over 300 candidate KPIs and validating them against stakeholder feedback resulted in a refined list of 21 high-value metrics. This validation process revealed a critical finding (Theme 6): stakeholders do not believe a single, universal KPI set can effectively measure the entire diversity of DoD services. Participants consistently emphasized that a one-size-fits-all approach risks creating meaningless data or incentivizing compliance over mission results.

However, to address the research question and enable enterprise visibility, the analysis identified a foundational core set of seven metrics. These metrics represent the "common language" of service health—universal enough to apply across most portfolios while remaining flexible enough to be tailored at the contract level. They are not intended to be the only measures used, but rather the standardized baseline that feeds into the broader Core-Plus-Tail framework proposed in Chapter V.

a. *Refined List of Validated KPIs*

Table 3 presents the 21 validated metrics, with seven core metrics highlighted as the foundational set for enterprise reporting.

Table 3.   Refined List of 21 Validated KPIs

KPI Number	Measure	Target
Cost Category		
1	Percentage of Contract spend obligated vs. Planned	95–105% alignment between planned and actual obligations.
Schedule Category		
2	Percentage of deliverables submitted early or ahead of schedule.	≥ 15% of total deliverables.
3	Percentage of multi-phase contracts where phase transitions occur on or before the planned date.	≥ 90% on-time phase transition.
4	Percentage of contracts meeting on-time delivery for urgent or high-priority tasks.	≥ 98% on-time delivery for urgent requirements.
5	Number of schedule re-baselining events per contract per year (excluding government-directed scope changes).	≤ 1 significant re-baseline event per contract year.
6	Percentage of contracts meeting key milestone dates (e.g., initial delivery, service start, projected phase completion)	≥ 90% milestone adherence
Performance Category		

7	Percentage of contracts meeting minimum acceptable technical performance thresholds per contract requirements.	≥ 95% compliance.
8	Percentage of contractor corrective actions closed on time after government-identified nonconformances.	≤ 10% critical role turnover annually.
9	Number of critical performance failures or service outages (defined by severity in the contract or QASP) per contract.	Zero critical failures; or ≤ X per quarter based on service type.
10	Percentage of government-identified non-conformances or deficiencies corrected within the required timeframe	≥ 95% corrected on time
11	Percentage of contract deliverable accepted on first submission without requiring rework	≥ 90% first-pass acceptance
12	Percentage of contracts meeting agreed-upon key performance indicators (KPIs) or service-level agreements (SLAs).	≥95% compliance
13	Percentage of contract-related issues (e.g., disputes, performance deficiencies) resolved within agreed timeframes	≥ 95% on-time resolution
14	Percentage of critical performance metrics and service levels, as defined within the specific service contract's Performance Work Statement (PWS), that are met or exceeded by the contractor.	≥95% of all defined critical SLAs met per reporting period.
15	Percentage of service-related issues/problems reported by users that are resolved by the contractor within the contractually agreed-upon timeframe and to the user's satisfaction	≥ 95% resolved within SLA timeframes; ≥ 90% satisfaction with resolution.
Customer Satisfaction Category		
16	Percentage of contractor personnel turnover impacting contract functions	≤ 10% critical role turnover annually
17	Percentage of Customer-reported issues or complaints resolved within agreed service levels	≥95% resolution within SLA
18	Percentage of stakeholder satisfaction ratings across all service portfolios	≥ 90% satisfaction ration

19	Average satisfaction rating (e.g., on a 1–5 or 1–10 scale) from verified end-users of the service regarding its quality, timeliness, and contribution to their tasks.	Average score of ≥ 4.0 out of 5.0 (or equivalent).
Regulatory Compliance Category		
20	Percentage of contractor personnel maintaining required certifications, licenses, or clearance	100% compliance across all service categories
21	Percentage of identified critical contract risks with documented, active, and effective mitigation plans (as assessed by progress against plan and risk reduction)	100% of critical risks have effective mitigation plans; ≥ 90% of mitigation actions on tract

b. Rationale for the 7 Core Selected Metrics

These seven metrics were selected because they bridge the gap between standardization and customization. They reflect a balanced view of performance (Theme 2) and incorporate both lagging outcomes and leading indicators (Theme 5) to predict health rather than just report history.

- Cost Category (KPI #1): Percentage of Contract Spend Obligated versus Planned
  - *Rationale:* While participants rejected "burn rate" as a standalone measure of success, this metric serves as a vital indicator of Financial Stewardship (Theme 4). It aligns with the requirement for fiscal accountability without incentivizing "cheapness" over quality. By tracking the variance between planned and actual obligations, this KPI provides a standardized view of financial predictability across the enterprise.
- Schedule Category (KPI #6): Percentage of Contracts Meeting Key Milestone Dates
  - *Rationale:* This metric operationalizes Mission-Timed Delivery (Theme 1). Unlike generic "on-time" metrics that can be gamed by delivering non-critical items early, this KPI focuses on the specific milestones that drive mission readiness. It ensures a direct link between the acquisition schedule and the operational capabilities required by the warfighter.
- Performance Category (KPI #9): Number of Critical Performance Failures or Service Outages
  - *Rationale:* Directly addressing Theme 3 (Service Quality as Outcome Evidence), this metric was identified by stakeholders as the "gold standard" for no-fail missions (e.g., medical, IT, safety).

It creates a definitive, non-negotiable floor for performance health that is far more actionable than generic compliance percentages.

- Performance Category (KPI #10): Gov-Identified Non-conformances Corrected Within Timeframe
  - *Rationale:* This metric balances accountability with responsiveness. It acknowledges that issues will arise but measures the contractor's recovery speed and effectiveness. This creates an objective proxy for "responsiveness" that is less subjective than satisfaction surveys, addressing the "squishy metric" concerns raised in Theme 3.

- Customer Satisfaction Category (KPI #16): Contractor Personnel Turnover Impacting Contract Functions
  - *Rationale:* Identified in Theme 5 (Structure-Process-Outcome Chain) as a primary Leading Indicator, this metric predicts future performance risks before they manifest as mission failures. High turnover in critical roles is often the first signal of instability, making this metric essential for proactive management, particularly in knowledge-based services.

- Customer Satisfaction Category (KPI #19): Average Satisfaction Rating from Verified End-Users
  - *Rationale:* Despite skepticism about the subjectivity of satisfaction ratings (Theme 7), a balanced scorecard (Theme 2) requires a customer perspective. This metric is included as a necessary counterweight to cost and schedule efficiency. To mitigate validity concerns, it is scoped to "verified end-users" to ensure feedback comes from those directly impacting the mission, rather than administrative staff.

- Regulatory Compliance Category (KPI #20): Percentage Maintaining Required Certifications
  - *Rationale:* This metric serves as both a Regulatory Floor and a Leading Indicator of capability (Theme 5). 100% compliance is essential not just for legality, but because a workforce lacking required certifications cannot legally perform the mission. It provides a binary, indisputable signal of "service health."

These seven metrics serve as the practical starting point for the standardized, outcome-oriented performance framework. They enable the DoD to aggregate health signals across the enterprise while leaving room for the "tail" metrics necessary for specific service categories.

D. SUMMARY

This chapter detailed the research methodology, analysis, and findings used to address the critical gap in standardized, outcome-based performance metrics for the DoD AoS Pathway. The study followed a qualitative, exploratory design to navigate the complexity of service acquisitions, utilizing a rigorous multi-phase approach that progressed from a broad literature review to the systematic validation of specific metrics. The analytical process combined comprehensive document review with the AI-assisted thematic analysis of 24 stakeholder interviews, ensuring that the resulting insights were both data-driven and grounded in practitioner reality.

The findings revealed a strong consensus that the current measurement system is overly process-focused and fragmented. Seven priority themes emerged from the analysis, establishing a new definition of "service health" anchored in mission contribution rather than administrative compliance. Stakeholders advocated for a balanced scorecard approach that integrates cost stewardship, schedule reliability, service quality, and customer satisfaction. A key analytical breakthrough was the identification of a causal structure-process-outcome chain, as described by Donabedian (1988), where leading indicators such as workforce stability and requirements clarity serve as early predictors of mission success. Furthermore, the data resolved the central dilemma of standardization versus customization by validating a Core-Plus-Tail governance model, which mandates a small set of enterprise metrics while allowing for category-specific tailoring.

Finally, the analysis produced a refined set of 21 high-value KPIs and identified seven foundational core metrics to serve as the baseline for enterprise reporting. These selected metrics address the critical domains of stewardship, readiness, and service quality while mitigating the risks of gaming and administrative burden identified during the interviews. Collectively, these findings and the validated KPI set provide the empirical foundation for the standardized performance measurement framework presented in Chapter V, which translates these insights into an actionable system for improving accountability and mission alignment across the DoD.

THIS PAGE INTENTIONALLY LEFT BLANK

V. CONCLUSIONS AND RECOMMENDATIONS

This final chapter synthesizes the study's findings to provide conclusions and actionable recommendations for the DoD. It directly addresses the core problem statement identified in Chapter I and by the OUSD(A&S) sponsor: the lack of standardized, outcome-oriented metrics to evaluate the health of service acquisitions.

The central contribution of this chapter is the Integrated Performance Ecosystem (IPE). This ecosystem is not a static list of metrics, but a dynamic system composed of three distinct, interdependent components derived from the systematic analysis:

1. **Governance (The "What"): Core-Plus-Tail.** The standardization logic that mandates what must be measured at the enterprise level while allowing flexibility for specific service categories.
2. **Structure (The "How"): Enterprise Performance Visibility (EPV).** The architectural framework that organizes those measures and defines how they are rolled up by echelon from the tactical contract level to the strategic top tier.
3. **Visualization (The "View"): The Flexible Balanced Scorecard (FBS).** The evaluation tool that allows leaders to see readiness and stewardship status at a glance.

Following the summary of findings and the direct answers to the research questions, this chapter details the IPE and offers actionable recommendations for implementation, addressing policy, technology, and cultural transformation.

A.   SUMMARY OF KEY FINDINGS

The recommendations in this chapter are grounded in the seven major thematic findings derived from the analysis in Chapter IV. These findings serve as the empirical evidence base for the framework:

- **Mission-Anchored Definition of Health (Theme 1):** Stakeholders redefined service health not as contract compliance, but as the service's measurable contribution to operational readiness.
- **Balanced, Multi-Perspective Scorecard (Theme 2):** There was overwhelming consensus that health cannot be assessed by a single metric. A balanced view across cost, schedule, performance, and customer satisfaction is required.

- **Service Quality and Reliability as Evidence (Theme 3):** Objective, right-first-time quality metrics were identified as the most credible evidence of outcome health.
- **Cost-to-Value Stewardship (Theme 4):** Stakeholders rejected simple burn rate metrics in favor of stewardship indicators that assess efficiency and right-sizing value.
- **Leading-Indicator "Structure-Process-Outcome Chain" (Theme 5):** Leading indicators—specifically requirements clarity and workforce stability—were found to be the strongest predictors of future service health.
- **Standardized Core-Plus-Tail Governance (Theme 6):** Participants validated a Core-Plus-Tail architecture: a mandated small core of enterprise metrics supplemented by flexible, category-specific tail metrics.
- **Cultural and Data Distortions (Theme 7):** The primary barriers to implementation are cultural risk aversion and administrative burden, requiring automated data collection.

B. RESEARCH QUESTIONS ANSWERED

This study set out to determine how the DoD can move beyond process compliance to measure the actual value of contracted services. The following section summarizes how the findings explicitly answer the research questions using the IPE components.

- **Primary Research Question:** What outcome-based performance measures should be used to assess the health of services?
- **Answer:** The study determined that a standardized Core 7 set of metrics is essential for enterprise governance. Based on the analysis, the specific measures that should be used are: The Core 7 Foundational Metrics. To ensure the FBS provides a consistent health grade across the enterprise (Tier 1), this study identified seven foundational metrics. These act as the core in the Core-Plus-Tail governance model (see Table 4).

Table 4.   Core 7 Metrics—Final KPI List

Category	KPI Name	Measure	Target	Rationale & Theoretical Link
Cost	**Cost Control**	% Spend Obligated vs. Planned	95–105%	Stewardship (Theme 4): Tracks financial predictability and discipline without incentivizing "cheapness" over value.

Schedule	Milestone Fidelity	% Contracts Meeting Key Milestones	≥ 90%	Mission-Timed Delivery (Theme 1): Focuses on specific milestones that drive mission readiness, avoiding the gaming of generic "on-time" metrics.
Performance	Critical Reliability	Critical Performance Failures	Zero	Reliability (Theme 3): Provides a definitive, non-negotiable floor for performance health in no-fail missions.
	Deficiency Correction	Non-conformances Corrected On-Time	≥ 95%	Responsiveness (Theme 3): An objective proxy for responsiveness that measures recovery speed.
Customer	Workforce Stability	Contractor Personnel Turnover	≤ 10%	SPO Leading Indicator (Theme 5): The primary predictor of future risk. High turnover in critical roles signals instability.
	End-User Satisfaction	Verified End-User Satisfaction	≥ 4.0/ 5.0	Stakeholder Value (Theme 2): Ensures a customer perspective is maintained via verified end-users.
Compliance	Credentialing	% Personnel Certified	100%	Capability (Theme 5): A binary leading indicator. A workforce lacking certifications cannot legally perform the mission.

Note. KPIs derived from the validated list of 21 as highlighted in Table 3.

While the research identified these metrics, the study also identified a "central dilemma" (Theme 6) regarding the inherent heterogeneity of service acquisitions. Specifically, a single "one-size-fits-all" list of metrics cannot cover the diversity of all DoD services (e.g., comparing janitorial services to cyber support) without becoming meaningless or overly burdensome. The solution is the Core-Plus-Tail governance model. This architecture mandates the Core 7 metrics for enterprise-level reporting (to ensure standardization and visibility) while explicitly allowing program managers to add flexible "Tail" metrics (to capture specific technical requirements and local context). These metrics are applied via the Core-Plus-Tail governance model, where they are mandatory

for reporting but supplemented by contract (Tier 3) and program (Tier 2) specific "tail" metrics.

Secondary Research Questions:

- **Secondary Question 1**. How can standardized outcome focused performance metrics improve assessment, oversight, and strategic decision-making across DoD service programs?

- **Answer:** Standardization improves assessment by creating a common language for "service health" through the Core-plus-tail governance model. The Enterprise Performance Visibility (EPV) structure improves oversight by enabling the rollup of data from disparate contracts into portfolio-level views. Strategic decision-making is improved by normalizing data (using the "Health Index" approach), which allows for "apples-to-apples" comparisons of stewardship and readiness contributions across different service portfolios.

- **Secondary Question 2.** What performance indicators used in service contracts can be adapted for DoD use?

- **Answer:** The literature and interviews identified SLAs, quality/reliability metrics, and customer satisfaction indexes as primary private-sector tools. The study adapted the private sector's focus on value co-creation into the Flexible Balanced Scorecard (FBS). Specifically, commercial indicators like "first pass yield" were adapted into defense-relevant metrics like "critical reliability."

- **Secondary Question 3.** What challenges exist for implementing standardized metrics, and how can they be mitigated?

- **Answer:** The primary challenges are administrative burden and cultural risk aversion. Stakeholders indicated that manual data collection leads to poor data quality and gaming. These challenges should be mitigated by automating data collection through authoritative systems and using Leading Indicators (Theme 5) to predict risks before they become failures.

C. THE INTEGRATED PERFORMANCE ECOSYSTEM

To resolve the fragmentation identified in the findings, this study proposes the Integrated Performance Ecosystem (IPE; see Figure 3). The IPE is the umbrella framework that unifies the "What," the "How," and the "View" of performance measurement.

The IPE functions through the integration of three specific components:

1. Governance (Core-Plus-Tail): Defines *what* is measured.
2. Structure (EPV): Defines *how* measures are organized and rolled up.

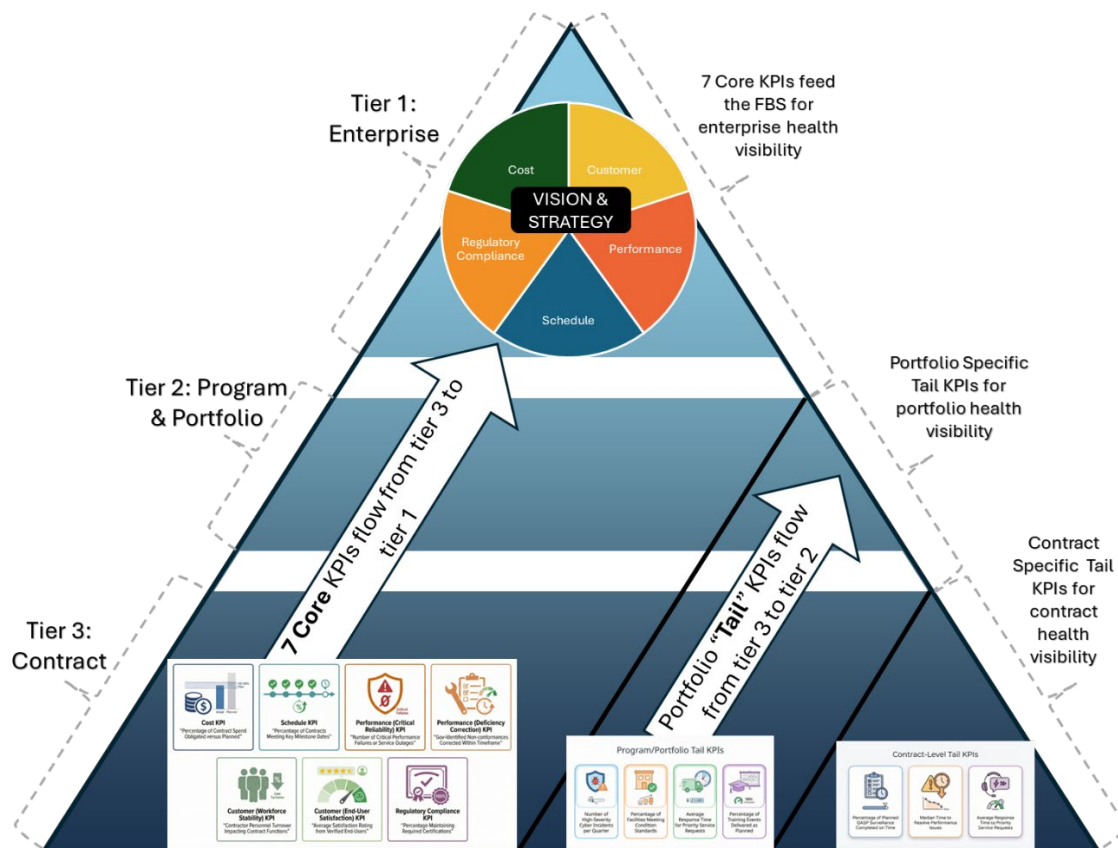3. Visualization (FBS): Defines how leaders *view* readiness and stewardship.

Figure 3. Integrated Performance Ecosystem Framework

D. GOVERNANCE: CORE-PLUS-TAIL

The foundation of the IPE is the Core-Plus-Tail governance model (see Figure 4). This component answers the question, "What must be measured?" It resolves the tension between standardization and customization identified in Theme 6 by creating a bi-level requirement.

Figure 4.   Core-Plus-Tail Model

The Core (Enterprise Mandate): A small set of 7 core KPIs (identified in Table 4) that are mandatory for all applicable contracts. This ensures that every service acquisition reports a consistent signal regarding Financial Stewardship, Schedule Adherence, Performance and Regulatory Compliance.

The Tail (Stakeholder-Owned Outcomes): Category-specific metrics defined by the portfolio managers and local requiring activity. These allow portfolio and managers to track unique technical requirements (e.g., "lines of code" for IT or "acres mowed" for facilities) without forcing irrelevant data up to the enterprise level.

Core-Plus-Tail Standardization: To operationalize the ecosystem effectively, the IPE framework relies on a specific governance architecture known as Core-Plus-Tail Standardization. This approach addresses the inherent heterogeneity of service acquisitions by balancing the need for enterprise visibility with the necessity of local context. Core-Plus-Tail Standardization is a governance architecture where the DoD adopts a small, enterprise-standard "Core" of outcome areas and KPIs that are roll-up safe, while explicitly allowing for category- and mission-criticality-tailored "Tail" KPIs at the contract level.

As a theoretical link, this architecture aligns with the performance prism model (Neely et al., 2002), which emphasizes stakeholder and context sensitivity, and the balanced scorecard (Kaplan & Norton, 1996), which provides standard perspectives.

E. STRUCTURE: ENTERPRISE PERFORMANCE VISIBILITY

The EPV framework (see Figure 5) is the structural "plumbing" of the ecosystem. This component answers the question: "How are measures organized and rolled up by echelon?." The EPV is the structural framework developed to resolve the central dilemma identified in Theme 6: the conflict between the need for enterprise standardization and the reality of diverse service categories. EPV provides the logic and plumbing that allows data to flow from a specific contract up to the FBS senior leadership dashboard. It relies on a three-tier implementation architecture to support the rollup of KPIs.
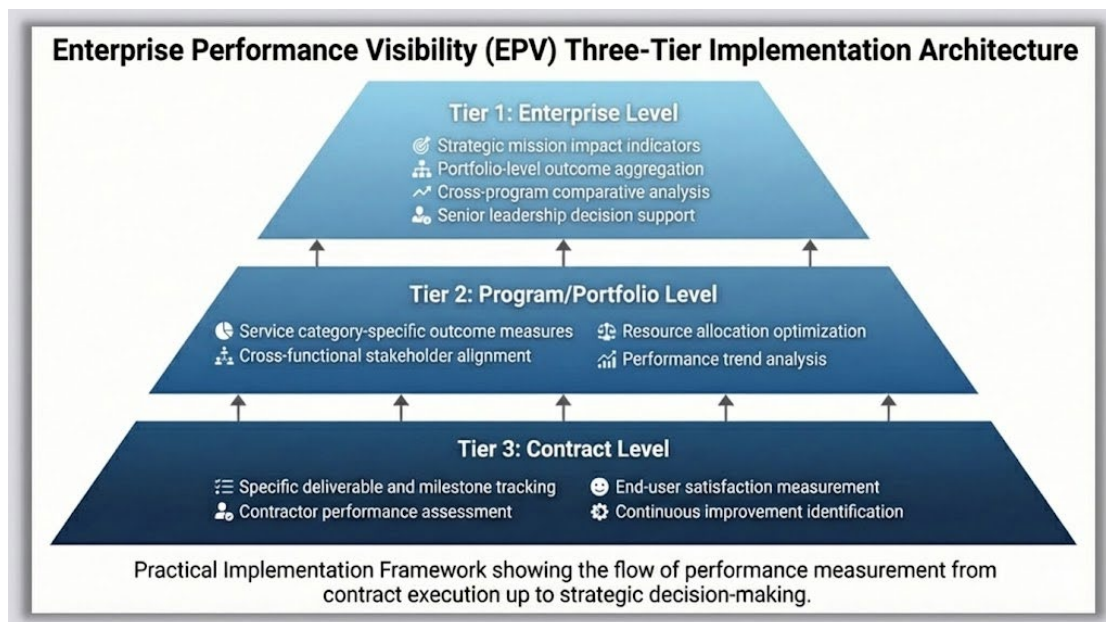
Figure 5.   Enterprise Performance Visibility Model

The Three-Tier Implementation Architecture: To move from theoretical models to a working system, EPV organizes data flow up through three distinct tiers of execution.

Tier 3: The Contract Level (Where Data is Born)

- *Who:* CORs, End Users, Contracting Officers, Contractor Leads.
- *Function:* This is the tactical layer where specific deliverables and outcomes are generated.
- *The Rollup Logic:* Practitioners define 3–7 outcome measures per contract. These measures capture specific Leading (e.g., staffing fill rate) and Lagging (e.g., service levels met) indicators.

- *Output:* A contract-level health assessment that feeds into the specific service category.

Tier 2: The Program/Portfolio Level (Where Meaning is Assembled)

- *Who:* PEOs, Portfolio Managers, Category Managers
- *Function:* This tier operationalizes the "adaptive contextualization" concept. It standardizes the categories of measurement (e.g., Reliability, Stewardship) rather than every single metric.
- *The Rollup Logic:* Aggregates contract-level data into portfolio dashboards. It allows for cross-program comparison, for example, comparing the "cost-to-outcome" efficiency of two different logistics contracts.
- *Output:* A portfolio health signal (Red/Amber/Green) that identifies systemic trends, such as regional labor shortages or supply chain gaps.

Tier 1: The Enterprise Level (Where Decisions Are Made)

- *Who:* OUSD(A&S), Service HQ, Senior Governance Boards.
- *Function:* This is the strategic layer where the FBS sits.
- *The Rollup Logic:* This tier does not ingest every raw data point. Instead, it aggregates the "health signals" from Tier 2 into a strategic view. It focuses on a small set of core indicators (see Table 4) to steer resources and develop reports.
- *Output:* An enterprise-wide view of service health, answering the strategic question, "are our service acquisition contributing to warfighter readiness?"

F. THE FLEXIBLE BALANCED SCORECARD: ASSESSING SERVICE HEALTH

While EPV provides the framework for data rollup, the FBS is the tool used to visualize and evaluate that data. It is the scorecard that displays the health of the service (see Figure 6).
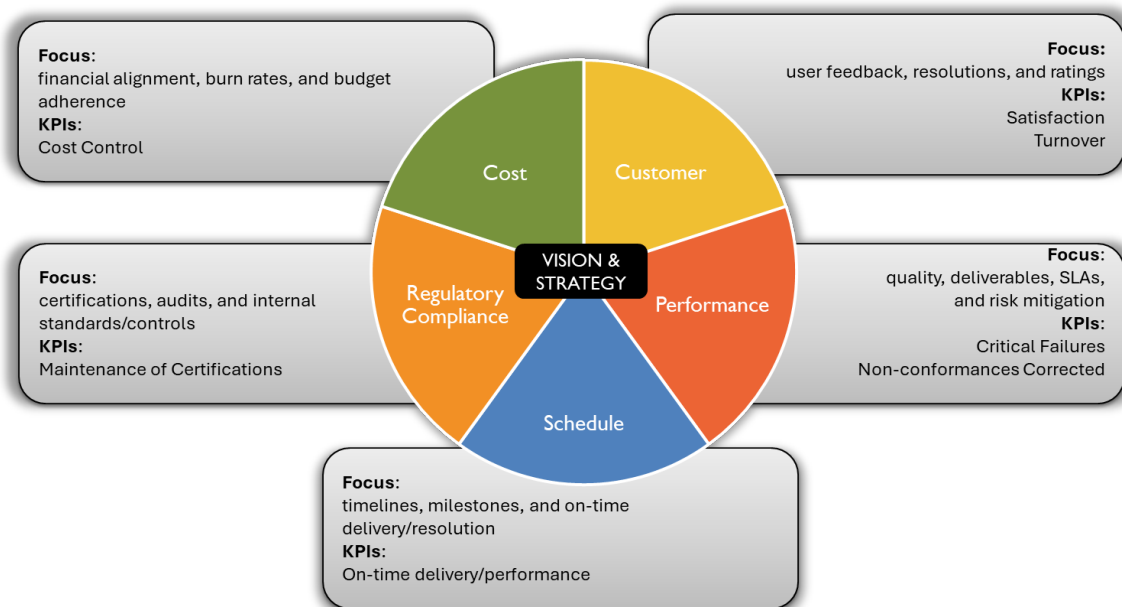
Figure 6. Flexible Balanced Scorecard Architecture. Source: Adapted from Kaplan (2010).

The FBS is designed to ensure the framework serves the multiple principals environment of the DoD. Unlike private-sector models driven by profit, the FBS which is adapted from Kaplan's (2010) public-sector balanced scorecard logic, expands the model to reflect how defense contracts are actually managed. It organizes "health" into five balanced perspectives:

- **Vision & Strategy (The Hub):** The central "North Star" is Warfighter Readiness. All other metrics are subordinate to this.
  - **Cost:** focuses on financial alignment and stewardship (not just burn rate)
  - **Schedule:** focuses on mission-timed delivery and milestone adherence
  - **Performance:** focuses on quality, reliability, and risk mitigation
  - **Customer:** focuses on end-user feedback and satisfaction
  - **Regulatory Compliance:** focuses on certifications, audits, and statutory requirements

G. INTEGRATED PERFORMANCE ECOSYSTEM IMPLEMENTATION

Concrete Vignette (Base Operations Support): To illustrate how this ecosystem functions in a routine service context (Tier 3 to Tier 1), consider a Base Grounds Maintenance contract (Landscaping). This vignette demonstrates the roll-up of Core-

Plus-Tail KPIs, showing how specific technical metrics at the tactical level inform strategic decisions.

Tier 3: The Contract Level (Base Landscaping)

- *The "Tail" (Category Specific):*
  - **BASH Compliance:** Percentage of airfield grass maintained below 7 inches (Bird Aircraft Strike Hazard).
  - **Storm Response:** Debris clearance from primary roads within 4 hours of storm cessation.
- *The "Core" (Enterprise Mandated):*
  - **Critical Reliability:** Zero obstructions to perimeter fence line visibility (Force Protection).
  - **End-User Satisfaction:** Verified rating > 4.0 from Base Housing Manager.
- *Result:* The contract scorecard mixes local needs (grass height) with enterprise imperatives (security/satisfaction).

Tier 2: The Portfolio Level (AFCEC Regional Command)

- *Tail Analysis:* "80% of bases in the Southeast Region are failing the 'Storm Response' Tail KPI" (Identifies a regional equipment shortage).
- *Core Analysis:* "Region-wide 'End-User Satisfaction' is 4.2 (Green), but 'Critical Reliability' is Red due to fence line obstructions at 3 bases."
- *Result:* The Portfolio Manager spots systemic issues that a single contract view would miss.

Tier 1: The Enterprise Level (HAF/A4)

- *Strategic View:* The "grass height" data stays at the tactical level, but its *impact* rolls up. The dashboard shows the Installation Support Portfolio is Green for Stewardship (Cost) but Amber for Readiness (Critical Reliability).
- *Decision:* Reallocate funds from routine aesthetic mowing (adjusting the Tail standard) to fence line security clearing (fixing the Core failure) to restore Readiness without increasing the budget.

H. DATA NORMALIZATION AND AGGREGATION LOGIC

A critical challenge identified during the study was the difficulty of comparing diverse services—the "apples-to-apples" problem (e.g., comparing a janitorial contract to a cyber warfare support contract). To resolve this, the EPV framework should employ index-based normalization, criticality weighting, and historical baselines to ensure valid aggregation at Tier 2 and Tier 1.

1. Normalization (The Health Index)

Adapting the aggregation logic from Franceschini et al. (2007)**,** the framework does not roll up raw data units (e.g., "hours worked" or "acres mowed"). Instead, it normalizes each metric into a percentage of variance from target. For example, if contract A has a target of 99% uptime, and achieves 98%, and contract B has a target of 5-day turnaround and achieves 6 days, both are converted to a standardized "health score" (e.g., 0–100 scale). This allows the Enterprise dashboard to aggregate the health of the contracts, rather than disparate data points.

2. Weighting Logic: Dollars versus Criticality

Aligning with Bititci et al.'s (2012) framework on the evolution of performance measurement from output-based to strategic outcome-based systems, the rollup logic applies a weighted value to the aggregation based on strategic alignment.

- *Mission Criticality (The Preferred Readiness Weight)*: For example, contracts designated as "Tier 1: No-Fail Mission Essential" are weighted heavier in the Readiness/Performance score than "Tier 3: Routine Commercial" services. This would ensure that a "Green" rating on a high-dollar landscaping contract does not mask a "Red" rating on a low-cost but critical medical support contract.
- *Financial Value (The Stewardship Weight)*: Conversely, for example, Contract Value ($) is used to weight the *cost* perspective. This ensures that a cost overrun on a billion-dollar program alerts leadership faster than a variance on a simplified acquisition, preserving financial stewardship.

3. Baseline Development (the Context Fix)

A static number (e.g., "92% Customer Satisfaction") is meaningless without context. The EPV framework mandates that targets be developed from historical baselines rather than arbitrary goals. This aligns with Behn's (2003) guidance that public sector performance measurement must be used for *learning* (trend analysis) rather than just *evaluation* (snapshots).

- *Trend Analysis:* By benchmarking current performance against a historical baseline (e.g., "The 3-year trailing average"), the framework can identify trends. This distinguishes between a "Red" program that is improving and a "Green" program that is slowly degrading, allowing leadership to intervene based on trajectory rather than just status.

4. Validity Guardrails

A scorecard is only as good as the trust in its data. To address Theme 7 (Cultural/ Data Distortions), the FBS operates under specific credibility conditions:

- **Clear Math:** All metrics must have mathematically precise denominators.
- **Balanced Sets:** Leading indicators (e.g., Turnover) must be paired with lagging outcomes (e.g., Critical Failures) to prevent gaming.
- **Automated Collection:** Data should be pulled from authoritative systems (e.g., FPDS, CPARS, or ADVANA) to minimize administrative burden.

I. ACTIONABLE RECOMMENDATIONS

The following recommendations provide a concrete path for implementing the EPV framework and the FBS scorecard.

1. Recommendation 1: Pilot the EPV Three-Tier Model

Do no attempt to "boil the ocean." The recommended pilot approach is:

- a). Pick 1–2 mission-critical service programs.
- b). Define the common EPV categories and starter KPI library.
- c). Run Tier-3 contract scorecards for two quarters.
- d). Stand up Tier-2 rollups and trend briefs
- e). Only after that, lock in Tier-1 enterprise indicators using FBS.

2. Recommendation 2: Update Policy to Mandate Core-Plus-Tail

The researchers recommend updating the DoDI 5000.74, *Defense Acquisition of Services,* to formally adopt the Core-Plus-Tail governance model embedded in EPV. This directly addresses the policy gap identified in Chapter II, where Section 4.5 of DoDI 5000.74 promotes performance measurement "to the maximum extent practicable" but fails to provide the standardized definitions necessary for implementation. Policy should mandate the collection of the Core 7 metrics for all service acquisitions above the simplified acquisition threshold to ensure enterprise visibility via the FBS, while granting program managers the authority to select tail metrics.

3. Recommendation 3: Automate the "Structure-Process-Outcome Chain" in Data Infrastructure

To address the "Burden of Data," the DoD must invest in integrating existing tools (e.g., ADVANA) to automatically track the structure-process-outcome Leading Indicators. Specifically, systems should be configured to flag "Workforce Instability" (e.g., Turnover) and "Schedule Volatility" (e.g., Re-baselining events) as automatic early warnings on the FBS dashboard.

4. Recommendation 4: Workforce Development on Pre-Award Discipline

The findings confirmed that metrics are a result of good requirements. Workforce training must pivot from "how to measure" to "how to write measurable requirements." This recommendation targets the critical workforce gap identified by Finkenstadt and Peterson (2011) and the 2024 AoS Panel Review (OUSD[A&S]), which found that insufficient analytic capacity prevents the effective use of data even when systems are available. Training should focus on the structure-process-outcome chain, teaching acquisition teams how to define the "structure" (i.e., Qualifications) and "process" (i.e., Surveillance) that lead to the desired "outcomes" displayed on the FBS, through the EPV and utilizing Core-Plus-Tail.

J. STUDY LIMITATIONS

This study's findings are based on a qualitative sample of 24+ interviews. While rich in detail, they are not statistically generalizable to the entire DoD acquisition workforce. The IPE framework is analytically derived from literature and expert consensus; it has not yet been empirically tested in a live acquisition environment. Its success hinges on the quality of underlying data in systems like ADVANA, CPARS, and FPDS.

K. RECOMMENDATIONS FOR FUTURE RESEARCH

Based on the limitations and the findings of this study, the following areas are recommended for future research to further advance the measurement of service acquisitions:

- **Quantitative Validation**: Conduct a statistical analysis linking the Core 7 KPIs to actual readiness ratings (e.g., DRRS) to empirically validate the causal link between service health and mission readiness.
- **Weighting Logic Validation**: Future research should empirically test different aggregation methods, specifically comparing Dollar-Weighted versus Criticality-Weighted rollups to determine which method provides the most accurate predictive signal for operational readiness.
- **Flex Library Development**: Research and develop the comprehensive "Tail" library of metrics for specific service taxonomies (e.g., Knowledge-Based Services, R&D, Construction) to populate Layer 2 of the framework.
- **Incentive Structure Analysis**: Investigate how different contract types (Fixed Price versus Cost Plus) influence the effectiveness of the FBS framework and the behavior of contractors regarding the Core KPIs.

L.   CONCLUSION

This research began by identifying a practical problem which is the absence of a consistent framework for assessing service health. The EPV solves this by providing the necessary architecture for data rollup, while the FBS provides the visual language of success. By adopting this dual approach, i.e., an ecosystem for structure and a scorecard for health, the DoD moves beyond the false choice of "standardization versus customization." It re-anchors the entire acquisition process to the only metric that truly matters: the readiness and lethality of the force.

"If you can't measure it, you can't change it." — Peter Drucker

"If you can not measure it, you can not improve it." — Lord Kelvin

"In business, the idea of measuring what you are doing, picking the measurements that count like customer satisfaction and performance… you thrive on that." — Bill Gates

GENERATIVE AI DISCLOSURE STATEMENT

Generative artificial intelligence tools were used in preparing this work. Specifically, AI was employed for the following purposes: improving sentence clarity, grammar, formatting, organization of tables and assistance in conducting systematic thematic analysis of transcript data. The use of these tools was approved according to institutional and faculty guidance, in alignment with NPS policy on generative AI use. ChatGPT, Grok, Gemini and Claude Sonnet 4 within NIPRGPT were the AI models used.

All recommendations and suggestions provided by the AI were reviewed and applied at the author's discretion, and final editorial decisions were made by the authors to ensure the accuracy, integrity, and originality of the content. All ideas, arguments, interpretations, and analyses presented in this work are entirely the author's own. Efforts were made to mitigate any risks related to AI-generated inaccuracies, over-reliance on automated recommendations, or misuse of field-specific terminology by verifying changes and consulting with academic resources when appropriate. All AI-generated outputs were independently verified for factual accuracy before inclusion.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX. INTERVIEW QUESTIONAIRE

Section 1: Background

Question 1: Could you describe your current role and your experience with service acquisition?

Follow-up prompts:

(1) How long have you been involved in service acquisition processes?

(2) Does your role primarily focus on outcome-based or process-based acquisition metrics?

Section 2: Agency Level Questions

Question 2: Does your agency currently have a standardized process for determining/ tracking the success of services put on contract?

If not – Follow-up prompt:

(1) Are there any informal processes or guidelines that different departments or programs use?

If yes – Follow-up prompts:

(1) What specific metrics or KPIs are required for each program to determine a contractor's success? How were these metrics chosen?

(2) How is this data reported and tracked through your agency? Who receives these reports and is responsible for reviewing the performance data of service contracts at the agency level?

(3) What actions are taken if the services don't meet expectations?

Section 3: Requirement Owner Questions

Question 3: What are the most critical factors that determine the success of a service contract?

Question 4: How do you currently measure the performance of service contracts within your program? What requirements do you levy in the PWS or data do you collect?

Question 5: Are the performance metrics you use specific to your program, or are they aligned with any agency-wide standards? Do you face challenges when measuring the success of service contracts in your program?

Question 6: What would an ideal set of metrics look like for determining the success of a service contract in your program, regardless of current limitations?

Section 4: Industry Representatives

Question 7: In your experience, what are the most effective metrics or KPIs for measuring the success of service contracts that you've had? (responses may be specific to individual functional areas)

Question 8: What types of data are typically collected to track your contractor's performance?

Question 9: Are there any industry-standard metrics or benchmarks that are commonly used to evaluate the success of service contracts?

Question 10: Are you aware of any private-sector methods or KPIs that could be adopted to improve the DoD's service acquisition performance measures?

Section 5: Closing & Additional Insights

Question 11: Are there any additional insights or perspectives you think are crucial for understanding how to improve outcome-based service acquisition metrics in the DoD?
    Follow-up prompts:
        (1) Is there anything we haven't discussed that you believe is essential?

        (2) Would you be willing to review our preliminary findings or be contacted for clarification if needed?

SUPPLEMENTAL: KPI UNIVERSE

This supplemental provides the full universe of 338 performance indicators compiled during Phase 1 of the research. The KPIs in the file are classified according to BSC categories and indicator type as well as other relevant information. Because the complete dataset exceeds the size limits for inclusion as an appendix, it is provided as a separate supplemental item to ensure transparency and replicability.

To access the supplemental material(s) listed here, contact the Dudley Knox Library or, for publicly releasable theses and supplementals only, visit the thesis pages in the library's Calhoun database.

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Akkermans, H., van Oppen, W., Wynstra, F., & Voss, C. (2019). Contracting outsourced services with collaborative key performance indicators. *Journal of Operations Management, 65*(1), 22–47. https://doi.org/10.1002/joom.1002

Alqahtani, F., Selviaridis, K., & Stevenson, M. (2023). The effectiveness of performance-based contracting in the defence sector: A systematic literature review. *Journal of Purchasing & Supply Management, 29*, 100877. https://doi.org/10.1016/j.pursup.2023.100877

American Customer Satisfaction Index. (2024, November 12). *ACSI® federal government study 2024: Citizen satisfaction with federal government services reaches 7-year high*. American Customer Satisfaction Index LLC. https://www.theacsi.org

Anton, P. S., Shelton, W., Ryseff, J., & Cohen, S. (2019). *The U.S. Air Force's ability to monitor contractor performance during contingency operations* (Report No. RR-3136). RAND. https://www.rand.org/content/dam/rand/pubs/research_reports/RR3100/RR3136/RAND_RR3136.pdf

Anton, P. S., Shelton, W., Ryseff, J., Cohen, S., Johnson, G., Joplin, S. B. … Vigo Camargo, A. (2022). *Early predictive indicators of contractor performance: A data-analytic approach* (Report No. RRA542-1). RAND. https://www.rand.org/pubs/research_reports/RRA542-1.html

Axelos. (2019). *ITIL foundation: ITIL 4 edition.* The Stationery Office.

Baker, J. T., Bono, M. J., & DeVoe, J. T. (2016). *LPTA versus tradeoff: Analysis of contract source selection strategies and performance outcomes* (Master's thesis, Naval Postgraduate School). NPS Archive: Calhoun. https://hdl.handle.net/10945/49498

Baldwin, L. H., Camm, F., & Moore, N. Y. (2000). *Strategic sourcing: Measuring and managing performance* (Report No. DB-287-AF). RAND. https://www.rand.org/pubs/documented_briefings/DB287.html

Baldwin, L.H., & Hunter, S. B., (2004). *Defining needs and managing performance of installation support contracts: Perspectives from the commercial sector* (Report No. MR-1812-OSD). RAND. https://www.rand.org/pubs/monograph_reports/MR1812.html

Barnd, D. J. (1989). *An evaluation of factors that influence service contract quality* (Master's thesis, Naval Postgraduate School). NPS Archive: Calhoun. https://calhoun.nps.edu/server/api/core/bitstreams/a13ccc20-110b-4c1d-ad84-8c1e56d9363c/content

Behn, R. D. (2003). Why measure performance? Different purposes require different measures. *Public Administration Review, 63*(5), 586–606. https://doi.org/10.1111/1540-6210.00322

Belotserkovskiy, R., Sewak, J., Teodorian, A., & Lietke, B. (2018, May 2). *Contracting for performance: Unlocking additional value.* McKinsey & Company. https://www.mckinsey.com/capabilities/operations/our-insights/contracting-for-performance-unlocking-additional-value

Bevan, G., & Hood, C. (2006). What's measured is what matters: Targets and gaming in the English public health care system. *Public Administration, 84*(3), 517–538. https://doi.org/10.1111/j.1467-9299.2006.00600.x

Bititci, U., Garengo, P., Dörfler, V., & Nudurupati, S. (2012). Performance measurement: Challenges for tomorrow. *International Journal of Management Reviews*, 14(3), 305–327.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brubaker, S. S., Dean, N., & Posey, D. L. (2018). *Analysis of performance metrics used in contracting agencies* (Master's thesis, Naval Postgraduate School). NPS Archive: Calhoun. https://hdl.handle.net/10945/60367

Bouckaert, G., & Halligan, J. (2008). *Managing performance: International comparisons*. Routledge.

Bullard, E. (2024). *Purposive sampling*. EBSCO Research Starters. https://www.ebsco.com/research-starters/social-sciences-and-humanities/purposive-sampling

Casey, W., Peck, W., Webb, N. J., & Quast, P. (2008). Are we driving strategic results or metric mania? Evaluating performance in the public sector. *International Public Management Review, 9*(2), 90–106. https://hdl.handle.net/10945/39496

Cavanagh, J., Lloyd, R., Logan, S., Sade, M., Schoenberg, A., Sochon, G., & Wheeler, E. (1999). *The balanced scorecard for managing procurement performance. Contract Management, 39*(12), 12–16.

Cooper, L. (2022, March 9). *ISO/IEC 20000: Project managers and service managers—enemies or close friends?* [Presentation slides]. BCS Spring School, PROMS-G Project Management Specialist Group & Service Management & IT Asset Management. https://www.bcs.org/media/8614/promsg-springschool-lcooper-090322.pdf

Cotter, P. F. (2000). *Design services contracting for the federal government* (Master's thesis, Naval Postgraduate School). NPS Archive: Calhoun. https://hdl.handle.net/10945/7671

Cullen, S. (2016). *The contract scorecard: Successful outsourcing by design.* Routledge.

Dean, A. M., & Kiu, C. (2002). Performance monitoring and quality outcomes in contracted services. *International Journal of Quality & Reliability Management*, *19*(4), 396–413. https://doi.org/10.1108/02656710210421571

Defense Acquisition University. (n.d.). *Acquisition of services pathway.* https://aaf.dau.edu/aaf/services/

Defense Business Board. (2022). *Recommendations for the next generation of business health metrics* (Report No. DBB FY23-01). Department of Defense. https://dbb.defense.gov/Portals/35/Documents/Reports/2023/DBB%20FY23-01%20Business%20Health%20Metrics_30%20Nov%202022_FINAL.pdf

Defense Science Board. (2011). *Report of the Defense Science Board task force on improvements to services contracting.* Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics. https://apps.dtic.mil/sti/pdfs/ADA550491.pdf

Department of the Air Force. (2024, January 2). *Air Force Instruction 63–138: Acquisition of services.* https://static.e-publishing.af.mil/production/1/saf_aq/publication/dafi63-138/dafi63-138.pdf

Department of the Army. (2010, July 30). *Army Regulation 70-13: Management and oversight of service acquisitions. Headquarters, Department of the Army.* https://armypubs.army.mil/ProductMaps/PubForm/Details.aspx?PUB_ID=85329

Department of Defense. (2022, October 27). *2022 National Defense Strategy of the United States of America.* https://media.defense.gov/2022/Oct/27/2003103845/-1/-1/1/2022-NATIONAL-DEFENSE-STRATEGY-NPR-MDR.PDF

Department of Defense. (2023). *National Defense Industrial Strategy.* https://www.dau.edu/sites/default/files/2024-02/2023%20NDIS_FINAL%20FOR%20PUBLICATION%201_0.pdf

Department of Energy. (2016). *Balanced scorecard: Performance measurement and performance management program for federal procurement and management and operating contractor purchasing systems.* https://www.energy.gov/sites/default/files/2022-09/Attachment%203%20BSC%20Program%20Description.pdf

Department of the Navy. (2019, May 14). *SECNAV Instruction 4200.37A: Organic Department of the Navy procurement system oversight and management*. https://www.secnav.navy.mil/doni/Directives/ 04000%20Logistical%20Support%20and%20Services/04- 200%20Contracting%20Services/4200.37A.pdf?ID=3827

Domingues, M. L., Reis, V., & Macário, R. (2015). A comprehensive framework for measuring performance in a third-party logistics provider. *Transportation Research Procedia, 10*, 662–672. https://doi.org/10.1016/j.trpro.2015.09.020

Donabedian, A. (1988). The quality of care: How can it be assessed? *JAMA, 260*(12), 1743–1748. https://doi.org/10.1001/jama.1988.03410120089033

Downer, J. (2019). *Operating metrics that effectively and efficiently measure contract performance operations within an organization* (Master's thesis, Naval Postgraduate School). NPS Archive: Calhoun. https://hdl.handle.net/10945/64143

Elbanna, S., Eid, R., & Kamel, H. (2015). Measuring hotel performance using the balanced scorecard: A theoretical construct development and its empirical validation. *International Journal of Hospitality Management, 51,* 105–114. https://doi.org/10.1016/j.ijhm.2015.09.004

Fagundes, L. (2016, July 12). *6 KPIs to measure the performance of your contracts*. ContractWorks. Retrieved July 23, 2025, from https://www.contractworks.com/ blog/6-kpis-to-measure-the-performance-of-your-contracts

Finkenstadt, D. J., & Peterson, A. J. (2011). *A benchmark study of the Air Force Program Executive Office for Combat and Mission Support* (Master's thesis, Naval Postgraduate School). NPS Archive: Calhoun. https://hdl.handle.net/10945/10599

Franceschini, F., Galetto, M., & Maisano, D. (2007). *Management by measurement: Designing key indicators and performance measurement systems*. Springer.

Franco-Santos, M., Kennerley, M., Micheli, P., Martinez, V., Mason, S., Marr, B. … Neely, A. (2007). Towards a definition of a business performance measurement system. *International Journal of Operations & Production Management, 27*(8), 784–801. https://doi.org/10.1108/01443570710763778

Gacenga, F., Cater-Steel, A., Toleman, M., & Tan, W.-G. (2010). An international analysis of IT service management benefits and performance measurement. *Journal of Global Information Technology Management, 13*(4), 28–63. https://doi.org/10.1080/1097198X.2010.10856525

Government Accountability Office. (2005). *Framework for assessing the acquisition function at federal agencies* (GAO-05-218G). https://www.gao.gov/products/gao- 05-218g

Government Accountability Office. (2006). *Defense acquisitions: Tailored approach needed to improve service acquisition outcomes* (GAO-07-20). https://www.gao.gov/assets/gao-07-20.pdf

Government Accountability Office. (2013, June 27). *Defense acquisitions: Goals and associated metrics needed to assess progress in improving service acquisition* (GAO-13-634). https://www.gao.gov/assets/gao-13-634.pdf

Government Accountability Office. (2015a, February). *Defense major automated information systems: Cost and schedule commitments need to be established earlier* (GAO-15-282). https://www.gao.gov/assets/gao-15-282.pdf

Government Accountability Office. (2015b, March). *Defense acquisitions: Better approach needed to account for number, cost, and performance of non-major programs* (GAO-15-188). https://www.gao.gov/assets/gao-15-188.pdf

Government Accountability Office. (2021a, February 16). *Service acquisitions: DoD's report to Congress identifies steps taken to improve management, but does not address some key planning issues* (GAO-21-267R). https://www.gao.gov/products/gao-21-267r

Government Accountability Office. (2021b, July). *Federal contracting: Senior leaders should use leading companies' key practices to improve performance* (GAO-21-491). https://www.gao.gov/assets/gao-21-491.pdf

Government Accountability Office. (2023a). *DoD service contracts: Actions needed to identify efficiencies and forecast budget needs* (GAO-23-106123). https://www.gao.gov/products/gao-23-106123

Government Accountability Office. (2023b). *High-risk series: Efforts made to achieve progress need to be maintained and expanded to fully address all areas* (GAO-23-106203). https://www.gao.gov/assets/gao-23-106203.pdf

Government Accountability Office. (2025, February 25). *High-risk series: Heightened attention could save billions more and improve government efficiency and effectiveness* (GAO-25-107743). https://www.gao.gov/products/gao-25-107743

Guajardo, J., Cohen, M. A., Kim, S.-H., & Netessine, S. (2012, May). Impact of performance-based contracting on product reliability: An empirical analysis. *Management Science, 58*(5), 961–979.

Gunasekaran, A., Irani, Z., Choy, K. L., Filippi, L., & Papadopoulos, T. (2015). Performance measures and metrics in outsourcing decisions: A review for research and applications. *International Journal of Production Economics, 161,* 153–166.

Hagan, P., Spede, J., & Sutton, T. (2012). *Defining and measuring the success of services contracts in the United States Navy* (Master's thesis, Naval Postgraduate School). NPS Archive: Calhoun. https://hdl.handle.net/10945/27838

Hatry, H. P., & Bryant, S. (1998). *Performance measurement: Getting results* (2nd ed.). Urban Institute Press. https://www.researchgate.net/publication/247140740_Performance_Measurement_Getting_Results

Herbert, L., & Colenso, J. (2018). *Performance-based contracting: The next step in sourcing* [ISG white paper]. Information Services Group. https://isg-one.com/docs/default-source/default-document-library/isg-white-paper-performance-based-contracting.pdf

Hristov, I., & Chirico, A. (2019). The role of sustainability key performance indicators (KPIs) in implementing sustainable strategies. *Sustainability, 11*, 5742. https://doi.org/10.3390/su11205742

Hunter, A. P., Sanders, G., Roberts, J., Graham, J., & McDonald, L. (2019). Measuring service contract performance: Preliminary findings on effects of service complexity, managerial capacity, and prior relationship. *Proceedings of the Sixteenth Annual Acquisition Research Symposium*, 231–250. https://dair.nps.edu/bitstream/123456789/1765/1/SYM-AM-19-074.pdf

Hwang, J. (2024). Strategic alignment and performance measurement in outsourced business processes: A comprehensive analysis of key performance indicators (KPIs) and streamlining strategies. *International Journal of Frontline Research in Multidisciplinary Studies, 4*(1), 63–73. https://frontlinejournals.com/ijfrms/sites/default/files/IJFRMS-2024-0034.pdf

Hypko, P., Tilebein, M., & Gleich, R. (2010). Benefits and uncertainties of performance-based contracting in manufacturing industries: An agency theory perspective. *Journal of Service Management, 21*(4), 460–489. https://doi.org/10.1108/09564231011066114

Icertis. (2025). What is contract performance? https://www.icertis.com/learn/what-is-contract-performance/

Iden, J., & Eikebrokk, T. R. (2013). Implementing IT service management: A systematic literature review. *International Journal of Information Management, 33*(3), 512–523. https://doi.org/10.1016/j.ijinfomgt.2013.01.004

Institute of Medicine Committee on Quality of Health Care in America. (2001). *Crossing the quality chasm: A new health system for the 21st century*. National Academies Press. https://pubmed.ncbi.nlm.nih.gov/25057539/

Jones, B. (2025, May 26). *4 important contract management performance metrics.* CobbleStone Software. https://www.cobblestonesoftware.com/blog/4-important-performance-metrics-for-contract-management

Kaplan, R. S. (2010). *Conceptual foundations of the balanced scorecard* (Working Paper No. 10–074). Harvard Business School. https://www.hbs.edu/ris/ Publication%20Files/10-074_0bf3c151-f82b-4592-b885-cdde7f5d97a6.pdf

Kaplan, R. S., & Norton, D. P. (1992). The balanced scorecard: Measures that drive performance. *Harvard Business Review, 70*(1), 71–79. https://hbr.org/1992/01/ the-balanced-scorecard-measures-that-drive-performance-2

Kaplan, R. S., & Norton, D. P. (1996). Using the balanced scorecard as a strategic management system. *Harvard Business Review*, *74*(1), 75–85. https://research-ebsco-com.nps.idm.oclc.org/c/vsq674/viewer/pdf/n377p5c5rr

Kelly, J. C., Earle, C. R., Lippitz, M. J., Shapiro, B. A., & Van Atta, R. H. (2015). *Commercial best practices in contracting for knowledge-based and equipment-related services* (IDA Paper P-5257). Institute for Defense Analyses. https://www.ida.org/-/media/feature/publications/c/co/commercial-best-practices-in-contracting-for-knowledge-based-and-equipment-related-services/p-5257.ashx

Kerzner, H. (2017) Project Management: A Systems Approach to Planning, Scheduling, and Controlling. 12th Edition, Wiley, Hoboken.

Kirvan, P. (2023, April 11). *What is an operational-level agreement?* TechTarget. https://www.techtarget.com/whatis/definition/operational-level-agreement-OLA

Lowry, K. K., & Droeske, M. A. (2022). *Analysis of the 2021 Office of Federal Procurement Policy definition of procurement acquisition lead time* (Master's thesis, Naval Postgraduate School). NPS Archive: Calhoun. https://hdl.handle.net/ 10945/70742

Lucyshyn, W., & Rigilano, J. (n.d.). Improvements in services contracting: The Defense Department's implementation of performance-based services acquisition. In *Government contracting: Promises and perils* (Chapter 10). Naval Postgraduate School. https://www.acquisitionresearch.net/publications/detail/1659/

McFall, T. G., & La, D. (2011). *Mission and Installation Contracting Command services acquisition: Empirical analysis of Army service contract management practices* (Master's thesis, Naval Postgraduate School). NPS Archive: Calhoun. https://hdl.handle.net/10945/10771

McKernan, M., Drezner, J. A., Arena, M. V., Wong, J. P., Shokh, Y., Lewis, A. … Newberry, S. (2022). *Using metrics to understand the performance of the adaptive acquisition framework* (Report No. RR-A1349-1). RAND. https://www.rand.org/content/dam/rand/pubs/research_reports/RRA1300/ RRA1349-1/RAND_RRA1349-1.pdf

Messick, S. (1994). *Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning* (Report No. RR-94-45). Educational Testing Service.

Miller, F., Newton, J., & D'Amato, S. (2012, June). *Defining and measuring the success of service contracts* (Master's thesis, Naval Postgraduate School). NPS Archive: Calhoun. https://calhoun.nps.edu/server/api/core/bitstreams/8c434e14-99d4-452b-8408-f92522dac870/content

Moore, N. Y., Baldwin, L. H., Camm, F., & Cook, C. R. (2002). *Implementing best purchasing and supply management practices: Lessons from innovative commercial firms* (Report No. DB-334-AF). RAND. https://www.rand.org/pubs/documented_briefings/DB334.html

Muir, W. A. (2010). *Determinants of services sourcing performance* (Master's thesis, Naval Postgraduate School). NPS Archive: Calhoun. https://hdl.handle.net/10945/13825

Naeem, M., Smith, T., & Thomas, L. (2025). Thematic analysis and artificial intelligence: A step-by-step process for using ChatGPT in thematic analysis. *International Journal of Qualitative Methods, 24*, 1–18. https://doi.org/10.1177/16094069251333886

Neely, A., Adams, C., & Kennerley, M. (2002). *The performance prism: The scorecard for measuring and managing business success*. Financial Times/Prentice Hall

Neely, A., & West, S. (2022). Aligning performance metrics in outcome-based contracts. *California Management Review.* https://cmr.berkeley.edu/2022/06/aligning-performance-metrics-in-outcome-based-contracts/

Ng, I. C. L., & Ding, X. (2010). *Outcome-based contract performance and value co-production in B2B maintenance and repair service* (Discussion Paper No. 10/01). University of Exeter Business School. https://www.researchgate.net/publication/41903650_Outcome-based_Contract_Performance_and_Value_Co-production_in_B2B_Maintenance_and_Repair_Service

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Organisation for Economic Co-operation and Development. (2025). *Government at a glance 2025*. OECD Publishing. https://doi.org/10.1787/0efd0bcd-en

Office of the Under Secretary of Defense for Acquisition & Sustainment. (2020, January 23). *Operation of the adaptive acquisition framework* (DoD Instruction 5000.02). U.S. Department of Defense.https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/500002p.pdf

Office of the Under Secretary of Defense for Acquisition and Sustainment. (2021, June 24). *Defense acquisition of services* (DoD Instruction 5000.74). Department of Defense. https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/500074p.pdf

Office of the Under Secretary of Defense for Acquisition and Sustainment. (2024b, July). *Adaptive acquisition framework: 120-day review report.* U.S. Department of Defense.

Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics. (2012a, June 5). *Guidebook for the acquisition of services.* U.S. Department of Defense. https://apps.dtic.mil/sti/citations/ADA606206

Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics. (2012b, August 27). *Taxonomy for the acquisition of services and supplies & equipment* [Memorandum]. Department of Defense. https://www.acq.osd.mil/dpap/policy/policyvault/USA004219-12-DPAP.pdf

O'Sullivan, D. F. (2003). *What model should be used to evaluate the efficiency and effectiveness of a field contracting office* (Master's thesis, Naval Postgraduate School). NPS Archive: Calhoun. https://hdl.handle.net/10945/937

Parekh, S., Shrewsbury, P., Khera, T., & Kulkarni, M. (n.d.). *The case for outcome-based partnerships in tech services*. Retrieved July 23, 2025, from https://www.kearney.com/industry/technology/the-case-for-outcome-based-partnerships-in-tech-services

Peters, H. M. (2016, January 5). *Using data to improve defense acquisitions: Background, analysis, and questions for Congress* (Report No. R44329). Congressional Research Service. https://crsreports.congress.gov/product/pdf/R/R44329

Peters, R. N., Stewart, S. L., & Weese, S. F. (2019). *Contract management maturity model assessment of the Air Force Nuclear Weapons Center* (Master's thesis, Naval Postgraduate School). NPS Archive: Calhoun. https://hdl.handle.net/10945/62719

Pope, J. K. (1997). *Measuring the effect of the Defense Acquisition Workforce Improvement Act* (Master's thesis, Naval Postgraduate School). NPS Archive: Calhoun. https://hdl.handle.net/10945/26061

Poppo, L., & Zenger, T. (2002). Do formal contracts and relational governance function as substitutes or complements? *Strategic Management Journal, 23*(8), 707–725. https://doi.org/10.1002/smj.249

Roehrich, J. K., Lewis, M. A., & George, G. (2014). Are public–private partnerships a healthy option? A systematic literature review. *Social Science & Medicine, 113*, 110–119. https://doi.org/10.1016/j.socscimed.2014.03.037

Singh, H. (2024, July 17). *Revolutionizing industries with outcome-based contracting: A blueprint for success. Forbes.* https://www.forbes.com/sites/forbesbusinesscouncil/2024/07/17/revolutionizing-industries-with-outcome-based-contracting-a-blueprint-for-success/

Stevens, T. A., Angere, I., & Colter, K. A. (2009). *An empirical study of the management and oversight of medical services acquisition within the Department of Defense* (Master's Thesis, Naval Postgraduate School). NPS Archive: Calhoun. https://hdl.handle.net/10945/10382

Strathern, M. (1997). 'Improving ratings': Audit in the British university system. *European Review, 5*(3), 305–321. https://doi.org/10.1002/(SICI)1234-981X(199707)5:3<305::AID-EURO184>3.0.CO;2-4

Symes, A. (1999). [Review of the book *Creating public value: Strategic management in government*, by M. Moore]. *International Public Management Journal, 2*(1), 158–167. https://doi.org/10.1016/S1096-7494(00)87438-3

Under Secretary of Defense for Acquisition and Sustainment. (2024a, February 24). *Adaptive acquisition framework pathways review* [Memorandum]. Department of Defense.

Under Secretary of Defense for Acquisition and Sustainment. (2024b, June 27). *FY 2024 acquisition of services review panel wrap-up* [Memorandum]. Department of Defense. https://www.acq.osd.mil/asda/dpc/announcements/docs/servicesreviewpanel.pdf

Wilhite, T., Stover, A., & Hart, J. (2013). *Management levers that drive services contracting success* (Master's thesis, Naval Postgraduate School). NPS Archive: Calhoun. https://hdl.handle.net/10945/38941