

PHILIP S. ANTON, WILLIAM SHELTON, JAMES RYSEFF, SAMANTHA COHEN, GRANT JOHNSON,
STEPHEN B. JOPLIN, DAVID KRAVITZ, MEGAN MCKERNAN, ALEJANDRO VIGO CAMARGO

Early Predictive Indicators of Contractor Performance

A Data-Analytic Approach



For more information on this publication, visit www.rand.org/t/RR542-1.

About RAND

The RAND Corporation is a research organization that develops solutions to public policy challenges to help make communities throughout the world safer and more secure, healthier and more prosperous. RAND is nonprofit, nonpartisan, and committed to the public interest. To learn more about RAND, visit www.rand.org.

Research Integrity

Our mission to help improve policy and decisionmaking through research and analysis is enabled through our core values of quality and objectivity and our unwavering commitment to the highest level of integrity and ethical behavior. To help ensure our research and analysis are rigorous, objective, and nonpartisan, we subject our research publications to a robust and exacting quality-assurance process; avoid both the appearance and reality of financial and other conflicts of interest through staff training, project screening, and a policy of mandatory disclosure; and pursue transparency in our research engagements through our commitment to the open publication of our research findings and recommendations, disclosure of the source of funding of published research, and policies to ensure intellectual independence. For more information, visit www.rand.org/about/research-integrity.

RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

Published by the RAND Corporation, Santa Monica, Calif.

© 2022 RAND Corporation

RAND® is a registered trademark.

Library of Congress Cataloging-in-Publication Data is available for this publication.

ISBN: 978-1-9774-0954-6

Cover: Airman 1st Class Jose Miguel T. Tamondong/U.S. Air Force; max_776/Adobe Stock.

Limited Print and Electronic Distribution Rights

This publication and trademark(s) contained herein are protected by law. This representation of RAND intellectual property is provided for noncommercial use only. Unauthorized posting of this publication online is prohibited; linking directly to its webpage on rand.org is encouraged. Permission is required from RAND to reproduce, or reuse in another form, any of its research products for commercial purposes. For information on reprint and reuse permissions, please visit www.rand.org/pubs/permissions.

About This Report

Early indication of potential contractor performance risks and contract execution issues is critical for proactive acquisition management. When contractors are in danger of not meeting contractual performance goals, Department of the Air Force (DAF) acquisition management might not be fully aware of the shortfall until, for example, a schedule deadline is missed, government testing indicates a system's poor technical performance, or costs exceed expectations.

In this report, the authors describe a new way to apply data science to a variety of disparate and disjointed government and external data sources to highlight the relative contractor performance risks and provide earlier indicators of performance issues in DAF acquisition contracts and programs than would normally be achieved in traditional formal reporting. Although the authors cannot definitively state this is the optimal approach, this method seeks to produce risk and performance indicators earlier than current information sources and metrics do. This is the final report for Phase II of an effort to test the approach outlined here by building a prototype that uses actual data to calculate contractor risk measures and performance metric values relative to those of their peers, the available contractor base, or fixed thresholds, presenting outliers to prototype users for further human investigation and assessment. This report summarizes the authors' findings, including (1) a taxonomy of contractor relative risks, (2) leading indicators of performance, (3) relevant data sources, (4) risk measures and equations, and (5) a prototype that implements some of the risk measures and equations using real data sources. Note that Phase I focused on relative risk related to enabling factors for contractor performance as opposed to those related to the design and technology involved in the delivered product or service. In Phase II, we are employing new types of leading indicator metrics to gain earlier insights into potential execution issues.

The intent of our process implemented in the accompanying prototype is to highlight areas for additional attention. It is incumbent on the acquisition professional to then focus appropriate management attention on this area based on its relevance to the program and the level of risk deemed acceptable to the government.

This research should be of interest to acquisition professionals and leadership who are looking for ways to improve acquisition performance through early identification of potential relative contractor risks and execution problems to inform active program management and mitigation of risks. The prototype should be of interest to acquisition officials (from program managers to milestone decision authorities) to help them access more data in an easy-to-understand way so they can focus their limited time on areas that require increased management attention. This approach should be useful during any phase of the acquisition process—from pre-Materiel Development Decision through disposal.

The research reported here was commissioned by the Deputy Assistant Secretary for Acquisition Integration, Office of the Assistant Secretary of the Air Force for Acquisition, Technology, and Logistics (SAF/AQX) and conducted within the Resource Management Program of RAND Project AIR FORCE (PAF) as part of a fiscal year 2019 project on *Mining Program Execution Performance Problems Using Advanced Analytics* (PA19D-R9A1). It built on a project on the *Early Indicators of Relative Contractor Performance Risk for Air Force Acquisition* (PA18H-R8A4). Research in these two projects was conducted from February 2018 through September 2020.

RAND Project AIR FORCE

RAND PAF, a division of the RAND Corporation, is the DAF's federally funded research and development center for studies and analyses, supporting both the United States Air Force and the United States Space Force. PAF provides the DAF with independent analyses of policy alternatives affecting the development, employment, combat readiness, and support of current and future air, space, and cyber forces. Research is conducted in four programs: Strategy and Doctrine; Force Modernization and Employment; Resource Management; and Workforce, Development, and Health. The research reported here was prepared under contract FA7014-16-D-1000.

Additional information about PAF is available on our website:
www.rand.org/paf/

This report documents work originally shared with the DAF on October 7, 2020. The draft report, issued in December 2020, was reviewed by formal peer reviewers and DAF subject-matter experts.

Acknowledgments

Many thanks go to our DAF sponsors John Miller, Senior Executive Service; Mark Murphy, Senior Executive Service; and William D. Bailey, Senior Executive Service; and our action officers Brian Knight and Joseph A. Alfano for their support, encouragement, and guidance in this research.

The approach and prototype discussed in this report build on those in an unpublished RAND PAF proof-of-concept effort in 2017 by our colleagues Chad J. R. Ohlandt, Timothy Stacey, Cole Sutura, Brian Dolan, and Stephen Joplin. Ohlandt developed the initial idea of using relative risk measures as a way to identify potential risks for further due diligence. The remaining team members explored approaches to automatically ingest and analyze unstructured public financial Securities and Exchange Commission filings and DAF Monthly Acquisition Reports to identify important facts, cross-linkages, and situations that could form the basis for subsequent risk analysis.

We also acknowledge the valuable contributions of the other members of our PAF project team who contributed to the design and implementation of this methodology: Edward Balkovich, Lisa Colabella, Suzanne Genc, John Parsons, and David Zhang. RAND Corporation colleagues Irv Blickstein and Jeff Drezner provided useful comments on the implementation tool. Elizabeth Hastings Roer, Mark Lorell, Mel Eisman, Thomas Light, and Peter Schirmer supplied useful reviews that helped improve the report and the associated prototype.

Thanks go to Obaid Younossi, Jim Powers, Patrick Mills, and Anu Narayanan at RAND for their support and encouragement. Our administrative assistant, Silas Dustin, kept us sane and functioning.

The authors claim any errors.

Summary

A recurring challenge to successful acquisition program execution is poor contractor performance. The Deputy Assistant Secretary for Acquisition Integration, Office of the Assistant Secretary of the Air Force for Acquisition, Technology, and Logistics (SAF/AQX) asked RAND Project AIR FORCE to develop and prototype data-analytic methods on a variety of disparate and disjointed government and external data sources to identify potential execution performance and relative contractor performance risks earlier than would be typically reported via traditional methods. In this report, the authors discuss initial insights and approaches in exploring and prototyping these methods. Subsequent work and further development are ongoing.

Issues

- What techniques can be employed to assess program execution performance and risks as early as possible?
- How can program managers (PMs) and stakeholders leverage available government and external data to assess these concerns?
- What data sources are available and useful?
- What practical concerns arise when implementing such assessments?

Approach

1. Review corporate performance and risk literature and existing business intelligence tools to determine how performance and risk can be measured and what methods exist to analyze these factors.
2. Identify leading indicators of performance (see Table S.1).
3. Develop a taxonomy of potential risk measures beyond those traditionally examined in program management and that use relative performance against those of their peers or fixed thresholds to highlight risk indicator outliers to government acquisition professionals.
4. Identify potential data sources and algorithms for these measures.
5. Obtain relevant data, including development of software to extract, transform, and load data into a custom database and processing environment.
6. Develop a research prototype database and software system to test and refine the concept. Cross-reference data sources to associate contracts and contractors with Department of the Air Force (DAF) programs and build a simple user interface to view results from both a contractor and program perspective.

Conclusions

- Automated tools could be created to ingest, aggregate, and analyze data that could focus managers' limited resources on early indicators of performance issues and potential risk indicators buried in large, diverse data; this could help inform mitigating actions by management based on effectiveness, program relevance, and risk tolerance.

- Some data that are important for assessing relative contractor risks are difficult to obtain—even for DAF officials and federally funded research and development centers.
- Further work is necessary in developing a prototype with a significant critical mass of data sources and measures to test and refine this approach. User feedback on utility and design is also needed.

Opportunities

- Integrating and assessing traditional and nontraditional data sources could provide useful indications of potential areas of concern.
- With further refinement, this approach could be a powerful tool to help PMs and stakeholders leverage vast amounts of data to identify performance and risk indicators for further due diligence, confirmation, and proactive management.
- Additional data and the inclusion and validation of more metrics and implementation details are needed to make these two factors more robust; data availability, accessibility, and analysis are key.
- This is a research prototype. Despite its limitations, this approach is more sophisticated in some ways (e.g., through company-level metrics, such as financial health or supply chain risks, and predictive indicators of future performance) than other available systems and may point to features or concepts that could be added to DAF or U.S. Department of Defense systems that assess potential contractor risks.

Table S.1. Performance Indicators and Associated Data

Metric	Description
Cross-cutting metrics	
Sentiment analysis	Assess whether the summary green, yellow, or red ratings for a report appear better than the text of Monthly Acquisition Report assessments
News reports	Filters news reports that appear to indicate performance issues or situations that could lead to future performance issues
Program Assessment Reports (PARs)	Program status reports for major programs from the Defense Contract Management Agency
Rayleigh Function ^a	Predicts the final costs and schedule by curve-fitting earned value (EV) data to date to a prototypical Rayleigh curve shape
Cost metrics	
Cost Performance Index (CPI) ^b	Measures the amount of completed work for every unit of planned cost spending (i.e., the inverse of actual spending to date relative to scheduled spending to date)
Cumulative CPI versus To-Complete Performance Index to Meet the Estimate at Completion (TCPI _{EAC})	Measures the amount of work that must be done for every dollar spent to meet the current EAC (i.e., the difference between cost efficiency to date and the needed efficiency to meet the EAC)
Range of independent estimates at completion (IEACs)	Estimates four different IEACs under the following assumptions: • Future cost performance will be the same as all past cost performance • Future cost performance will be influenced by past schedule performance • Future cost performance will be influenced by past schedule and cost performance

Metric	Description
Cost-over-target growth	Similar to previous assumption, except increased weight is placed on CPI. Measures the cost over the current contract cost target (i.e., “overruns” from a certain perspective)
Schedule metrics	
Schedule Performance Index (SPI) ^b	Measures the percentage of budgeted work actually performed to date
Time-Based Schedule Performance Index (SPI _t)	Measures the average percentage of scheduled work actually being completed to date
Schedule Slippage Indicator (SPI _t vs. To-Complete Schedule Performance Index to Meet the Estimate Duration [TSPI _{ed}])	Estimates the difference between schedule progress to date (SPI _t) and remaining planned efficiency (i.e., does the current remaining schedule reflect the same schedule slippage rates?)
Independent Estimated Completion Date based on earned schedule (IECD _{es})	Generates a predicted project completion date assuming future schedule slippage will match current slippage in percentage
Contract health metrics	
Work-content growth	Measures the growth in the amount of work content added to the contract to date

NOTES:

^a The Rayleigh Function has been modeled but not implemented in the research prototype.

^b This is a traditional EV management metric.

Contents

About This Report.....	iii
Summary	vi
Figures.....	xi
Tables.....	xii
1. The Challenge: Early Identification of Contractor Execution Problems in Acquisition	1
Early Identification of Execution Problems	1
Validity of the Measures.....	4
Complementary Identification of Relative Contractor Risks	5
Operationalization: Business Intelligence for Acquisition Professionals.....	8
Data Streams and Risk Evaluation	9
Operationalization of a Prototype.....	13
Organization of This Report	14
2. Indicators of Performance	15
Qualitative Indicators: Detecting Performance Issues Buried in Textual Reports.....	17
Quantitative Indicators of Performance.....	21
Cost Metrics.....	24
Schedule Metrics	31
Joint Cost and Schedule Metrics	37
Contract Health Metric	40
Opportunities for Future Capabilities	41
Summary.....	42
3. Prototype Structure and Approach.....	43
User Interface	43
Intended Uses	44
Combining Performance Measures	53
Observations	57
Architecture Overview	57
4. Insights, Conclusions, and Next Steps.....	59
Insights on Data Utility and Availability.....	59
Conclusions	60
Next Steps.....	61
Appendix A. Sentiment Modeling and Analysis and Modeling of Free-Text Assessments and Associated Ratings	64
Appendix B. Assessing Final Cost and Schedule Using a Rayleigh Function	74
Abbreviations.....	85
References.....	88

Figures

Figure 1.1. The Concept of Integrating Data to Measure Contractor Risks	6
Figure 2.1. Output of the Sentiment Model	19
Figure 2.2. EV Variables over Time	23
Figure 2.3. CPI Notional Example.....	24
Figure 2.4. CPI vs. $TCPI_{EAC}$ Notional Example	26
Figure 2.5. Range of IEACs (Notional Example).....	28
Figure 2.6. Relationship Between Total Contract Cost Growth, COTG, and WCG	30
Figure 2.7. Notional Example of SPI.....	31
Figure 2.8. Calculating ES and AD	33
Figure 2.9. Illustration of PDWR and RD	34
Figure 2.10. Illustration of $IECD_{es}$	36
Figure 3.1. Prototype User Interface Structure: Contractor and Program Views	44
Figure 3.2. Example Prototype User Interface: All Contractors Page (anonymized).....	47
Figure 3.3. Example Prototype User Interface: Single Contractor Page (anonymized)	49
Figure 3.4. Prototype User Interface: All Programs Page (anonymized)	51
Figure 3.5. Prototype User Interface: Single Program Page (anonymized).....	52
Figure 3.6. Relative Risk Rating and Score.....	54
Figure 3.7. Combining Indicators Within and Across Categories to Reveal Worst Potential Risks	55
Figure 3.8. Relative Risk Rating Based on Combined Scores for Multiple Measures	56
Figure 3.9. The Prototype's Architecture	57
Figure A.1. Example MAR Scores and Assessments (anonymized).....	65
Figure A.2. Semantic Text Assessment Results (using LightGBM)	71
Figure B.1. Example Rayleigh Cumulative Plot.....	74
Figure B.2. Example Rayleigh Cost and Duration Predictions	78
Figure B.3. Multiple Curves Fitting for Work-Added Adjustments.....	79
Figure B.4. R^2 for Example Contracts That Display an S-Curve Shape Over Time	80
Figure B.5. Prediction Plots with Confidence Intervals	83

Tables

Table S.1. Performance Indicators and Associated Data	vii
Table 1.1. Assessed Leading Indicators; Implemented Viable Candidates in the Prototype	3
Table 1.2. Implemented Risk Measures by Category	7
Table 1.3. Components of Risk Factors Considered.....	12
Table 2.1. Performance Indicators and Associated Data	15
Table 2.2. Data Sources Incorporated in the Prototype	16
Table 2.3. Keywords for News Story Filter	20
Table 2.4. EVM Terms	22
Table 2.5. CPI Metric Thresholds in the Prototype	25
Table 2.6. CPI vs. $TCPI_{EAC}$ Metric Thresholds in the Prototype.....	26
Table 2.7. Range of IEAC Thresholds in the Prototype	29
Table 2.8. COTG Metric Thresholds in the Prototype.....	30
Table 2.9. SPI Metric Thresholds in the Prototype.....	32
Table 2.10. SPI_t Metric Thresholds in the Prototype.....	33
Table 2.11. SPI_t vs. $TSPI_{ed}$ Metric Thresholds in the Prototype.....	35
Table 2.12. ECD_{CTR} vs. $IECD_{es}$ Metric Thresholds in the Prototype	37
Table 2.13. Possible Rayleigh Cost and Schedule Assessment Criteria.....	40
Table 2.14. WCG Metric Thresholds in the Prototype	41
Table 2.15. Candidate Metrics for Future Implementation in the Prototype	42
Table A.1. Program Types and Sample Size	66
Table A.2. Number of Programs by ACAT Level in MAR Sample.....	67
Table A.3. Color Rating Assessment Categories in MAR Sample.....	67
Table A.4. Average Feature Importance for the Four Text and Two Category Variables	72
Table B.1. Percentage Difference Between Rayleigh Predictions and Contract Estimates.....	77
Table B.2. EVM Data Summary Statistics	82

1. The Challenge: Early Identification of Contractor Execution Problems in Acquisition

Acquisition professionals deal with a complex set of initial conditions, inputs, and processes to deliver an article within a specified cost and schedule with a desired level of performance. This effort requires integration of a large amount of information, the ability to discern and focus on critical items, and the ability to act responsively and as responsibly as possible. This allows the acquisition professional to mitigate, if not prevent, additional issues, keep bad situations from worsening, and reduce the impact on resources (cost and schedule) to achieve the desired output (performance).

Early Identification of Execution Problems

Concerns continue to be raised about cost and schedule growth in acquisition and some postulate about a lack of knowledge about the status of acquisition programs.¹ To help address these larger concerns, we examined metrics to identify emerging execution problems earlier than traditional acquisition oversight systems to enable more-proactive risk and performance management. Risk is a key component to any business transaction, especially transactions worth millions, if not billions, of dollars. Buyers and suppliers view risks differently, although their risks are interrelated. As a buyer, the Department of the Air Force (DAF) faces both internal and external risks (Anton et al., 2022). The first iteration of this work focused on external, relative risks associated with the contractors from whom the DAF acquires goods and services. Acquisition professionals, by tracking business intelligence (BI), receive notification of external risks (Anton et al., 2022). The tracking and notification are outward-looking and, therefore, are a much harder problem to solve because of the difficulties of attaining sufficient data and managing those data over thousands of contractors. Anton and colleagues aimed to solve this problem by creating useful outward-looking BI for the DAF to use in its acquisition management and oversight (Anton et al., 2022). The work described in this report continues that effort by focusing on metrics to identify emerging problems early in program execution. In other words, the previous report on risk identification seeks to identify potential problems, whereas, in this report, we add execution problem identification to determine whether potential problems have come to fruition and to enable proactive management.

Some of the DAF's external risks originate from contractors. Thus, Anton et al., 2022, used various contractor-level characteristics as indicators of potential relative risks associated

¹ See, for example, U.S. Government Accountability Office (2019, 2020b); Schwartz and Peters, 2018.

conceptually with a contractor's business performance. In this context, risk is a characteristic of a DAF contractor that, under certain conditions, could contribute to poor business outcomes and thus negatively affect schedule, cost, or performance in a DAF contract or associated acquisition program.² The relative aspect of this approach compares metric values with those of their peers, the available contractor base, or fixed thresholds, presenting outliers to prototype users for further human investigation and assessment. Relative performance helps avoid the problem of always needing an absolute level of known risk (likelihood and consequence) that correlates with bad outcomes. In this report, we are simply identifying performance indicators in which a company is performing worse than other companies as a potential indicator of a problem. Whether that indicates an actual problem would depend on further examination (often by the program manager [PM] or other officials) given other information not available in the data used in the prototype. Some of these relative measures are statistical in nature, while others use fixed thresholds based on best practices or experiences with the measure in question—yet, even in these cases, they remain potential indicators of problems whose causes and consequence need to be determined. The prototype's objective is to filter out and alert on these potentials from the much larger set of contractor performance values.

We build on this framework by identifying metrics that could show risk early in the life of the contract or program. These metrics focus on program execution risk and are inherently more inward-looking. First, the National Defense Industrial Association's (NDIA's) Integrated Program Management Division published a report in 2017 on using predictive measures as program management tools. As part of our research, we implemented measures (based on this 2017 NDIA report) for which we had data; these measures reflected logical relationships between past and future performance as measured by earned-value (EV) and other data. Second, we used the research by Davis, Christle, and Abba, 2009, that uses the Norden-Rayleigh distribution as a model to predict total cost and schedule at completion of program. The Norden-Rayleigh model takes a few data points on cost and schedule and predicts total cost and schedule over the contract's lifetime. Finally, we used machine learning and sentiment analysis to explore whether the text in monthly program evaluation reports aligns with the program's ratings. Later in this chapter, we discuss the metrics we explored further and how we integrated them into our

² The U.S. Department of Defense (DoD) defines *risks in acquisition* as “potential future events or conditions that may have a negative effect on achieving program objectives for cost, schedule, and performance” (Deputy Assistant Secretary of Defense for Systems Engineering, 2017). When fully understood and measured, risks have a probability, consequence, impact, or severity of the undesired event on cost, schedule, and performance were the potential event or condition to occur. The probability reflects both variability and uncertainty. Our approach focuses on the *risk identification* phase in DoD's risk process planning cycle, whereas subsequent due diligence of alerted relative risks would involve the *risk analysis* to determine the likelihood and consequence of the risk (Deputy Assistant Secretary of Defense for Systems Engineering, 2017). Therefore, our concept is a data-driven *risk identification methodology* focused on inclusion of data not normally employed in such identification and on prime and lower-tier contractors within acquisition programs.

prototype called Performance Indicators and Execution Risks of Contractors Early On (PIERCE). These leading indicators are presented in Table 1.1. Note that these indicators use prior performance as the basis to predict future performance. This serves two purposes: (1) to reveal performance issues as early as possible and (2) to try and predict the resulting total effect if problems persist.

Table 1.1. Assessed Leading Indicators; Implemented Viable Candidates in the Prototype

Cross-cutting metrics – MARs sentiment analysis* – Defense Acquisition Executive Summary (DAES) sentiment analysis** – Selected Acquisition Reports (SARs) sentiment analysis** – News reports* – PAR ratings* – MAR and DAES ratings** – Rayleigh Function** – Obligations vs. program plan*** – Expenditures vs. program plan***	Requirements metrics – Requirements volatility** – TBD/TBR burn down*** Contract health metric – Contract mods* – Baseline revisions** – Program funding plan*** – Research, development, test, and evaluation (RDT&E) actual billings vs. forecast billings***
Cost metrics – Cost performance index (CPI)* – Cumulative CPI vs. TCPI_{EAC}* – Range of IEACs (independent estimates at completion)* – Cost-over-target growth (COTG)*	Supply chain metrics – Parts demand fulfillment*** – Supplier acceptance rate*** – Supplier late starts***
Schedule metrics – Schedule Performance Index (SPI)* – Time-Based Schedule Performance Index (SPI_t)* – Schedule slippage (SPI_t vs. TSPI_{ed})* – Contractor Schedule Estimate Accuracy* – Bid protest (alert on protest)**	

NOTES: * = Text in green denotes indicators prototyped. ** = Text in purple denotes a future prototype option.

*** = Text in black denotes a data challenge.

MAR = Monthly Acquisition Report; PAR = Program Assessment Report; TBD/TBR = to be determined/to be resolved; TCPI_{EAC} = To-Complete Performance Index to Meet the Estimate at Completion; TSPI_{ed} = To-Complete Schedule Performance Index to Meet the Estimated Duration.

In 2017, NDIA published a guide to managing programs using leading indicators. Leading indicators are predictive measures that signal future outcomes based on the modeling of past cost, schedule, and performance metrics. The guide describes developing predictive measures to manage program success and predict program performance by identifying root causes of problems and action items to correct performance downfalls. From this work, the guide recommends a standardized set of 24 predictive metrics. It defines the metrics, presents methodology on how they are calculated, reviews their outputs and what rating thresholds the NDIA team typically found meaningful, the predictive information for each metric, questions that an acquisition professional using the metrics might have, and any limitations (NDIA, 2017). Our effort explores whether and how these and other metrics could be implemented with actual data to identify risk indicators of potential concern for further due diligence earlier than more traditional reporting within DoD.

Validity of the Measures

In an ideal world, one would test proposed metrics (such as those in this report) by seeking statistical correlation with contract outcomes. Davis, Christle, and Abba, 2009, reported such evidence that their Rayleigh Function metrics correlate with final outcomes.

Also, we discuss in this report that expert evaluation determined that the sentiment analysis model was able to identify cases where the human submitter's rating was more favorable than indicated by the situation discussed in the free text; however, we have not empirically tested the long-term relationship between the performance ratings (green, yellow, or red) and near-term final program outcomes.

Finally, the predictive measures that we implemented from the 2017 NDIA report reflect logical relationships between past and future performance as measured by EV and other data. Various studies have examined the performance of the EV-based metrics as predictive of final values. For example, other researchers provide analyses examining the predictive performance of the earned schedule (ES)-based concept underlying three of the metrics implemented in our prototype (e.g., Lipke, 2009; Henderson, 2004, 2005, and 2007; and Petter, 2014). They found the results were better than those in more traditional EV metrics. The predictability of the estimates improves significantly after 20 percent of the project completion and tend to be generally accurate after about 60 percent (Henderson and Zwikael, 2008; Lipke, 2009). Also, Henderson and Zwikael, 2008, reviewed and assessed whether early values of CPI and SPI_t remain stable over time and thus help predict the final CPI values. Coleman et al., 2003, examined the extent to which early CPI values are predictive of final values. The U.S. Government Accountability Office (GAO, 2020a) also discusses that CPI, CPI vs. TCPI, range of IEACs, and the Rayleigh model can be used to help predict the final contract cost. Various other studies (e.g., Christensen and Payne, 1992; Christensen and Heise, 1993; Christensen and Templin, 2002; Petter, 2014; and Clayson, Thal, and White, 2018) provide additional statistical analyses on the correlation of early CPI with final CPI (i.e., CPI stability).

It should also be noted that actual projects with poor performance are subject to intervention by PMs. It is difficult to obtain variables that track management action to ascertain the degree to which management intervention affects final outcomes compared with the initial situation. Such intervention is the point of our work and the larger EV management (EVM) approach—specifically, to identify performance issues for PMs so that they can take proactive measures to help mitigate the issues. These measures use performance and issues to date as the best (and sometimes only) indicator of future problems, so they are a combination of facts together with logical extrapolations into the future. Evidence points to their validity, but the interventions make it nearly impossible to state absolute statistical correlations because of intervening variables for which we have no data.

Complementary Identification of Relative Contractor Risks

Earlier identification of execution issues is the flip side of the coin of our earlier work identifying relative contractor risks (Anton et al., 2022). Risk identification seeks to identify potential problems, whereas execution problem identification seeks to identify whether potential problems have come to fruition. Our analytic approach adds new execution indicators to the same prototype. We briefly discuss the issues and approaches related to identifying *relative* contractor risks. We refer the reader to our other report (Anton et al., 2022) for details on relative contractor risks.

Contractor Risks in Acquisition

Risk management in acquisition requires the understanding of factors from various sources associated with the contractor, program development, and workforce. Tracking data sources not typically monitored by acquisition professionals can help synthesize these factors to better manage risk. Often, the DAF becomes aware of a risk after the damage has been done. For instance, when contractors are in danger of a cost overrun, acquisition professionals might not be aware until it is realized. To streamline risk management and track data sources not typically monitored, the research team built a web-based application prototype that uses nontraditional data sources to measure relative risk levels for contractors and programs associated with the DAF (we refer to this work as *Phase I*). We build on Phase I by identifying further data resources and methods to identify potential contractor risks and execution problems early.

Risks in acquisition associated with the contractor, program development, and workforce can affect acquisition program success through schedule, cost, and performance. For example, if the contractor is having difficulty hiring personnel critical to program execution, this could result in technical development difficulties and delays which could eventually cause contract cost increases or schedule delays. If the government program office had earlier indications of this risk, then it could apply appropriate management attention and/or develop a mitigation strategy. Doing this sooner could reduce, or even eliminate, adverse cost and schedule effects and keep the program on track.

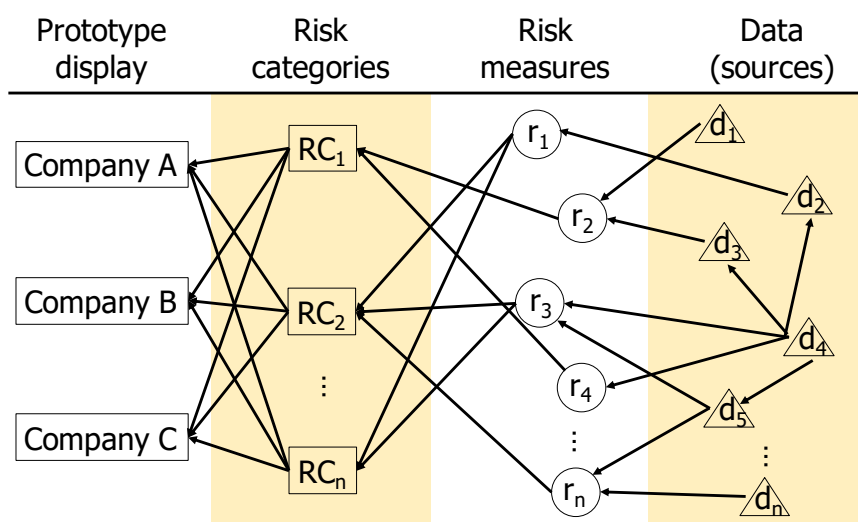
Traditional acquisition oversight data are collected to promote transparency in spending, manage costs, report on defense spending, achieve small-business goals, manage fraud, improve decisionmaking, and archive acquisition programs (McKernan et al., 2017). The data systems designed to achieve these goals were developed in response to different policy directives, congressional mandates, and requirements in DoD Instruction (DoDI) 5000.02. McKernan et al., 2017, found the disjointed nature of these data sources can make it challenging for acquisition professionals to synthesize and achieve the goals that these data systems were meant to meet. Additionally, many of the data systems, especially the monthly reporting ones, impose a reporting burden on the contractor and the program office. The inefficiencies of synthesizing data siloed in disparate systems that require unique permissions to access in addition to the

timely requirements necessary for data upkeep creates barriers to effective risk management. This report describes the work done to continue developing a prototype that integrates these data streams and provides acquisition professionals with program and contractor assessments that can be used to manage risk more efficiently.

Indicators of Potential Risk

Our prior work built a taxonomy of contractor performance risks that could be measured by integrating internal DAF and other DoD data (including proprietary and other sensitive data), public government and commercial data, and licensed commercial data. We then developed and employed analytic techniques and algorithms to combine these structured and unstructured data to identify relative contractor performance risks. A unique characteristic of our approach is the relative aspect of contractor risk. Our methodology compares contractors with their peers to identify outliers. These comparisons can indicate areas where DAF acquisition professionals should focus time and due diligence to determine how relevant the risks are and how they can be mitigated. The general concept is illustrated in Figure 1.1.

Figure 1.1. The Concept of Integrating Data to Measure Contractor Risks



Relative risk measures in our other work are summarized in Table 1.2.

Table 1.2. Implemented Risk Measures by Category

Workforce	Influence
– Hiring (U.S. Bureau of Labor Statistics [BLS] unemployment rate)	– Insignificant customer (DAF, DoD, federal)
– Hiring and retention (job openings)	Performance
– Experience in key work area (product and service code [PSC])	– Past or current contractor performance (Defense Contract Management Agency [DCMA] PARs)
Cost (price)	– Experience in key work area (PSC)
– Overhead (declining revenue)	– Prior experience working for DAF, DoD, or other federal agency
– Invest to increase capacity (jump in revenue)	Capacity
Financial	– Invest to increase capacity (jump in revenue)
– Financial metrics (overall; operational, solvency, and liquidity)	Cross-cutting
– Customer base: declining revenue	– News alerts (keyword filters)
Stability	Supply chain
– Excluded contractor	– Corporate-wide suppliers (risks recursively applied)
– Recent and pending mergers and acquisitions (M&As)	– Contract-specific suppliers (risks recursively applied)

SOURCE: Anton et al., 2022.

NOTE: Measures are unordered within each category, and there may be partial correlations between different measures within each category.

Relative (Not Absolute) Risks

A key aspect of our risk-identification approach is the emphasis on relative³ risks. Here we are comparing a company relative to its peers to identify outliers where DAF PMs and executives should focus time and due diligence to determine risks' relevancy and potential mitigation strategies.

This approach allows the user to compare one company with either a targeted peer set or with the hundreds to thousands of other companies in the data set, thereby making outliers more significant. Although this approach appears potentially powerful and useful, it does have several limitations: Most of the data are available only for publicly traded companies, we are dependent on others for data collection and accuracy, and the approach does not identify program-specific risks in an absolute manner. Whether the risks have significant probabilities or consequences for

³ It is key to emphasize the focus on relative versus absolute risks. The information to assess absolute risk in an acquisition is typically not available outside a specific program office or the associated contractor(s). Therefore, performing an absolute risk assessment over a large set of contractors and programs would entail repeated data calls. This would be problematic for several reasons, a primary one being the additional burden on the program office and contractor with the resulting increased costs. Additionally, collecting this for all DAF programs continuously is simply impractical. By comparing companies with peers, the PM has insight as to how the program's contractor(s) perform relative to a specific industry (e.g., major aerospace companies) or to the defense industrial base as a whole and can focus on outliers. Currently, the prototype generally compares performance against the entire available industrial base, but further research could determine ways to target a narrow peer set more effectively. For example, CapIQ offers a peer set when rating a company's financial health.

cost, schedule, or performance, how these concerns may affect the government's level of acceptable risk, and whether actions that are already being taken by the government or contractors are sufficient to mitigate the risks are things to be determined by the government acquisition professionals.

In addition to the data streams traditionally used by acquisition professionals, we identify nonconventional data streams that provide information to manage risk at the program and contractor level. Public data sources, such as web-based news and government statistics, can provide information on contractor performance and stability. Additionally, integrating the disparate DoD data sources could make risk management more efficient and allow acquisition professionals to identify risk earlier. Typically, acquisition professionals have the tools to assess risk related to the design and technology of the acquisition product or service. Traditional approaches include DoD risk management and assessment procedures (e.g., DoDI 8510.01, 2017; Deputy Assistant Secretary of Defense for Systems Engineering, 2017) and program and contract management controls (e.g., DoD, 1988; Parker, 2011; Bounds et al., 2014; Naval Air Systems Command, 2015; Marine Corps Systems Command, 2020). The prototype described in this report aims to complement these traditional methods with novel approaches from nonconventional data sources.

Operationalization: Business Intelligence for Acquisition Professionals

At a broad level, BI is a tool that organizations use to manage risk through collecting and analyzing data on various characteristics of a business's internal and external relations. Some organizations conduct BI in-house; however, there are many third-party organizations whose mission is to provide BI for others. These include self-service platforms, such as SAP Crystal Reports or Alteryx (Alteryx, undated), which provide software with tools to organize and analyze data but leave those steps up to the company using their product (SAP Crystal Reports, undated). BI companies that focus on data visualization specifically include iDashboards (iDashboards, undated) and Dundas BI (Dundas BI, undated). Other BI interfaces focus on data warehousing and cloud computing. These include Sisense and Oracle BI (Orad, 2020; Oracle Solutions Business Intelligence, 2019). Lastly, there are BI companies that not only provide software to manage and analyze data but also use modeling and machine learning to draw insights themselves. These include such companies as InsightSquared and Domo (InsightSquared, undated; Domo, undated).

Although the research team acknowledges the existence of commercial BI tools, we did not find one that solves the issue of relative contractor risk management in the context of acquisition management and oversight for the DAF or that uses statistical analysis to determine and prove which measures correlate with undesirable performance outcomes. We referred earlier to such published analysis for performance indicators, but more work could be done in this regard. Arguably, some of the platforms could be useful for inserting data to manage, model, and

visualize. Additionally, various platforms offer risk identification and analysis through deep learning. But in the context of acquisition management and oversight for the DAF, it is likely that these platforms would neither collect the needed data because of government data restrictions nor use methods to model the relative contractor risk for acquisition oversight managers to monitor program success. Additionally, the existing BI tools tend to be inward-looking. As previously mentioned, the DAF as a buyer needs outward-looking BI to parse potential risks over thousands of vendors—starting with prime contractors and including as many suppliers as can be identified with available data. This requires additional complexity to BI where the DAF must collect data over a wide variety of sources, often ones that are sensitive and proprietary, to effectively proxy or capture relevant contractor risks.

Data Streams and Risk Evaluation

In addition to the NDIA metrics, we use Rayleigh Functions to forecast total cost and schedule at contract completion. The Rayleigh Function has been shown to estimate total cost and contract duration using only three time and cost data points from the contract (Davis, Christle, and Abba, 2009; Lee, Hogue, and Gallagher, 1993; Dukovich, Houser, and Lee, 1999). It can easily be updated as new data become available and produce forecasts of total cost and contract duration. The benefits of this approach are that the data requirements are minimal, the results can provide early warnings of potential cost overruns or schedule slips, and the approach can be used to evaluate realistic program RDT&E funding profiles. The accuracy of confidence intervals constructed under this method relies on the assumption that the data follows a Rayleigh distribution, making it most relevant for research and development (R&D) contracts and those for which there is a start-up and ramp-down effort (Davis, Christle, and Abba, 2009; Dukovich, Houser, and Lee, 1999). We built a Rayleigh model to predict cost and schedule for programs and use the results to determine whether a program is at risk of cost or schedule overruns. The functional form and methods are discussed further in Chapter 2.

Our third and final addition to PIERCE uses a machine learning model that we developed to analyze the sentiment⁴ of free text in MARs. MARs are reports from a government reporting system that provide program statuses and summaries of issues, risks, costs, performance, and life-cycle sustainment. The reports include free text written by the PM, program executive

⁴ By *sentiment*, we mean the general assessment of the status and risk of the program or its contracts as conveyed by the text in the report.

officer (PEO), and “3-Letter,”⁵ and seven ratings of green, yellow, or red⁶—one each for the PM, PEO, and 3-Letter/other’s overall assessment and a focused overall program rating of cost, schedule, performance, and funding. We hypothesized the sentiment of the text descriptions occasionally disagrees with the color ratings. For instance, a program could be rated green for schedule, but the description in the text written by the PM, PEO, or 3-Letter could suggest the program should be rated as yellow or red. Thus, we used machine learning techniques to take the text descriptions of a MAR and predict what the ratings should be for the seven metrics. The results show that there are instances where the model predicts a worse rating than the PM, PEO, or 3-Letter assigned. In PIERCE, we take these cases and present the outcome on the relevant program page. Acquisition professionals using PIERCE will be notified that the model predicted a worse rating than reported in the MAR and have the opportunity to read the actual MAR text. This contributes to early identification of potential risks and issues for those who are not already intimately familiar with the situation in the program (e.g., at the assistant secretary level and associated staff), alerting them to issues that might have otherwise stayed under the radar. The data and model are described more in Chapter 2.

⁵ In the DAF (and in much of DoD), leaders and their organizations are abbreviated with letters. Higher-level leaders and their organizations have fewer letters, while their subordinates usually add a letter to their symbols. Thus, for example, SAF/AQX (the Deputy Assistant Secretary for Acquisition Integration in the Office of the Assistant Secretary of the Air Force for Acquisition Technology and Logistics) is the 3-letter organization under SAF/AQ (the Assistant Secretary of the Air Force for Acquisition Technology and Logistics). In MARs, assessments are often provided by 3-Letter staff organizations on the status of the program independent of those by the PM and PEO.

⁶ These colors reflect the subjective expert judgments of the human assessor. Their criteria generally follow the guidance published for the DAES reports (Defense Acquisition University [DAU], 2018, CH 1–5.1.1.1):

On-Track (GREEN). The program is progressing satisfactorily in the given assessment area. Some minor problems may exist, but appropriate solutions to those problems are available; none of the problems are expected to affect overall program cost, schedule, interdependency risk, and performance requirements; and none are expected to require managerial attention or action.

Potential or Actual Problem (YELLOW). Some event, action, or delay has occurred that may impair progress against major objectives in one or more segments of the program, and may affect the program’s ability to meet overall cost schedule and performance requirements or other major program objective, or may eventually result in a breach of Nunn-McCurdy unit cost targets.

Critical (RED). An event, action, or delay has occurred that, if not corrected, poses a serious risk to the program’s ability to meet overall cost, schedule, and performance requirements, or other major program objective; would require significant funding changes; or may result in a breach of Nunn-McCurdy unit cost thresholds or a Chapter 144A Significant or Critical Change.

Note that these criteria are more detailed but align generally with those in the old, superseded version of Air Force Instruction 63-101, 2011, for Initial Contract Performance Reviews.

How This Methodology Might Be Used

The goal of this approach is to provide acquisition professionals, from PMs to milestone decision authorities (MDAs), a way to access more data in an easy-to-understand way so they can focus their limited time on areas that require increased attention. This would be useful throughout the acquisition life cycle—from pre-materiel development decision through disposal.⁷ However, it is not a replacement for risk management and must be supplemented by the user's acquisition expertise, program knowledge, and priorities (both internal and external) to appropriately highlight areas for additional focus.

Because the usefulness of these measures relies on PMs' and oversight managers' insight and expertise rather than on absolute and verified risk correlations, the usefulness of various measures might be situation-dependent, and lessons on their usefulness may change over time. Therefore, as experience is gained in using these measures, the users might want to adjust the weights for each measure to reflect each situation or this learning.

Use by PMs

Defense acquisition PMs are typically focused on executing their specific program and meeting the cost, schedule, and performance goals set and agreed to in the acquisition program baseline (APB), sometimes referred to as the contract between the PM and the MDA. Although the APB specifies the major goals for the program, it does not explicitly discuss the many underlying factors that contribute to meeting these objectives. Therefore, the PM can become singularly focused on these objectives and, unbeknownst to this person, miss other risk indicators until they manifest as issues.

Because this methodology uses data from a variety of government-controlled, public, and commercial sources⁸ and identifies relative risks compared with peers, it reveals areas that PMs should consider in evaluating the risks and whether the contractor is mitigating them. This approach disaggregates the traditional areas of cost, schedule, and performance risk into 11 categories, which are further decomposed into factors as shown in Table 1.3. Details are discussed in Anton et al., 2022.

⁷ If this type of approach is going to be used in a source selection, the government should consider whether to identify this in Section M of the associated request for proposal when factors for evaluation are identified. This approach could supplement the past performance information and any risk assessment.

⁸ For example, this methodology accesses data from the BLS, Federal Procurement Data System—Next Generation (FPDS-NG), the Federal Funding Accountability and Transparency Act Subaward Reporting System (FSRS), U.S. Securities and Exchange Commission (SEC) filings, news reporting, and Standard & Poor's Global Market Intelligence, and more-traditional DAF program management information systems (e.g., the Project Management Resource Tools [PMRT]).

Table 1.3. Components of Risk Factors Considered

Workforce	Cost (price)	Financial	Corporate Stability
Cleared staff, meeting staffing plan, attractiveness of work location, inflation at work location, prior federal work, experience (leadership and staff), R&D spending to support design teams, retention (salary growth)	Inflation in primary work location(s), large revenue increases (does capacity need to increase?), decreases in revenue (increases in overhead), production costs, inflation (materiel)	Credit health score (solvency, liquidity, operational), declining profits, solidity of customer base	Recent and pending M&As, declining stock price, C-Suite turnover, lawsuits (pending or losses)
Supply Chain	Influence	Contractor Performance	Security
Suppliers (corporatewide, contract-specific), reduced availability of sources, quality assurance, parts (tampering, counterfeit)	Significant customer (DAF, DoD, federal)	Prior work experience (specific product, service, or technology), federal, DoD, commercial reputation	Recent cyber compromises, meeting cyber requirements, sufficient cleared staff, sufficient cleared workspace
Infrastructure	Capacity	Future	
Production stability, new infrastructure required	Production capacity (economic order quantities), large increases in revenue (more needed capacity)	Low R&D investments (corporate, government-funded), published technology works, patent applications	

SOURCE: Anton et al., 2022.

PMs typically prefer knowing about risks before they become issues. One example is anticipating execution risk. One source of execution risk is related to the contractor's staffing—sufficient quantities of critical skillsets. The PM might not know that the contractor is having difficulty filling critical personnel positions until he or she sees a degradation in EV metrics, e.g., the schedule slips because work packages on the critical path are not completed on time. If sufficient data were available, this approach conceptually could indicate the Workforce category is rated worse than good, allowing the PM to drill down and see that there are several vacancies the contractor is trying to fill and where those vacancies are, or that the contractor received a larger contract from another government entity that requires the same kinds of skills. This could indicate that the PM's contractor team could be redirected to another effort. Knowing these are possibilities supported by data, the PM could address these risks with the contractor before they manifest as issues.

Use by an Analyst

Because PIERCE allows the user to display information in myriad ways, analysts would be able to answer a variety of questions. For instance, if it were necessary to understand how pervasive the risks experienced by a single or group of contractors may be, then the analyst could navigate to the contractor's page, identify the DAF programs that the contractor is participating in, and assess that risk. Or, if a program was being considered for cancellation or significant reductions, the analyst could navigate to the program page and identify the prime contractor and the reported subcontractors. This, coupled with the analysts' expertise, would allow him or her to make a first order approximation of a program's cancellation on that sector of the defense industrial base.

These are but two potential uses. Imaginative analysts will find other uses as they gain familiarity with PIERCE, more data are made available, and more metrics are implemented.

Use by an Oversight Manager

This approach could be even more useful for an oversight manager responsible for a portfolio. When overseeing multiple programs, the manager is dependent on program reporting or data calls. Although PIERCE does not identify absolute risks, it does use diverse data sources and combines them to highlight potential areas for increased management attention. This supplements traditional reporting (e.g., MARs, SARs, and DAES), allowing the oversight manager to quickly assess the relative contractor performance risks across multiple programs (or contractors) via a dashboard. The manager will see the categories in Table 1.3 for each contractor of interest and see how many DAF programs this contractor participates in. This would rapidly give the oversight manager the scale of a risk and potential impact to his or her portfolio. Armed with these indications, additional management attention can be applied to the individual programs or to the contractor's corporate leadership, as appropriate.

As with the PM, this approach supplements other available information and needs to be reviewed within current priorities, both internal and external to the portfolio. This insight would help focus the acquisition executive on the more pressing issues of the day.

Operationalization of a Prototype

This effort began the operationalization of PIERCE through a series of beta testing. Using PIERCE as developed in Phase I and implementing the three additional risk categories discussed above; NDIA leading indicators, Rayleigh functions, and sentiment analysis of MARs; we had RAND researchers and DAF subject-matter experts explore PIERCE and report back on their user experience and evaluate the information presented. These beta users accessed PIERCE through a secure, common access card (CAC)-enabled web-based application. We collected informal feedback and then made improvements to both the user interface and risk reporting methods. Although the user interface and reporting of risk metrics has improved, there are still

areas for improvement. First, the design of the user interface could have a more modern look. The current version took a utilitarian approach to user interface, focusing time and resources to developing and implementing the risk metrics. Second, certain data limitations prevent us from having some risk metrics update automatically. Thus, the relevant risk metric results are somewhat outdated until new data are manually updated. Lastly, there are limitations where the data are missing for specific contractors and programs across the metrics described above. We did not have resources for extensive user testing and examination of the utility of the performance metrics themselves; future exercise of the prototype would be useful to gain this feedback. Despite these limitations, however, PIERCE is uniquely positioned to inform DAF professionals on risk metrics at the program and contractor level early on and from nontraditional data streams.

Organization of This Report

In Chapter 2, we describe two kinds of algorithms implemented in PIERCE: those using textual analysis to indicate potential performance problems and those using quantitative performance leading indicators.

In Chapter 3, we describe the architecture and approach taken to prototype these leading indicators in combination with the prior relative risk methodology. We also describe how the performance indicators are combined to obtain a top-level score that highlights the worst indicator out of all the metrics for each program.

In Chapter 4, we summarize the insights and lessons learned along with suggested next steps for implementing relative risk measures using these expanding data sets.

In Appendix A, we provide methodological details on the sentiment modeling and analysis of free-text assessments and their associated ratings.

In Appendix B, we provide methodological details on our Rayleigh model (not yet integrated into PIERCE).

2. Indicators of Performance

In this chapter, we discuss the qualitative (textual) and quantitative indicators of early performance we assessed and implemented in PIERCE. These metrics were identified using a review of the available literature on leading indicators and brainstorming within our project.

Table 2.1 lists the performance indicators from Table 1.1 that we implemented in the current prototype or for which we conducted preparatory analysis; these metrics are described in this chapter. These indicators measure performance in various dimensions: cost, schedule, contract health, or cross-cutting textual analysis. All but the Rayleigh Function metric are implemented in our research prototype; we experimented with a Rayleigh model but have not integrated it into PIERCE. Table 2.2 provides further details on the data loaded into PIERCE for these indicators and the relative risk indicators listed in Table 1.2 as described in Anton et al., 2022.

Table 2.1. Performance Indicators and Associated Data

Metric	Description	Data
Cross-cutting metrics		
Sentiment analysis	Assess whether the summary green, yellow, or red ratings for a report appear better than the text of MAR assessments	MARs
News report analysis	Filters news reports that appear to indicate performance issues or situations that could lead to future performance issues	News feed
PARs	Program status reports for major programs from the DCMA	PARs
Rayleigh Function ^a	Predicts the final costs and schedule by curve-fitting EV data values to date to a prototypical Rayleigh curve shape	EV (level 1)
Cost metrics		
CPI ^b	This measures the amount of completed work for every unit of planned cost spending (i.e., the inverse of actual spending to date relative to scheduled spending to date)	EV (level 1)
Cumulative CPI versus TCPI _{EAC}	Measures the amount of work that must be done for every dollar spent to meet the current estimate at completion (EAC) (i.e., the difference between cost efficiency to date and the needed efficiency to meet the EAC)	EV (level 1)
Range of IEACs	Estimates four different IEACs under the following assumptions: • Future cost performance will be the same as all past cost performance • Future cost performance will be influenced by past schedule performance • Future cost performance will be influenced by past schedule and cost performance • Similar to previous assumption, except increased weight is placed on CPI	EV (level 1)
COTG	Measures the cost over the current contract cost target (i.e., overruns from a certain perspective)	EV (level1)
Schedule metrics		

Metric	Description	Data
SPI ^b	Measures the fraction of budgeted work actually performed to date	EV (level 1)
ES versus actual duration (AD)	Measures the amount of planned schedule earned to date (i.e., the amount of time that was originally planned to take to reach the current level of work performed) compared with the AD	EV (level 1)
SPI _t	Measures the average percentage of scheduled work actually being completed to date	EV (level 1)
Schedule slippage indicator (SPI _t vs. TSPI _{ed})	Estimates the difference between progress schedule performance to date (SPI _t) and remaining planned efficiency (i.e., does the current remaining schedule reflect the same schedule slippage rates?)	EV (level 1)
IECD _{es}	Generates a predicted project completion date assuming future schedule slippage will match current slippage in percentage	EV (level 1)
Contract health metrics		
Work-content growth	Measures the growth in the amount of work content added to the contract to date	EV (level 1)

^a The Rayleigh Function has been modeled but not implemented in the research prototype.

^b This is a traditional EV management metric.

Table 2.2. Data Sources Incorporated in the Prototype

Name	Description and Uses	Data Loaded	Last Updated
FPDS-NG	Detailed information on all federal contracts with an estimated value of \$10,000 or more ^a ; includes contract numbers, prime contractor, and the primary place of performance	~140,000 contract details	September 2018
PMRT	Internal DAF reporting on programs and contracts; used to link U.S. Air Force programs and prime contract numbers	~500 U.S. Air Force programs	June 2020
USA Spending	FSRS data on subcontractors for federal contracts	27 gigabytes of data	April 2020
Data Universal Numbering System (DUNS)	Identify companies by DUNS number	~230,000 DUNS numbers	February 2019
Monster.com	Open job positions, including how long those positions have been vacant	Two snapshots from 2019 of ~405,000 job postings	April 2019
CapIQ	Commercial vendor system providing data about companies: financial metrics, M&As, suppliers, and news articles about the companies	735 gigabytes of data	August 2020
CapIQ News	Press articles and other media sources on companies, filtered by relevance	~9,000 news stories	August 2020
BLS	Unemployment and inflation rates by location. Used for places of performance for U.S. Air Force contractors as recorded in FPDS-NG		2019

Name	Description and Uses	Data Loaded	Last Updated
SAM.gov	Excluded companies (barred from receiving federal contracts)		
DAES	EV data snapshot of selected major defense acquisition program (MDAP) EV data	~26,000 program reports	November 2019
EVM Central Repository (EVM-CR)	EV data dating back to 2007, primarily for MDAPs	~43,000 program reports across 1,161 contract tasks	July 2020
DCMA PARs	Contractor PARs	640 program evaluations	May 2019

NOTE: ^a FPDS-NG, undated.

Note that data availability and completeness vary widely for each contract or situation. For example, EV data are generally not collected for DoD contracts valued at less than \$20 million (then-year dollars) but may be applied at the PM's discretion based on risk to the government (Defense Federal Acquisition Regulation Subpart 234.2, 2011). FPDS-NG data, on the other hand, are available for every federal contract with an estimated value of \$10,000 or more and thus are very comprehensive.

Taken together, we obtained at least some type of data on about 168,000 public and privately held companies in total. This gave us enough data to calculate at least one risk or performance measure for about 16,000 (primarily public) contractors.

Qualitative Indicators: Detecting Performance Issues Buried in Textual Reports

In addition to the quantitative indicators described below, we employed two indicators that assess unstructured, natural-language text to identify cases where problems may be described. The first employs machine learning and sentiment analysis to train a model to associate (or predict) a green, yellow, or red rating that summarizes an acquisition program's status as provided by the PM, PEO, or 3-Letter organization in the MARs. The second uses simpler keyword filtering of news to identify stories that discuss problems.

Sentiment Analysis and Modeling of Free-Text Assessments

In our prior experience, we found that textual reports on the status of a program sometimes contain indications of problems before the author is willing to make the program yellow or red. This led us to hypothesize the potential to train a natural language "sentiment analysis" model to learn the general correlation between free-text descriptions and the green, yellow, or red ratings. Then, if the model rates a text with a worse value than the human author did (e.g., yellow instead of green), we may have a useful indication of a concern or problem brewing in the program. In

other words, we hypothesize that some text reports in each of the MARs might indicate problems that are worse than the corresponding color (green, yellow, or red) evaluations reported in the same MAR.

To test this hypothesis, we trained an experimental model on 80 percent of the 3,530 DAF MARs available to us, tested it on the remaining MARs, and displayed cases in PIERCE where the model predicted a worse rating than the PM, PEO, or 3-Letter did.⁹ To help see whether the model is indicating concerns in a reasonable set of these cases, we asked subject-matter experts on our project team to review these indicated cases; they found that a surprising number of these contained items of concern in their professional judgment.

Figure 2.1 illustrates how this operates. Across the diagonal of the plot, the model and the human submitter had similar scores for most of the cases. However, there are a few cases marked as “candidates for due diligence” in which the sentiment analysis model identified a situation worthy of a yellow rating, whereas the human submitter scored it as green. These cases are displayed in PIERCE as potential earlier indicators of performance issues for the user to assess. Conceptually, this step would allow the busy manager to pick out the reports that might contain problematic language from among the numerous green reports.

Because we are trying to find earlier indicators of potential problems, we did not highlight the cases in which the model was more lenient than the human assessor, giving the text a better rating. In these cases, the human assessor had already flagged such problems with a worse rating, so the alert exists for the user to review. There could be future value in examining these cases of more-lenient grading if further research seeks to build better models of grading textual assessments of a contract’s or program’s status.

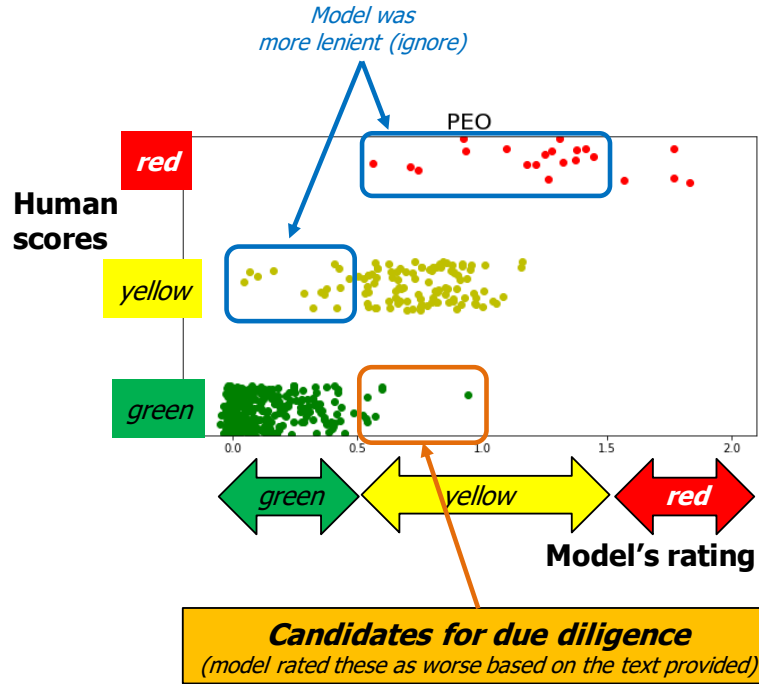
⁹ Although the training set of MARs may include cases where the color rating does not reflect the actual program performance (i.e., it may include cases where the color rating does not reflect the language provided), we assume the sample is large enough that most of the time the language matches the color rating, so the model has a possibility of learning the general relationship despite these less-frequent disconnects.

If, during the operation of the system, a PM used different language than most do (e.g., because of cultural differences), then the model and thus the system would be expected to make erroneous conclusions, e.g., failing to report some cases where the color reflects a better situation than reality. It may also report false positives (where the manager’s color accurately reflects the situation). This is one reason why the reported cases require further investigation by the user.

The model may be imperfect in analyzing the text (i.e., may report some false positives) and may fail to report cases where the situation is worse than the coloring indicates. This is another reason why the reported cases require further investigation by the user.

In the end, the model, at best, reflects what the training population of cases indicates about the relationship between language and color ratings (green, yellow, and red). The model and data (text and colors) should reflect the actual performance of the program, but the system does not further investigate whether other program performance metrics agree with the text and colors. We assume most acquisition professionals do relate green, yellow, and red ratings to actual program performance, and that their textual reports are true. However, data, model, and system do not assess ground truth but instead rely on this assumption as the basis for operation and alerting of cases where problems may be worse than indicated, in general, by the color provided in the MARs.

Figure 2.1. Output of the Sentiment Model



Details on the implementation of this sentiment analysis model are provided in Appendix A, which provides detailed descriptions of our approach to modeling the association between unconstrained natural-language text summarizing the status of a program and an overall summary rating of green, yellow, or red. Note that this preliminary research focused only on whether a model can identify some problematic cases in the set of green reports. Further research would be needed to determine whether there are other green cases that also have problems buried in the text (i.e., are false positives). Also, this is a relatively small data set; further research should examine a broader set of textual reports. We also do not know exactly how the model is identifying green cases with problematic language (e.g., whether it is simply picking up on problematic keywords (e.g., the approach we discuss in the next section about filtering news stories) or is somehow ascertaining sentiment in a more sophisticated manner. Finally, we do not know the effect of variation between the human MAR writers (e.g., depth or transparency of discussion) and the ability of the model to detect problems with green reports.

Detecting Potential Concerns in News Stories

With respect to using news filtering to find indicators of relative contractor risk (see Anton et al., 2022), we employed a keyword-based news filter to provide early indicators of potential performance problems. This news alert subsystem seeks to provide users with timely information about major events affecting contractors working on DAF programs within the context of other potential risk and performance indicators identified by PIERCE. It obtains news from CapIQ's

service for specified companies (S&P Global, undated). CapIQ's service is employed in PIERCE as a proof of concept; if PIERCE were to become operational, other news services, such as Proquest or Factiva, could be used in lieu of CapIQ's service or as supplemental data. Other newspapers, journals, or industry-specific magazines also could be added.

Only stories that contain at least one of the keywords listed in Table 2.3 are shown by PIERCE to the user. This keyword list was developed using the team's acquisition expertise.¹⁰ When observed in a news story, these keywords can be highlighted to help the reader see what potential performance concern may be indicated. Additional research could be conducted to refine the keywords list or to develop a sentiment analysis model that could focus on news stories of concern.

Table 2.3. Keywords for News Story Filter

Attrition	Fail*	Loss	Requalif*
Audit*	Foreign component	Lost	Resign*
Bad	Foreign supplier	M&A	Restructure
Bankruptcy	Fired	Merger	Rework
Breach	Firing*	Penalty	Scandal
Charge	Hack*	Poor performance	Schedule growth
Close*	Harassment	Price growth	Schedule slip
Compromis*	Halt	Price increases	Scrap
Cost growth	Hearing	Problem	Scrap rate
Cost increases	Hostile	Outsource*	Security breach
Counterfeit	High* overhead	Overhead increases	Security problem
Court	Immatur*	Overrun	Shortcoming
Damage	Incident	Phishing	Shortfall
Death	Inexperience*	Poor	Slip
Default on	Inflate	Privacy	Stop work
Deficien*	Inflated	Protest	Suit
Deficit	Injury	Rebuild	Surplus
Delay	Insider threat	Reconstruct	Takeover
Denied	Instability	Redesign	Tampering
Denial	Insufficient	Reorganiz*	Theft
Destroy	Judge*	Repair	Threat*
Destroyed	Lack	Replace*	Turnover
Destruction	Late	Replan	Unstable
Disbar*	Lawsuit	Reputation	Vulnerab*
Disqualified			

NOTE: Some of these keywords (such as *cost growth*, *poor performance*, *schedule growth*, and *overrun*) may be more associated with program-level issues. Further research is needed to determine whether these produce news stories with useful new information or whether such stories are duplicative of information already known to the PM and PEO. Asterisks are wild cards representing zero or more characters.

¹⁰ If we had a corpus of news stories partitioned (labeled) as either containing bad news or neutral-to-good news, then future research might employ clustering algorithms (machine learning) to obtain an improved keywords list or do the filtering directly.

In addition to identifying potential program concerns, this approach has been used recently to provide indicators of performance issues in the defense industrial base associated with the coronavirus disease 2019 (COVID-19) global pandemic.

Quantitative Indicators of Performance

Most of the quantitative indicators used in this project are summarized in the NDIA's guidebook on leading indicators (NDIA, 2017). Data systems at DAF headquarters or in the Office of the Secretary of Defense (OSD)—primarily level-1 EV data¹¹ from OSD's EVM-CR for Major Defense Acquisition Programs—contain what is needed to perform these calculations. Other indicators in the NDIA guidebook require data only available in the program office or possibly in other operational organizations, such as the DCMA. Next, we describe each metric along with the data used and rating thresholds.

Thresholds for Ratings

Quantitative thresholds define the rating boundaries between the quantitative metric values and the green, yellow, or red ratings; some measures also include an orange rating between yellow and red. Currently, we use a mix of different approaches for setting threshold levels.

In some cases, we use a population statistic—the standard deviation—to identify cases that are beyond the general behavior of the population in the metric; this approach has the benefit of alerting to cases in which the performance is beyond the norm of the population, with the yellow, orange, red ratings indicating an increasing distance from the norm. The drawback of this approach is that the norm might not necessarily be acceptable—therefore, some green ratings might be problematic but not revealed.

Alternatively, in other cases, we employ a fixed set of thresholds. This approach has the benefit of potentially revealing cases in which the norm is not acceptable. Next, we hypothesize and implement thresholds postulated by published literature or based on our expertise and insight. However, setting these thresholds is somewhat subjective because there are no absolute policies or guidance in the literature or in DoD policy as to what levels are acceptable for these metrics. For example, there is no a priori level of “goodness” or “badness” associated with CPI beyond the definition of CPI: “A CPI value greater than 1.0 indicates the work accomplished cost less than planned, while a value less than 1.0 indicates the work accomplished cost more than planned” (DoD, 2019a, p. 66). A CPI of 0.95 indicates that 95 percent of planned work is performed relative to plans. Whether that is acceptable involves a judgment about the quality of the original plan, the difficulty and uncertainty of the work involved, the tolerance of cost growth

¹¹ Level-1 EV data are for the entire effort (as opposed to smaller work packages in the EVM work breakdown structure).

relative to plans, and other mitigating factors.¹² Thus, setting thresholds is complicated and not generally fixed in the literature.

Also, the number of rating levels itself is still an open question. For example, whether there is value in having an intermediary orange rating between yellow and red for a particular metric is still not known. In some cases, we are experimenting with the inclusion of this intermediary level, but further research is needed to determine whether those additional levels are useful or not. Therefore, the values and approaches described next are initial starting points for further examination and possible adjustment.

EVM Terms

Given that many of the metrics discussed below use various EVM terms, we define them in in Table 2.4 and illustrate them in Figure 2.2.

Table 2.4. EVM Terms

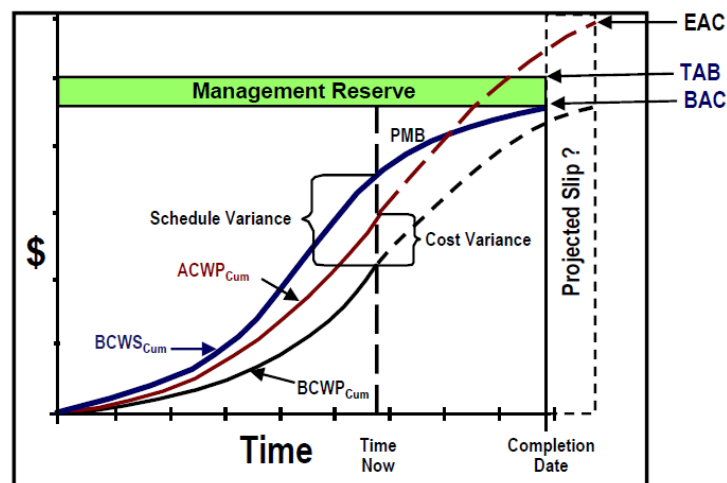
Acronym	EVM Term	Description
ACWP	Actual Cost of Work Performed	Cost actually incurred in accomplishing work performed to date
AUW	Authorized Unpriced Work	Work contractually approved, but not yet negotiated and definitized (priced and finalized on the contract)
BAC	Budget at Completion	The sum of all budgets for the contract through any given Work Breakdown Structure (WBS)/Organizational Breakdown Structure (OBS) level
BCWP	Budgeted Cost for Work Performed	Value of completed work in terms of the work's assigned budget; defined as the <i>earned value</i>
BCWS	Budgeted Cost for Work Scheduled	Time-phased budget plan for work scheduled; defines the planned value
CBB	Contract Budget Base	The sum of Negotiated Contract Cost (NCC) and AUW (i.e., the target cost of the contract)
EAC	Estimate at Completion	Estimate of total cost for the contract through any given level. EACs are usually designated by the entity that produced the estimate (e.g., the government PM, the contractor, or DCMA)
ETC	Estimate to Complete	Estimated cost of remaining work
LRE	Latest Revised Estimate	The contractor's EAC
MR	Management Reserve	Budget withheld by the contractor for managing known and unknown risks

¹² For comparison, there are some established cost and schedule thresholds in policy, but these are at the program level. For example, the Joint Staff threshold (called a “tripwire”) for revalidation is 10 percent program cost growth over the current acquisition program baseline or 25 percent over the original baseline program; for schedule slippage, the threshold is 12 months delay from initial operational capability (Joint Staff J-8, 2018, pp. C-13 and C-14). Nunn-McCurdy breach thresholds are 15 percent and 25 percent cost growth against current baselines for significant and critical breaches, respectively (10 U.S.C. Sec. 2433, 2020). However, these are also somewhat arbitrary and apply only at the acquisition program level, not at the contract level.

Acronym	EVM Term	Description
NCC	Negotiated Contract Cost	The contract price less profit or fee(s)
OBS	Organizational Breakdown Structure	The arrangement of the organization established to manage the resources tasked with performing the work on a specific contract or program.
OTB	Over Target Baseline	The sum of the CBB and additional budget approved for remaining work
PAC	Price at Completion	The EAC plus any (adjusted) profit or fee
PMB	Performance Measurement Baseline	The contract's time-phased budget plan
TAB	Total Allocated Budget	Sum of all budgets (NCC, CBB, or OTB) for work on the contract
TCPI	To-Complete Performance Index	The future cost efficiency needed from now forward to achieve a cost target (e.g., a BAC, LRE, or EAC)
WBS	Work Breakdown Structure	A hierarchical product-oriented division of program tasks depicting the breakdown of work scope for work authorization, tracking, and reporting purposes
WP	Work Package	Natural subdivision of control accounts. A WP is simply a task/activity or grouping of work. A WP is the point at which work is planned, progress is measured, and EV is computed.

SOURCES: DAU, 2018; DoD, 2019b.

Figure 2.2. EV Variables over Time



SOURCE: DAU, 2018.

NOTE: The *cost variance* is relatively straightforward, being the difference between the budgeted and actual cost of the work performed (to date). *Schedule variance*, on the other hand, in EV terms is less intuitive in that it is expressed in dollar rather than temporal terms, being the difference between the budgeted cost of work performed and scheduled to date (i.e., how much of the scheduled work that should have been completed by now has been performed in dollars).

Cost Metrics

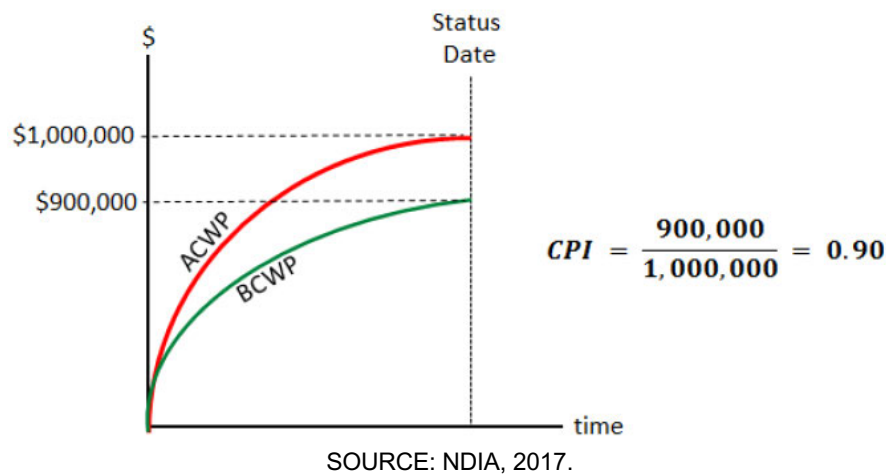
CPI

CPI is a traditional EVM cost metric we included here for completeness and comparison. It measures the amount of completed work to date for every unit of planned cost spending to date (i.e., the inverse of actual spending to date relative to scheduled spending to date). In other words, “CPI is a cost efficiency indicator; a CPI of 0.80 indicates that for every \$1.00 spent, only \$0.80 worth of planned work was accomplished” (National Aeronautics and Space Administration [NASA], 2019, p. 15-1). CPI is defined as follows:

$$CPI = BCWP/ACWP$$

Figure 2.3 illustrates this relationship.

Figure 2.3. CPI Notional Example



Thresholds. As an index, a CPI value of 1.0 is on track with plans. A value greater than 1.0 is favorable (execution to date is more efficient than planned) and a value less than 1.0 is unfavorable (execution to date is less efficient than planned). As a starting point, we implemented similar thresholds to those suggested by NDIA, 2017, but set the break between yellow and red at 0.9 (see Table 2.5). Setting these thresholds lower or higher (relative to 1.0) will generally trigger fewer or more contracts, respectively. As discussed earlier, there are no general DoD guidelines on CPI thresholds for warnings. Thus, continued experimentation with the prototype and a possible upgrade to the interface to allow the user to adjust the thresholds and the associated number of contracts with non-green ratings will permit customization as needed.

Table 2.5. CPI Metric Thresholds in the Prototype

CPI Value	Rating	Implication
≥1.0	Green	Efficiency is as planned or better
0.90–0.99	Yellow	Efficiency is somewhat worse than planned
<0.90	Red	Poor efficiency or planning

SOURCE: Adapted from NDIA, 2017.

NOTE: Future research is needed to determine whether these are the best values or whether thresholds should be set by the user.

NDIA, 2017, identified some possible management questions when this indicator is triggered (e.g., is CPI trending up or down? What is driving these trends? What can the PM do to help recover? Are such actions reasonable and realistic?).

Cumulative CPI vs. $TCPI_{EAC}$

In this efficiency measure, we compare the amount of work that must be done for every dollar spent to meet the current EAC (see NDIA, 2017). That is, we compare the difference between cost efficiency to date (CPI) and the needed efficiency to meet the estimate at completion—the TCPI to meet the EAC ($TCPI_{EAC}$). The latter (NDIA, 2017; NASA, 2019) is defined as

$$TCPI_{EAC} = (BAC - BCWP) / (EAC - ACWP)$$

The absolute difference between these two indexes is then calculated to aid the comparison:

$$|CPI - TCPI_{EAC}|$$

Thresholds. As a starting point, we used a threshold of 0.10 as the initial breaking point as suggested by NDIA, 2017, with higher breaking points for higher values (see Table 2.6), but as NDIA also points out, these thresholds can be adjusted. Figure 2.4 illustrates how a 0.10 band around the CPI could indicate a concern as the difference between these two indexes diverges. In this example, the metric would be green through May but then turn yellow in June and be on the brink of turning orange later in November.¹³ Here, we added an orange rating between yellow and red. This is somewhat arbitrary but was done to begin experimenting with the value of

¹³ As with CPI, these threshold values are somewhat arbitrary and can be adjusted based on the user's surveillance requirements. For example, the GAO mentions a rule of thumb of 0.05 as an initial threshold (GAO, 2020a, p. 258). Further experimentation is needed to determine how many contracts are triggered relative to the desired levels. Future upgrades to the prototype interface could also allow the user to adjust these levels.

having an additional threshold level. Further research is needed to determine the value of an additional rating level and the threshold values themselves.

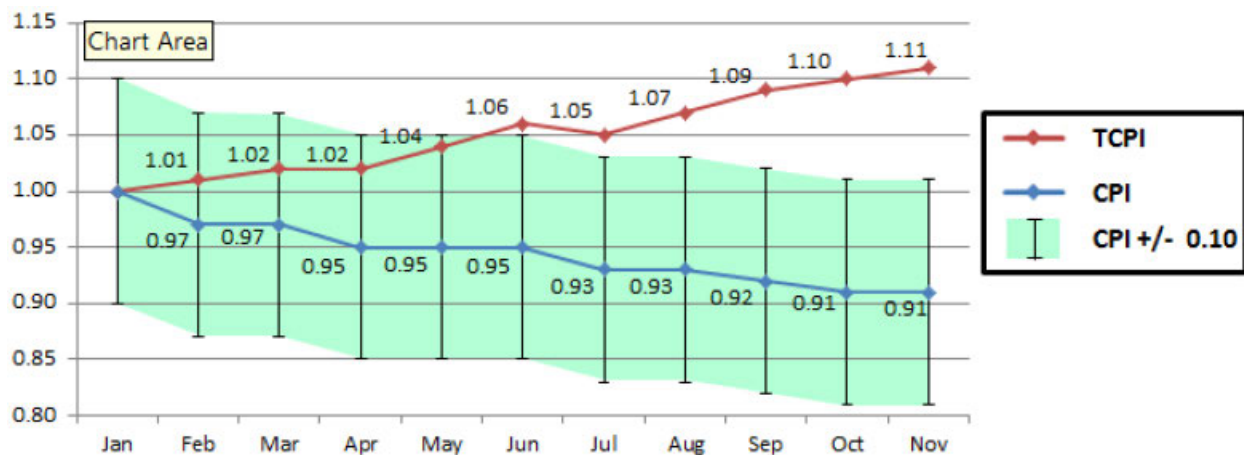
Table 2.6. CPI vs. TCPI_{EAC} Metric Thresholds in the Prototype

$ CPI - TCPI_{EAC} $ Value	Rating	Implication
<0.10	Green	Needed future efficiency is close to the cost efficiency seen to date
0.10–0.19	Yellow	Needed future efficiency is high compared with that seen so far
0.20–0.29	Orange	Needed future efficiency is very high compared with that seen so far
≥ 0.30	Red	Needed future efficiency is radically higher than seen to date.

SOURCE: Adapted from NDIA, 2017.

NOTE: Future research is needed to determine whether these are the best values or whether thresholds should be set by the user. Also, the benefits of having an intermediary *orange* rating remain in question.

Figure 2.4. CPI vs. TCPI_{EAC} Notional Example



SOURCE: NDIA, 2017.

NDIA, 2017, identified some possible management questions that might be asked when examining these two metrics (e.g., are they diverging, which may indicate an unrealistic EAC or warrants explaining why future cost efficiency is expected to be different than efficiency to date; If TCPI_{EAC} is close to 1.0, is that accurate, or might it not reflect challenging tasks later in the effort? If CPI is below 1.0 and TCPI_{EAC} above 1.0, are future ETCs being artificially shrunk?).

Range of Independent Estimates at Completion

These metrics use four different approaches to estimate EACs independent of those produced by the PM, contractor, or any other organization. Thus, they provide four different ways to see how well the PM and contractor EACs compare with those that follow these quantitative

approaches. Figure 2.5 illustrates how the IEAC range can be used to determine whether the PM’s or contractor’s EAC is in range, optimistic, or pessimistic compared with the set of four IEACs.

The four metrics follow these assumptions (NDIA, 2017):

1. Future cost performance will be the same as all past cost performance
2. Future cost performance will be influenced by past schedule performance
3. Future cost performance will be influenced by past schedule and cost performance
4. Similar to the previous assumption, except increased weight is placed on CPI.

$$IEAC_1 = ACWP + (BAC - BCWP_{cum})/CPI$$

$$IEAC_2 = ACWP + (BAC - BCWP_{cum})/SPI$$

$$IEAC_3 = ACWP + (BAC - BCWP_{cum})/(SPI * CPI)$$

$$IEAC_4 = ACWP + (BAC - BCWP_{cum})/(0.2 * SPI + 0.8 * CPI)$$

where

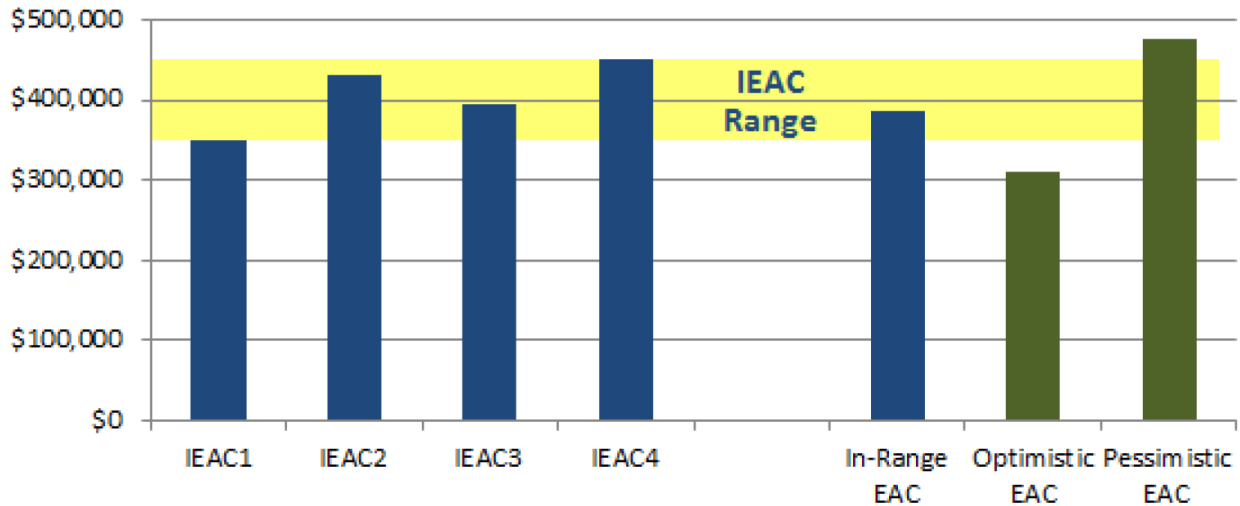
$BCWP_{cum}$ is the cumulative BCWP to date, and

$BAC - BCWP_{cum}$ is the remaining amount of budgeted cost to execute.

Thus, these equations take the ACWP and scale the remaining budgeted cost according to four different assumptions of efficiency: cost performance to date (CPI), schedule performance to date (SPI), the product of both cost and schedule performance to date (CPI and SPI), and a combination of cost and schedule performance but weighted toward cost performance (CPI). Here, the idea is to use different metrics to try and detect performance problems that might not be revealed in a single metric.¹⁴ Both Lipke, 2004, and GAO, 2020a, discuss the periods in which different metrics might be more insightful in the life of the contract.

¹⁴ Further discussion about these various metrics and their reasoning can be found in Lipke, 2004; NDIA, 2017, pp. 45–47; and GAO, 2020a, pp. 265–268.

Figure 2.5. Range of IEACs (Notional Example)



SOURCE: NDIA, 2017.

Thresholds. First, we calculated all four IEACs to determine the IEAC range. As a starting point, we prototyped the thresholds in Table 2.7 for comparing the PM’s best estimate EAC with the range of IEACs. We added a 10-percent boundary to the range of IEACs as the thresholds for a green rating.¹⁵ Also, we expect users will be more interested in predictions indicating the PM’s EAC is lower than the IEAC range than when it is higher. Thus, we show a broader color rating range extending to orange and red as the PM’s EAC gets further below range. On the other side, there is the case where the PM’s EAC appears higher than predicted by all four IEACs. However, the PM might be indicating conditions not being detected by the IEACs, so we grouped all these together as a yellow rating.

¹⁵ This approach is somewhat more liberal than the range used by NDIA, 2017, which suggested using the lowest and highest IEAC as the thresholds for a green rating. We felt that being close to the lowest or highest IEAC is probably reasonable to still have a green rating. Nevertheless, further analysis and use of PIERCE are needed to determine whether this is warranted and where to best set the threshold breaks. For example, if the user wants more contracts to be triggered, then tightening the thresholds could be warranted.

Table 2.7. Range of IEAC Thresholds in the Prototype

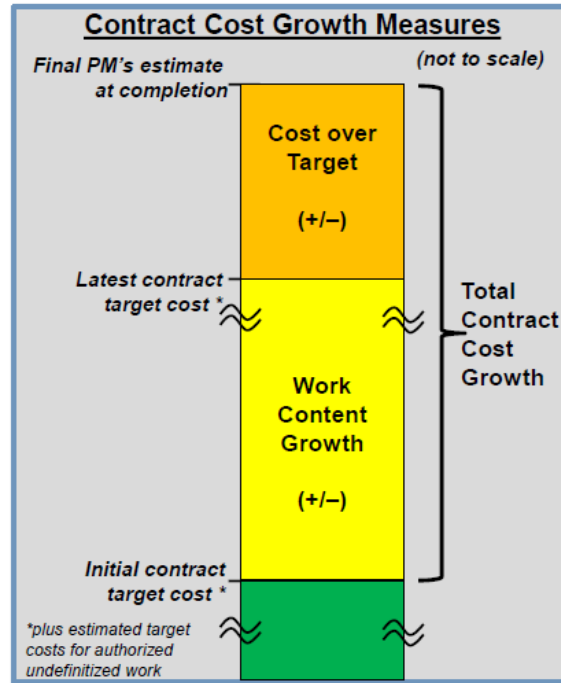
PM's EAC	Rating	Implication
Within 10% of the IEAC range	Green	EAC is consistent with performance to date—or almost so
Outside the IEAC range by > +10%	Yellow	EAC is pessimistic compared with performance to date
Outside the IEAC range by –20% to –10%	Yellow	EAC is somewhat optimistic compared with performance to date
Outside the IEAC range by –50% to –20%	Orange	EAC is very optimistic compared with performance to date
Outside the IEAC range by below –50%	Red	EAC is highly optimistic compared with performance to date

NOTE: Future research is needed to determine whether these are the best values or whether thresholds should be set by the user. Also, the benefits of having an intermediary orange rating remain in question.

NDIA, 2017, identified some possible management questions that might be asked when this indicator is triggered (e.g., what might be causing these calculated IEACs to differ significantly from the projected EAC? Has recent cost efficiency changed significantly from the overall efficiency, and, if so, would the recent efficiency be a better basis for calculating IEACs' COTG?).

This indicator measures the cost over the current contract cost target—i.e., “overruns,” from a certain perspective (Under Secretary of Defense for Acquisition, Technology, and Logistics [USD(AT&L)], 2013, pp. 37–39). Figure 2.6 illustrates that total contract cost growth can be split between total contract cost growth (COTG) and work content growth (WCG) using level-1 EV data. These metrics have been used extensively in DoD analyses of the performance of the defense acquisition system (USD[AT&L], 2013, 2014, 2015, 2016). Whether COTG or WCG is justifiable cannot be determined *a priori*, but they help identify the general sources of cost growth and thus can inform further investigation by the PM and others, especially if these growths are unexpected and not in line with plans. If COTG is greater than zero, it indicates that the contractor is not meeting the negotiated contract target cost; whether that is a major problem depends on such factors as the reliability of the target cost and its associated uncertainties and risks. We will discuss WCG separately later in the chapter.

Figure 2.6. Relationship Between Total Contract Cost Growth, COTG, and WCG



SOURCE: USD(AT&L), 2013.

NOTE: Initial contract target cost is the initial CBB.

COTG is calculated using the following equation:

$$COTG = [EAC_{PM}(t) - CBB(t)] / CBB(0)$$

where:

$EAC_{PM}(t)$ is the PM's EAC at time t , and

$CBB(t)$ is the CBB at time t .

Thresholds. As a starting point for the proof of concept, we used thresholds in increments of 10 percentage points for PIERCE (see Table 2.8). Further use of PIERCE may suggest different thresholds (e.g., based on the statistical distribution across contracts) to adjust the level of triggering.

Table 2.8. COTG Metric Thresholds in the Prototype

COTG Value	Rating	Implication
<10%	Green	Negative or negligible cost growth relative to target
10–20%	Yellow	Modest cost growth relative to target
>20%	Red	Moderate to high cost growth relative to the current target

NOTE: Future research is needed to determine whether these are the best values or whether thresholds should be set by the user.

Conceptually, a high COTG greater than zero could lead to possible management questions (e.g., what is driving the contractor to miss the contract target price? Was the target price unrealistic or full of uncertainties, or has the contractor not performed satisfactorily? If the latter, why has the contractor not performed, and are there steps that the government or contractor PM can take to improve performance? If the target price had high uncertainties, are there framing assumptions underlying the cost estimate that proved false, and what are the larger implications of those assumptions being false? If the contract is part of a larger program, does a high COTG on this contract put the program at risk of program-level cost growth?).

Schedule Metrics

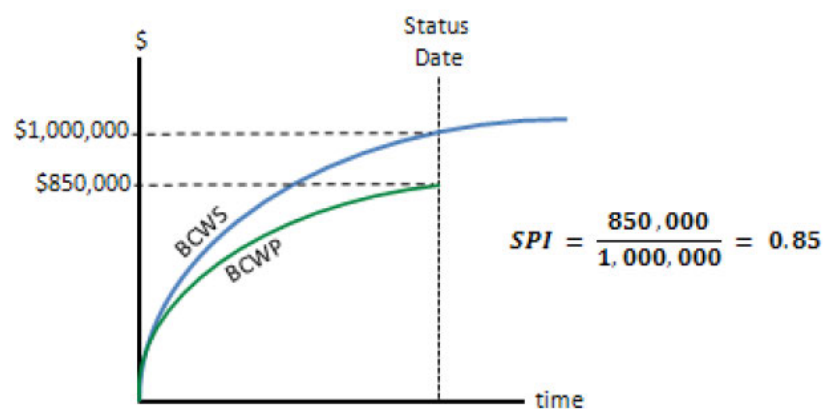
SPI

SPI is another traditional EVM cost metric we included here for completeness and comparison. It measures the fraction of budgeted work actually performed to date. In other words, “SPI is a schedule efficiency indicator; an SPI of 0.80 indicates that for every \$1.00 planned, only \$0.80 worth of planned work was accomplished” (NASA, 2019, p. 15-2). SPI is defined as follows (DoD, 2019b; DAU, 2018; NASA, 2020):

$$SPI = BCWP / BCWS$$

Figure 2.7 illustrates this relationship.

Figure 2.7. Notional Example of SPI



SOURCE: NDIA, 2017.

Thresholds. As an index, an SPI value of 1.0 is on track with plans. A value greater than 1.0 is favorable (execution to date is faster than planned) but a value less than 1.0 is unfavorable (execution to date is slower than planned). As a starting point, we implemented similar

thresholds to those suggested by NDIA (2017) but set the break between yellow and red at 0.9 (see Table 2.9).

Note that, by definition, the value of SPI is 1.0 for all completed contracts no matter how late the contract completion may be and thus loses its predictive value later in the contract—generally in the last third of the contract (NDIA, 2017; GAO, 2020a). Thus, we stop reporting SPIs in PIERCE when two-thirds of the contract schedule has elapsed to avoid giving late-running contracts an artificially high SPI rating toward the end of the contract. Further research is needed to determine whether this initial two-thirds point is a good stopping point for examining SPI (i.e., mathematically or statistically, at what point does SPI’s value become unrepresentative of schedule delays?).

Table 2.9. SPI Metric Thresholds in the Prototype

SPI Value	Rating	Implication
If $t < 0.67$:		
≥ 1.0	Green	Efficiency is as planned or better
0.90–0.99	Yellow	Efficiency is somewhat worse than planned
< 0.90	Red	Poor efficiency or planning
If $t \geq 0.67$	No rating	

SOURCE: Adapted from NDIA, 2017.

NOTE: Future research is needed to determine whether these are the best values or whether thresholds should be set by the user.

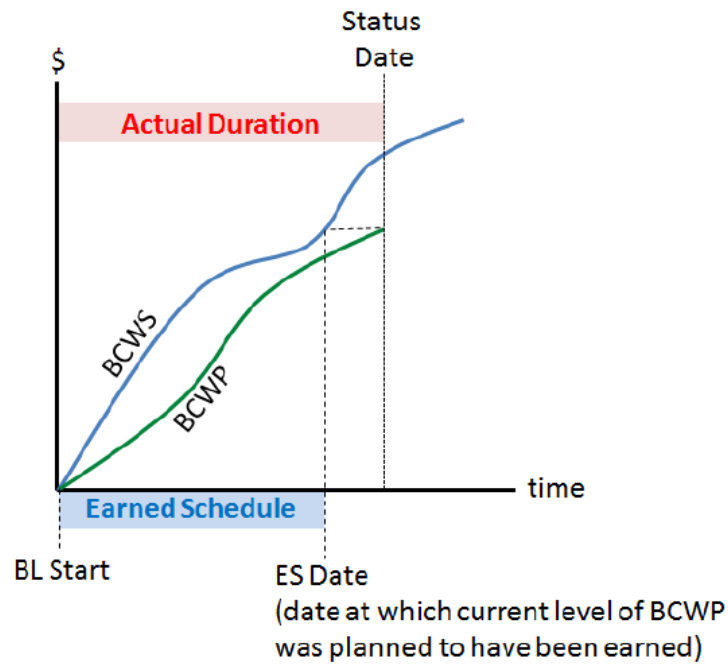
NDIA, 2017, identified some possible management questions that might be asked when this indicator is triggered (e.g., is SPI trending up or down, and why? Do the SPI values reflect performance on the critical path for the contract, and, if no, why? Which WBS elements are performing the worst, and why? If SPI is below 1.0, are there steps that the PM can take to improve it? Is SPI performance to date in line with other performance indicators?).

SPI_t

This measure calculates the average percent of scheduled work actually completed to date. In some sense, SPI_t is SPI but measured in time rather than in dollars. It compares ES—the amount of planned schedule earned to date (i.e., the amount of time originally planned to take to reach the current level of work performed)—with the AD. Conceptually, this compares the schedule-related complement with EV—that is, how much schedule you have earned to date—with how long the work has actually taken (Lipke, 2003, 2014; NASA, 2020; NDIA, 2017, International Standard 21508, 2018).

Figure 2.8 illustrates how ES and AD are calculated. ES is the duration on the BCWS plot to reach the current level of BCWP measured from program start.

Figure 2.8. Calculating ES and AD



SOURCE: NDIA, 2017.

SPI_t is defined as the ratio of ES to AD (NASA, 2020; NDIA, 2017):

$$SPI_t = ES / AD .$$

Thresholds. Table 2.10 lists the thresholds we used to alert deviations of ES compared with AD. If SPI_t is 1.0 or larger, then the contract is on or ahead of schedule. SPI_t below 1.0 indicates that the work is behind schedule. We used 0.90 as the threshold between yellow and red.

Table 2.10. SPI_t Metric Thresholds in the Prototype

SPI_t Value	Rating	Implication
0.98–1.05	Green	Efficiency is exceptional or planning may have been padded
0.95–0.98 or >1.05	Yellow	Efficiency is somewhat worse than planned or too good (could be risky)
0.90–0.95	Orange	Efficiency is worse than planned or too good
<0.90	Red	Poor efficiency or planning

NOTE: Future research is needed to determine whether these are the best values or whether thresholds should be set by the user. Also, the benefits of having an intermediary *orange* rating remain in question.

NDIA, 2017, identified similar management questions to consider for SPI_t as for SPI (e.g., is SPI_t trending up or down, and why? Do the SPI_t values reflect performance on the critical path for the contract, and, if no, why? Which WBS elements are performing the worst, and why? If SPI_t is below 1.0, are there steps that the PM can take to improve it? Is SPI_t performance to date in line with other performance indicators?). Here, the benefit is that SPI_t remains a useful metric through the end of the contract.

Schedule Slippage Indicator (SPI_t vs. $TSPI_{ed}$)

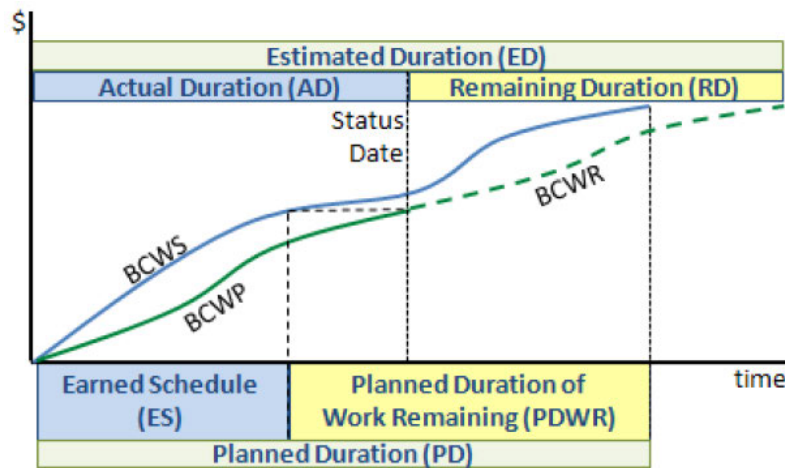
This metric estimates the difference between SPI_t and remaining planned efficiency (i.e., does the current remaining schedule reflect the same schedule slippage rates?) (Lipke, 2014; NDIA, 2017). Specifically, we compare SPI_t with $TSPI_{ed}$. The latter of these is the future schedule efficiency needed to achieve the current ED:

$$TSPI_{ed} = PDWR / RD$$

where

PDWR is the (originally) *planned duration of work remaining*, and
RD is the *remaining duration* based on current estimates (see Figure 2.9).

Figure 2.9. Illustration of PDWR and RD



SOURCE: NDIA, 2017.

In other words, $TSPI_{ed}$ measures the amount of scaling of the PDWR in current estimates. Thus, by comparing SPI_t with $TSPI_{ed}$, we can see whether this scaling is comparable with the efficiency seen to date (SPI_t).

Thresholds. Table 2.11 lists the thresholds we used when comparing the differences between SPI_t and $TSPI_{ed}$, using thresholds that are 10 percentage points apart.

Table 2.11. SPI_t vs. $TSPI_{ed}$ Metric Thresholds in the Prototype

$SPI_t - TSPI_{ed}$	Rating	Implication
-10% to 10%	Green	$TSPI_{ed}$ is roughly in line with schedule efficiency seen to date
-20% to -10% or 10% to 20%	Yellow	$TSPI_{ed}$ may be overly pessimistic compared with efficiency to date $TSPI_{ed}$ may be overly optimistic compared with efficiency to date
-30% to -20% or 20% to 30%	Orange	$TSPI_{ed}$ may be very pessimistic compared with efficiency to date $TSPI_{ed}$ may be very optimistic compared with efficiency to date
< -30% or > 30%	Red	$TSPI_{ed}$ may be highly pessimistic compared with efficiency to date $TSPI_{ed}$ may be highly optimistic compared with efficiency to date

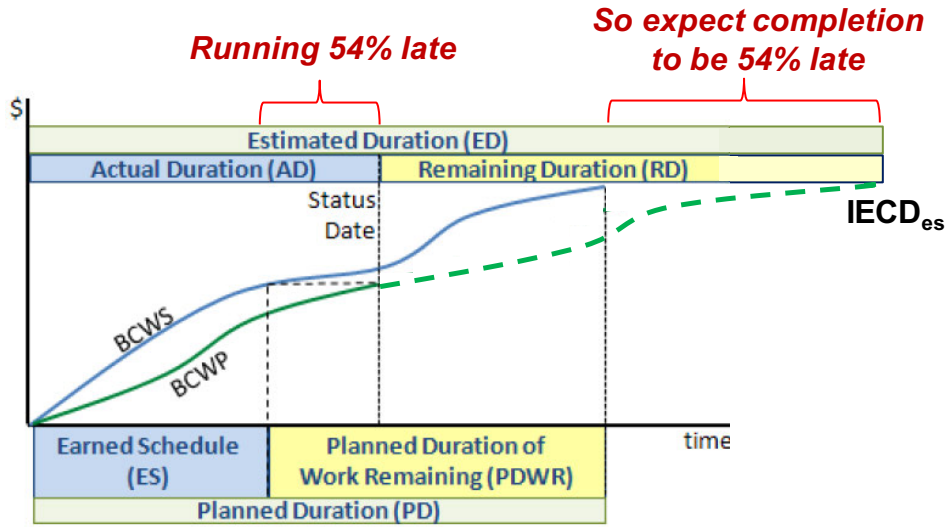
NOTE: Future research is needed to determine whether these are the best values or whether thresholds should be set by the user. Also, the benefits of having an intermediary *orange* rating remain in question.

NDIA, 2017, lists some possible management questions that might be asked when this indicator is triggered (e.g., what factors might be causing future schedule efficiency to differ from what has been demonstrated to date? Change in resources and staffing? Change in facilities/capacity? Change in technology? Change in plan [OTB]?).

IECD_{es}

This metric generates a predicted project completion date assuming future schedule slippage will match current slippage in percent. This provides an independent value to compare with the PM's or contractor's expected completion data. Here, we use the schedule progress rate to date ($SPI_t = ES/AD$) to scale the schedule to the end of the contract (NDIA, 2017). This is illustrated in Figure 2.10.

Figure 2.10. Illustration of $IECD_{es}$



SOURCE: Adapted from NDIA, 2017.

$IECD_{es}$ is defined as follows (NDIA, 2017):

$$IECD_{es} = StatusDate + PDWR * AD / ES$$

which is equivalent to

$$IECD_{es} = StartDate + PD * AD / ES$$

where:

StatusDate is the date of the latest EV report

PDWR is the planned duration of work remaining

StartDate is the date of the beginning of the effort, and

PD is the planned duration of the entire contract effort.

The absolute difference between the durations based on the $IECD_{es}$ and the contractor's current estimated completion date (ECD_{CTR}) is then calculated to see how different they are:

$$ECD_{CTR_IECD_error} = | (IECD_{es} - StartDate) - (ECD_{CTR} - StartDate) | / PD$$

For example, if the $IECD_{es}$ is eight years after the start date, the ECD_{CTR} is six years after the start date, and the *PD* is also six years, then the $ECD_{CTR_IECD_error}$ is $|8-6|/6 = 33.3$ percent.

Thresholds. Table 2.12 lists the thresholds we used when comparing the duration differences calculations, using thresholds that are 10 percentage points apart.

Table 2.12. ECD_{CTR} vs. IECD_{es} Metric Thresholds in the Prototype

<i>ECD_{CTR}_IECD_{error}</i>	Rating	Implication
<10%	Green	ECD _{CTR} is roughly in line with IECD _{es}
10–20%	Yellow	ECD _{CTR} is somewhat out of line with IECD _{es}
20–30%	Orange	ECD _{CTR} is very out of line with IECD _{es}
>30%	Red	ECD _{CTR} is far out of line with IECD _{es}

NOTE: Future research is needed to determine whether these are the best values or whether thresholds should be set by the user. Also, the benefits of having an intermediary orange rating remain in question.

NDIA, 2017, lists some possible management questions that might be asked when this indicator is triggered (e.g., what might be causing the calculated IECD_{es} to differ from the contractor's ECD? Are there changes or improvements afoot? Is the performance along the critical path similar to the overall schedule performance? Has recent performance been significantly better or worse than the overall performance?).

Joint Cost and Schedule Metrics

We explored one metric that jointly predicts final cost and schedule using a Rayleigh Function. This metric was developed but there was not sufficient time on this project to integrate it into PIERCE; we indicate later that integration could be considered in future work.

Rayleigh Function

This metric predicts final contract cost and schedule by curve fitting EV data values to a prototypical Rayleigh curve shape. In a broad sense, the strategy is to leverage data on how cumulative costs (as measured by ACWP) for acquisition contracts tend to unfold over time as projects ramp up at the beginning and ramp down at the end. Published research demonstrates that cumulative costs—particularly for RDT&E contracts—tend to follow an S-shaped curve that is well described by the Norden-Rayleigh Function, which consists of two parameters that determine its shape and scale.¹⁶ Hence, once enough data are available to reliably estimate the two parameters, one can construct a forward-looking curve representing the projected path of future costs through contract completion. The result is an independent, data-driven prediction of final cost and schedule that can be compared with internal estimates to provide an early indicator of prospective cost and schedule overruns.

The Norden-Rayleigh Function is given by

¹⁶ See, for example, Watkins, 1982; Davis, Christle, and Abba, 2009; Tracy and White, 2011; Brown et al., 2015; and Yang et al., 2019.

$$c(t) = d (1 - e^{-\alpha t^2}),$$

where:

- $c(t)$ denotes the ACWP as of month t ,
- α signifies the shape parameter, and
- d represents the scale parameter.

The shape parameter—as its name suggests—governs the shape of the function; in other words, the precise nature of the “S” shape. In Appendix B, we demonstrate how this parameter is numerically related to the estimated total duration of the contract. The scale parameter is related to the height of the curve and, therefore, is directly linked to the contract’s expected final cost. As the above equation highlights, the key data that feed the function are ACWP by month, where month t represents the number of months that have elapsed since the contract began. Once enough reports have accrued for the curve to begin taking shape, we estimate α and d to obtain the specific Rayleigh Function that best fits the data. We then extrapolate based on the estimated function to predict final cost and schedule. We update our parameter estimates and predictions as each new month of data becomes available; the predictions should improve as more data accumulate. Given that there is variability surrounding the final cost and schedule point estimates, we also estimate confidence regions to capture the degree of uncertainty accompanying our forecasts.

There are two primary obstacles that we must overcome to improve the fidelity of our estimates. First, not every contract will—or should—follow a Rayleigh curve.¹⁷ Accordingly, we must develop a procedure for categorizing contracts that adequately fit the curve and those that do not. Second, we must deal with breaks in the curve. For example, if major work content is added partway through the contract, our estimates will be biased in the absence of adjustment to explicitly account for the change. In effect, we would be fitting a single Rayleigh curve that treats pre- and post-work content growth observations as coming from a contract with a single, fixed amount of final target work.

We address the first concern by developing an automated rule for keeping or excluding contracts based on whether they meet a statistical goodness of fit threshold after sufficiently

¹⁷ The Rayleigh model does not apply to all contracts. Generally, it tends to apply more to development contracts where there is an initial ramp-up and a final closeout on the effort, yielding a cumulative S-curve. Luckily, these are also the type of contracts that tend to report EV data. Other contracts, such as production contracts where there is prior and future activity on different contracts, will not be expected to have this kind of start-up and shutdown behavior and thus not exhibit the shape of the Rayleigh Function. See Davis, Christle, and Abba, 2009, for further discussion.

many cost reports have accrued.¹⁸ We address the second concern by fitting a new curve any time we detect a jump in the CBB greater than or equal to 20 percent. Appendix B presents further information on these adjustments and also describes how the resulting model performs on a sample of completed contracts. On the whole, the model performs reasonably well at predicting final cost and schedule, although there is still variability in the estimates. Future work could refine the rules for excluding contracts, refit curves based on work content growth, and examine alternative or potentially more-sophisticated approaches for constructing confidence intervals.

These cost and schedule forecasts could be translated into predictive (leading) performance indicators by using thresholds with separate ratings for cost and schedule. One candidate set of threshold levels are those used for significant and critical Nunn-McCurdy breaches, which are triggered at 15 percent and 25 percent, respectively (10 U.S.C. 2433, 2020). Using this approach, a contract's cost or schedule risk could be classified as green if the Rayleigh estimate is within 15 percent of the internal estimate—the same as the Nunn-McCurdy significant threshold.¹⁹ Given that our Rayleigh algorithm produces some oscillation, we suggest triggering based on a combination of persistency and percentage from the internal cost estimate.²⁰ Thus, a contract's cost or schedule risk could be classified as yellow if the Rayleigh estimate has been at least 15 percent greater than the internal estimate for at least two consecutive reports. At the higher level, a contract's cost or schedule risk indicator could be classified as red if the Rayleigh estimate has been at least 25 percent greater than the internal estimate for at least two consecutive reports. This possible classification scheme is summarized in Table 2.13.

Other alternatives might be to use a temporal threshold for schedule. For example, 10 U.S.C. 2431a (2016) defines a “significant change to the schedule” as a “schedule delay greater than six months.”

¹⁸ Specifically, we calculate R^2 for each contract after ten months of cost reports have accumulated. If R^2 is below 0.8, the curve fit is deemed insufficient, and the contract is excluded from the analysis. The R^2 calculation stems from the regression estimating the two parameters of the Rayleigh curve. R^2 is a commonly used measure that captures how well our estimated curve aligns with the empirical data for each contract. See Davis, Christle, and Abba, 2009.

¹⁹ We use the PM's EAC as the internal cost estimate and estimated completion date as the internal schedule estimate.

²⁰ Some oscillation may be the result of outliers in the data. Presently, outliers may also cause poor model fit (low R^2) that results in certain contracts being dropped from the analysis. If an outlier is the result of inaccuracy in the data, it is possible that these contracts should not be dropped. We recommend examining outliers and correcting the data when possible so that the methodology can be applied to all suitable contracts.

Table 2.13. Possible Rayleigh Cost and Schedule Assessment Criteria

Rating	Criteria
Green	Percentage difference on most recent report is below 15 percent
Yellow	Percentage difference on the past two or more consecutive reports is between 15 percent and 25 percent
Red	Percentage difference on the past two or more consecutive reports is at least 25 percent

NOTE: Future research is needed to determine whether these are the best values or whether thresholds should be set by the user.

Conceptually, significant differences between the Rayleigh predictions and current cost and schedule estimates by the government PM or contractor might lead to possible management questions (e.g., if the Rayleigh model indicates a slow start, then what is the cause? Is it a lack of staff or slow initial funding? Is there a justifiable reason to believe that the contractor can catch up later after the slow start? Is there reason to believe that the contract should not exhibit an S-curve profile and thus the Rayleigh model is not appropriate in the first place?).

Additional details on our Rayleigh Function modeling are provided in Appendix B.

Contract Health Metric

We now discuss two metrics that relate to the general health of a contract. The first examines the stability of the work content, and the second examines the stability of the baseline of budgeted costs scheduled.

WCG

This indicator measures the amount of WCG added to the contract to date through contract modifications (USD[AT&L], 2013, pp. 37–39). Significant WCG might indicate instability in the contract and the larger program that it supports. WCG could also reflect a planned incremental addition of work onto a contract, so additional insight from the program office is important. Figure 2.6 illustrates the relationship between total contract cost growth, WCG, and COTG. As with COTG, WCG has been used extensively in DoD analyses of the performance of the defense acquisition system (e.g., USD[AT&L], 2013, 2014, 2015, and 2016). These reports tend to report the facts of WCG and attributes that distinguish it from COTG rather than trying to determine whether a specific WCG is good or bad; COTG may arguably be more of a concern, but even there some caution and insight is necessary to understand what is causing the cost growth over the negotiated target.

WCG is calculated by comparing the current CBB with the initial CBB:

$$WCG = [CBB(t) - CBB(0)] / CBB(0) ,$$

where

$CBB(t)$ is the CBB at time t .

Thresholds. As with COTG, we used thresholds in increments of 10 percentage points as a starting point for WCG in PIERCE (see Table 2.14). Further use of PIERCE might suggest different thresholds (e.g., based on the statistical distribution across contracts) to adjust the level of triggering.

Table 2.14. WCG Metric Thresholds in the Prototype

WCG Value	Rating	Implication
<10%	Green	Negative or negligible work-content growth
10–20%	Yellow	Modest work-content growth
>20%	Red	Moderate to high work-content growth

NOTE: Future research is needed to determine whether these are the best values or whether thresholds should be set by the user.

Although it is less clear whether WCG is necessarily a problem, NDIA, 2017, and our analysis identified some possible management questions for when this indicator is triggered (e.g., what is causing the changes in the baseline work? Are they to address new requirements or threats? Was the contract planned from the beginning to have work added over time? Does oversight management know about these changes, and do they approve? Are there budgetary issues associated with the increased work?).

Opportunities for Future Capabilities

Although these metrics, other than the Rayleigh Function, have been implemented in PIERCE, additional work in collecting more longitudinal data—especially from the FPDS-NG and either PMRT or the Defense Acquisition Visibility Environment (DAVE) with the Defense Acquisition Management Information Retrieval (DAMIR)—and exercising PIERCE to better understand the utility of these measures and the overall approach is warranted. Also, Table 2.15 lists some candidates we identified as promising new metrics that could be added to the system. Some of these would require obtaining access to data sources, such as DCMA’s Electronic Tools (eTools) system or OSD’s Procurement Integrated Enterprise Environment (PIEE).

Table 2.15. Candidate Metrics for Future Implementation in the Prototype

Metric	General Approach
Rayleigh Function	As discussed earlier, we explored approaches for curve-fitting Rayleigh Functions to sequential EV reports to project the cost and schedule EAC. Further work would be needed to implement and test these approaches in PIERCE.
DAES and SARs sentiment analysis	Apply the sentiment analysis model used in PIERCE on MAR text reports to free-text reports and ratings in DAES and SARs. ^a Also, report the MAR, DAES, and SAR rating values for every program to compare with the indicators in PIERCE.
Baseline revisions	Detect changes in the monthly, noncumulative BCWS budgeted costs to trigger due diligence of the cause (e.g., investigate whether revisions are from "contractual modifications, the use of management reserve, application of undistributed budget, replanning, or formal reprogramming" [NDIA, 2017, p. 96]).
Requirements volatility	This measure could examine the stability of the requirements as reported in OSD's DAVE/DAMIR system.
Bid protest reports	Bid protests to the GAO are not readily available at the PEO, DAF headquarters, or OSD levels. Our understanding is that all such protests may be archived in the PEE system, but further work would be needed to obtain access to PEE data and develop mechanisms to integrate these bid-protest data into PIERCE as an early indicator of program problems.
Parts demand fulfillment	DCMA eTools or other sources may have data on parts fulfillment, in which case fulfillments could be measured as an indicator of supply-chain performance.
Supplier acceptance rate	Likewise, DCMA eTools or other sources may have data on the acceptance (or rejection) rates for supplies being delivered to a prime contractor.
Supplier late starts	In addition, if supplier data, such as late starts, are available in eTools, then metrics could be developed if access to eTools was obtained.

^a Note, however, that preliminary research by Joseph and Sconion, 2020, did not find promising results so far for SAR-level correlations with final cost.

Summary

In this chapter, we identified several leading indicators of performance to provide early information to PMs and oversight leadership. Many of these were codified with actual data in the PIERCE to begin exploring their utility and what triggering thresholds might be useful. Other metrics were explored, and still others were identified as possible candidates for future upgrades or uses in other systems. In Chapter 3, we present an overview of the PIERCE prototype structure and user interface.

3. Prototype Structure and Approach

We implemented the metrics discussed in Chapter 2—combined with the relative risk indicator metrics discussed in our companion report (Anton et al., 2022)—in a functioning research prototype to test and demonstrate these ideas. In this chapter, we provide an overview of PIERCE structure and intended usage.

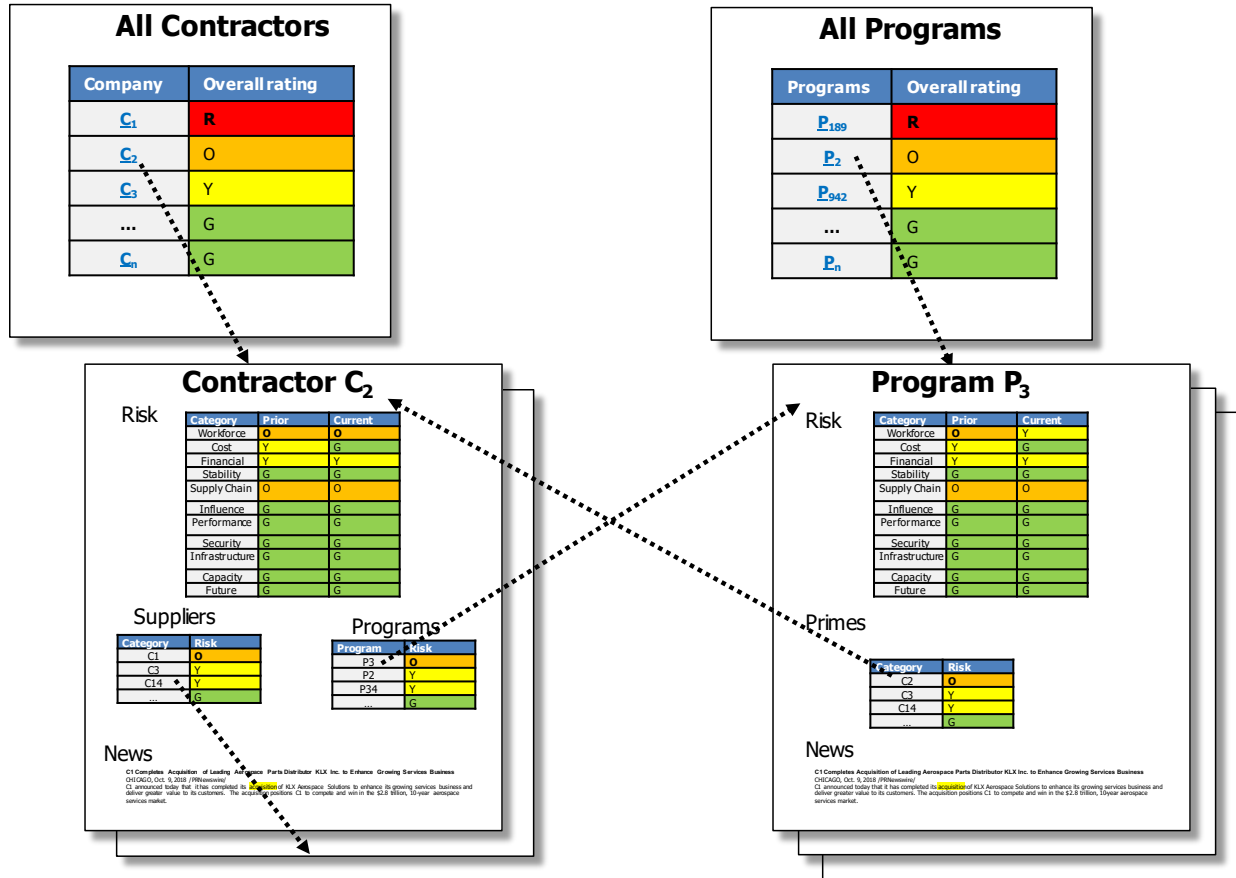
User Interface

Prototype users interact with the data and results through a secure, CAC-enabled web interface. It provides four primary views:

- an All Contractors page
- numerous underlying Single Contractor pages
- an All Programs page
- numerous underlying Single Program pages.

Information on each of these pages is often linked—for example, the Single Contractor page displays information about the specific contractor, other contractors with whom their contractor has a relationship, and DAF programs on which the contractor works. Users can easily navigate between these pages, as shown in Figure 3.1.

Figure 3.1. Prototype User Interface Structure: Contractor and Program Views



NOTES: The arrows illustrate linkages in PIERCE between views. Thus, the Program P₃ element on the Contractor C₂ page points to the Program P₃ page. The arrow in the lower left from Supplier C₁₄ points to the Contractor C₁₄ page lower in the stack. G = green; O = orange; R = red; Y = yellow.

Additionally, every page in PIERCE has global navigation controls. Each page has a hyperlink at the top to the All Contractors page and the All Programs page, allowing the user to quickly return to these broad overviews. Also, every page has a search box in the upper-right corner where a user can search by name for the page of a specific contractor or program; the search box displays suggestions in a drop-down menu as the user enters text.

Intended Uses

The research prototype designed and implemented for this project serves two primary categories of users and their business needs. The first user category consists of DAF PMs or portfolio managers who oversee one or more DAF programs. The second user category consists of DAF analysts or other personnel assigned to the headquarters component who must keep track of the overall state of all DAF acquisition programs.

Given these two categories of users, the PIERCE interface centers around two common uses. First, an analyst (e.g., at SAF/AQX) might need to quickly identify the most problematic DAF contractors compared with peers and discover whether there are any new problems for closer examination. This analyst could navigate to the All Contractors page that lists the status of all DAF contractors in order of their highest potential relative risk or execution indicator. This page also lists contractors whose worst indicators have increased since the previous evaluation period, followed by a list of contractors whose worst indicators have decreased since the past evaluation period. This allows the analyst to quickly discover contractors with new risks or points of concern that might require further investigation. After this, the application lists all of the known contractors (from those with the worst indicators to those with the best), allowing an analyst to quickly determine which DAF contractors currently present the greatest potential concern.

The analyst can click on any of the listed contractors to see its Single Contractor page containing additional information about the nature of the indicators associated with that contractor. The analyst can also search for a contractor of interest by typing a partial match for its name in the search box at the top of PIERCE and selecting from the dropdown menu. The Single Contractor page contains information on

- the worst indicator rating for the contractor
- current relative risk indicators for the contractor
- history of risks for this contractor across categories
- programs on which the contractor works
- subcontractors to the contractor and their worst indicator rating
- news stories about potential risk and execution problems for this contractor.

Second, an acquisition professional with responsibility for DAF programs may want to understand the state of the overall acquisition portfolio. Two types of pages facilitate this usage. The Single Program page lists all of the information PIERCE knows about the individual program in question. In particular, this page includes available information on:

- the prime contractor(s) that work on the program
- any subcontractor(s) that are known to work on the program
- the worst indicators of those primes and subcontractors
- contract execution metrics
- any recent PAR evaluations for the program.

The page also lists recent news stories about the program that are relevant to acquisition personnel. If one of the program's contractors has an elevated level of risk, the user can see details about the contractor's identified risks by clicking on their name and reaching the associated Single Contractor page. This will list the contractor's overall risks in each category and include an explanation of the specific conditions detected. It will also list all known subcontractors for this contractor (regardless of which program(s) those subcontractors work on,

if any²¹) and any recent news stories about this contractor that are relevant to acquisition personnel. The Single Program page and the Single Contractor page are both reachable from the search box in the application or by clicking on any reference to the contractor or program from the All Contractors page or the All Programs page.

All Contractors Page

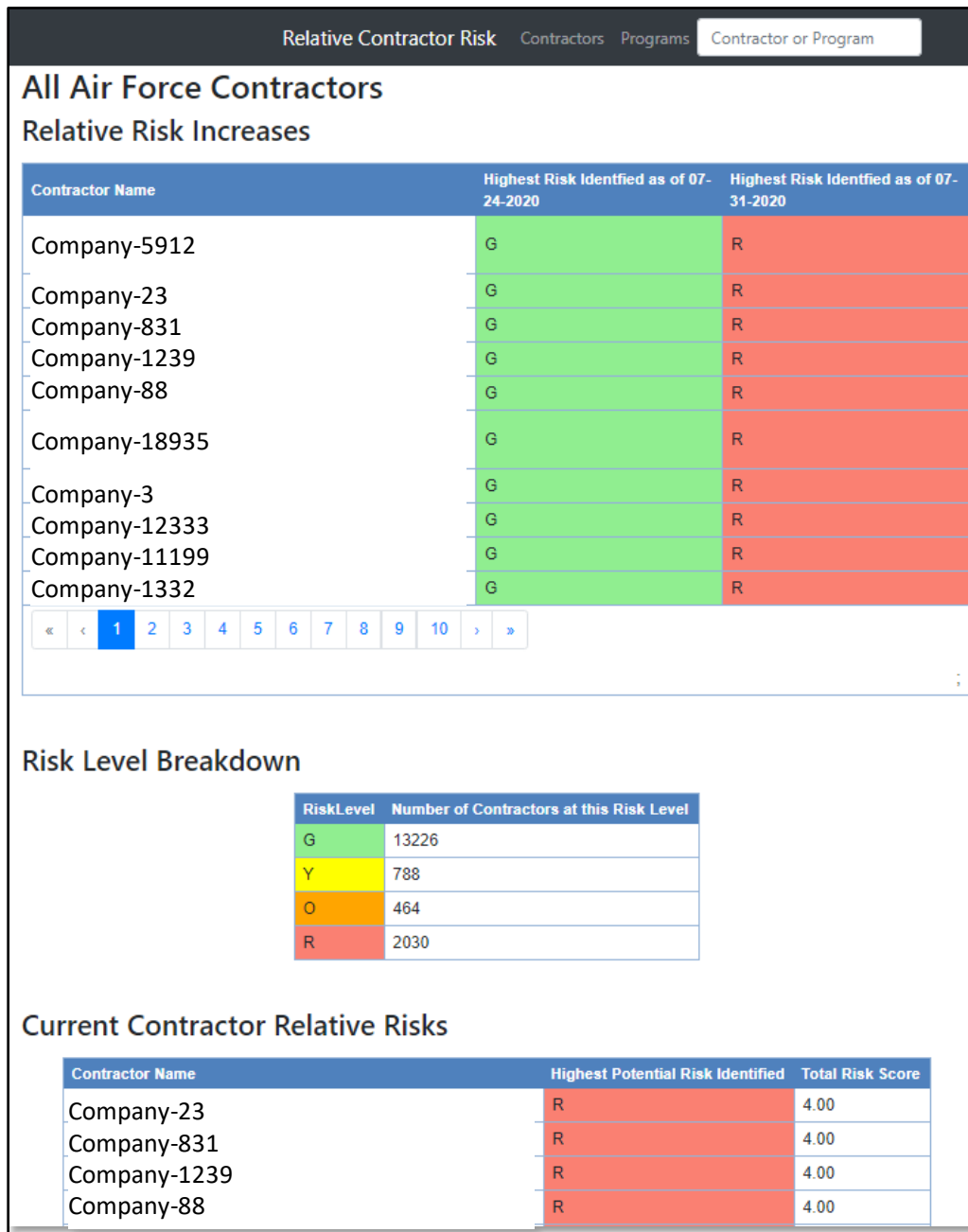
The All Contractors page quickly gives an overview of DAF contractors sorted by their highest potential risk indicators. Figure 3.2 shows an example screenshot of the page with anonymized results. At the top, the “Relative Risk Increases” table displays contractors whose overall relative risk level has increased since the previous evaluation period. These might be good starting points for further investigating new cases of higher potential risk. A user can select a company in the table to go to the Single Contractor page for the company.

Next, the “Risk Level Breakdown” table shows how many government contractors have been classified at each risk level in PIERCE. Green is the lowest potential relative risk level and red is the highest potential relative risk level.

Finally, the “Current Contractor Relative Risks” table lists every contractor for which the application can evaluate at least one potential risk condition. A Relative Risk score is displayed to give the user an idea of how large a gap may exist between two contractors. This score is not intended to focus the user on any particular score or number; rather, the scores are displayed simply to give the user a general understanding of the gap between one contractor’s relative risk and another’s. For example, a company that has two red risk indicators will have a total score that is higher than that for a company that has one red risk indicator. Although we do not have a theoretical basis for understanding whether two reds are worse than one (i.e., one indicator might be much more significant than another), this is a rough heuristic that might prove useful in prioritizing which company to review first. Future research on this factor is needed.

²¹ Some of our data sources indicate whether the suppliers work on U.S. Air Force or larger DoD contracts (and thus possibly programs). Other sources indicate general corporate supplier relationships—some of which might not involve defense work.

Figure 3.2. Example Prototype User Interface: All Contractors Page (anonymized)



NOTE: This notional screen shot is from an early version of PIERCE with test data. The numerical values in this figure are not illustrative of operation. The section at the top of the page is intended to indicate companies that have shown the largest increases in risk between reporting periods. Here we see that ten of the approximately 16,000 companies in our prototype had their worst ratings jump from green to red. Underlying pages (revealed by clicking on the page numbers across the bar in the middle of the page) would show companies with lower current ratings. The underlying metrics that result in the ratings shown on this page are visible on the Single Contractor page described later in this chapter.

Single Contractor Page

The Single Contractor page provides an overview of a particular contractor and the relative risk indicators for the company as a whole that might affect its performance. Figure 3.3 shows an example screenshot of such a page with anonymized results. At the top of the page, the contractor's level of highest potential risks identified by the data is shown by its color. For context, a coding key is provided to explain the four relative risk levels: red, orange, yellow, and green.

Below the overall risk level key on the page, the “Current Risks for This Contractor” table provides specific information about every potential relative risk identified for this contractor. As discussed in our companion report, each risk factor is assigned to one of the 11 risk categories identified (Anton et al., 2022). The “Risk Description” column provides an explanation of what this specific indicator is measuring and how this contractor rates relative to that indicator.

Below this table, the application displays a table showing the history of relative risks scores for this contractor. Once populated, the intent is to help the user identify new trends or values not noticed previously.

Next, PIERCE displays a table showing all of the DAF programs on which this contractor works. It also displays the total count of programs for which this company is a contractor or subcontractor. The user can click on any program name to see its Single Program page.

The last table displays all the companies known to be this company's subcontractors, regardless of program(s), along with their highest relative risk indicator score. Only on a Single Program page does PIERCE restrict subcontractors to those working on that program's contract. The user can click on any company name to see the Single Contractor page for the subcontractor selected.

Finally, the application displays news stories of interest about this particular contractor. These news stories originate from CapIQ but could come from a variety of publications. News is filtered so that PIERCE shows only stories that are relevant to a specific company and contain one or more keywords of potential concern listed earlier in Table 2.3.

Figure 3.3. Example Prototype User Interface: Single Contractor Page (anonymized)

Relative Contractor Risk
Contractors
Programs
Logout

Contractor or Program
Search

Contractor Name: Company1239

Contractor's Highest Potential Risk Identified:

Y

Coloring key

R

O

Y

G

Current Risks for this Contractor

Relative Risk Level	Risk Type	Risk Description
O	Financial	Change in moving average DoD obligations between FYs 2015-17 and FYs 2016-18 was between 20% and 50%, posing risk if contractor did not ramp up accordingly
G	Financial	Change in average revenue between CYs 2016-18 and CYs 2013-15 was within one standard deviation of average, posing minimal risk.
G	Hiring	Company1239 has no unusually high hiring risks based on Monster.com data

History of risks for this Contractor across categories

Risk Category	04-01-2019	06-01-2019	08-01-2019	10-01-2019
Financial	G	R	R	R
Hiring	G	G	G	G
Influence	Y	Y	Y	O
Performance	R	O	Y	G

Programs for Contractor:

Number of Air Force programs: 2

Programs

Program454

Program722

Subcontractors of Contractor:

Number of subcontractors: 3

Corporate Subcontractor	Highest Risk	Relative Risk Score
Company2	O	700
Company6	Y	3
Company13	G	1

News Stories about potential risks

Missile Defense Agency Awards Company1239 ...
PR Newswire
The Missile Defense Agency awarded Company1239 a \$... million contract to support the ...
...

Relative Contractor Risk Application
Copyright RAND Corporation 2018 - for prototype use only
For Official Use Only - May contain Proprietary Data

NOTE: Subcontractors listed are all those that have any relationship with this contractor.

All Programs Page

The All Programs page lists all known DAF programs so the user can quickly see which DAF programs have one or more contractors or subcontractors with an elevated relative risk rating. It simultaneously provides a quick overview of the contractors with the highest potential risks that are working on each of the DAF programs. Again, here we do not integrate ratings across companies but list the worst identified. Also, some ratings (the general company risk ratings shown on this page) are generic to a company and might apply to this program; this is why we indicate that the PM and others need to apply their expertise and further diligence to assess whether this risk is of concern. Other performance ratings (shown on the individual program page in the next section) are contract-specific and relate to the program in which the contract resides. Figure 3.4 shows a screenshot of an example page with anonymized results. Programs are ordered by the highest combined potential risk scores of each contractor who works on them. Note that the risk scores are for the contractors, not the program. The user can click on either the program or company name to get to the Single Program page with details on that program (see the next subsection).

Figure 3.4. Prototype User Interface: All Programs Page (anonymized)

Program Name	Contractor Name	Highest Risk Identified
Program 133	Company-1239	R
Program 133	Company-11821	R
Program 874	Company-1239	R
Program 874	Company-8339	R
Program 121	Company-23	R
Program 121	Company-13923	R
Program 982	Company-23	R
Program 982	Company-23	R
Program 133	Company-422	O
Program 133	Company-9882	O
Program 554	Company-3930	O
Program 321	Company-3930	O
Program 090	Company-4220	O
Program 454	Company-690	O
Program 454	Company-8990	O
Program 722	Company-690	O

NOTE: The risk scores are for the contractor(s) that work on the program, not the program per se. Also, here we do not attempt to integrate risks for a single program across companies but rather show any that have higher risks. For example, Program 133 has two companies with at least one red rating (Company-1239 and Company-11821) and two companies with at least one orange rating (Company-422 and Company-9882).

Single Program Page

The Single Program page lists a series of relevant details for a single, specific DAF program. Figure 3.5 shows a screenshot of an example page with anonymized results. First, it lists the known prime contractor for this program, derived from FPDS-NG and PMRT data. It also lists the highest relative risk identified for the prime contractor. Next, it displays all known subcontractors on this program, displaying the highest relative risk identified for each subcontractor (when available).²² Third, the page displays the available contract performance execution metrics for each major contract on the program. These are the new performance metrics discussed in Chapter 2.

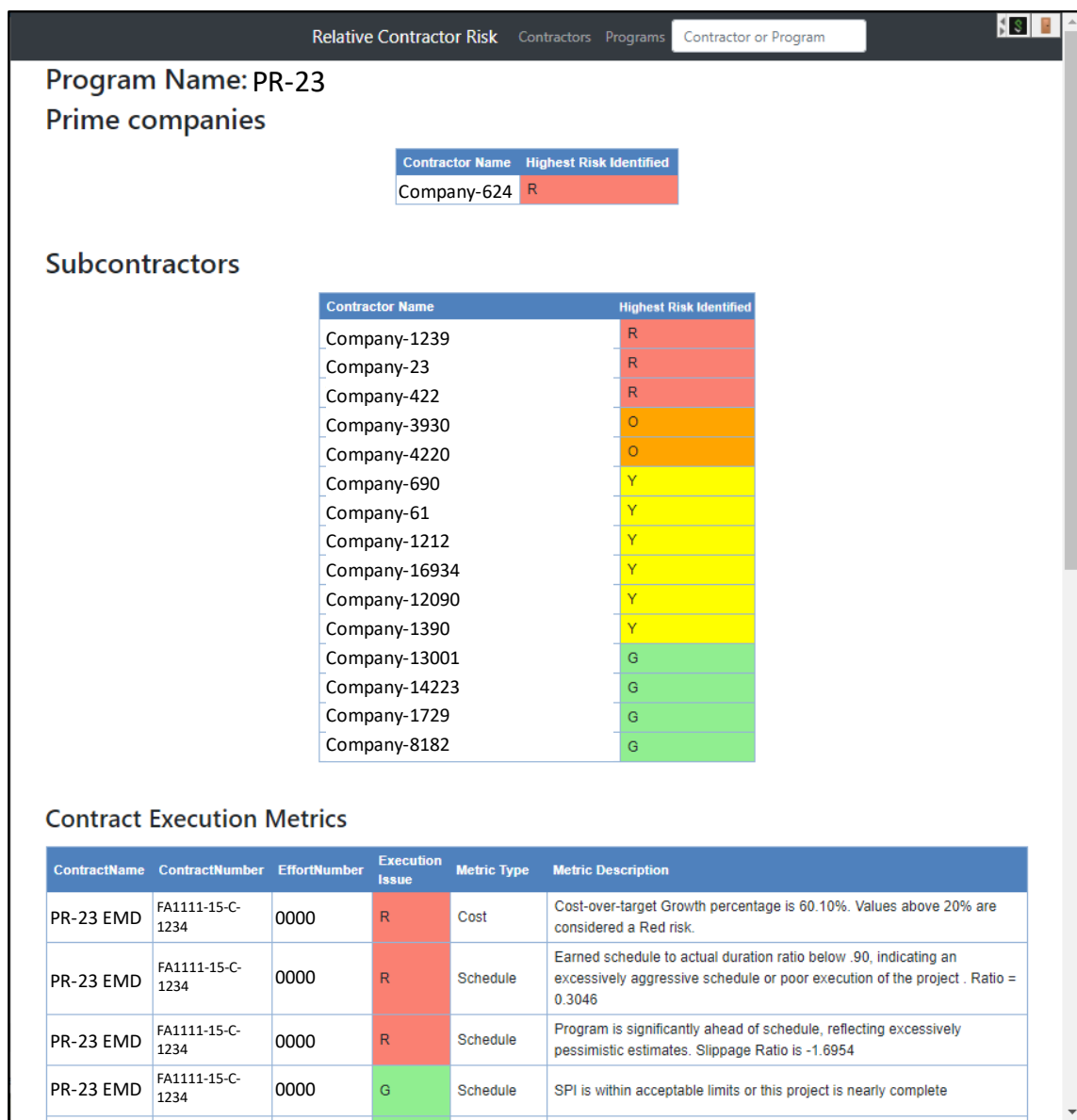
²² Currently, the PIERCE prototype does not tag individual data with their source, but that may be a consideration in future versions to help the user understand the pedigree of the data.

Below this information, PIERCE also lists the three PAR program evaluation scores, when available. These are marked as potentially containing proprietary information and are not shown in Figure 3.5. Currently, this is the only proprietary information in PIERCE.

Finally, it displays any instances where the sentiment analysis model predicts a worse rating than reported in the MARs. If there are no such instances, then nothing further is shown.

Each of the rows in the “Prime Companies” and “Subcontractors” tables allows the user to drill down into the details of contractors’ metrics by clicking on the name of a specific contractor.

Figure 3.5. Prototype User Interface: Single Program Page (anonymized)



Combining Performance Measures

To combine relative risk and performance measures, we first converted the ratings to a numerical score and then used various functions to combine those scores to obtain an overall relative risk measure. This allows us to reveal underlying risks so that they are not hidden by other scores, but it does not necessarily reflect a reliable overall risk rating for a company or program. That would require deeper understanding of the relative importance of different risks and performance metrics. Instead, our approach here filters through the vast data available to identify areas of potential concern for the PM and others to further investigate and assess.

Scoring Relative Population Risks

For our initial prototype, we used a simple algorithm to convert a single indicator value to a number. We used the associated population fraction involved if the population distribution were Gaussian:²³

$$Relative_score = 1 / Population_fraction$$

Here the score is simply the inverse of the fraction, as illustrated in Figure 3.6. For example, if the score was yellow, then that score would be associated with standard deviations between -1 and $+2$. In a Gaussian distribution, about 13.6 percent of the population would be in that range. Thus, the score is $1 / 0.136 = 7.4$. Of course, we could use the exact standard deviation value rather than the grouping value to obtain a more precise score, but this illustrates the concept. In addition to reflecting how far the risk measure is from the population, this yields a nonlinear scoring value that reflects extreme outliers, which is what we want.

²³ We realize that we do not know whether the distribution is Gaussian, but this simple approach gives what we need: a strongly nonlinear scoring that heavily emphasizes companies that are three standard deviations away from the mean over those that are two standard deviations away (and similarly compared with those that are one standard deviation away). Further research and effort could be made in subsequent iterations of the prototype to empirically assess the actual distributions and associated fractions, but this simple approach was useful for a first iteration to get an operational prototype functioning and allow us to begin assessing these finer points.

Figure 3.6. Relative Risk Rating and Score

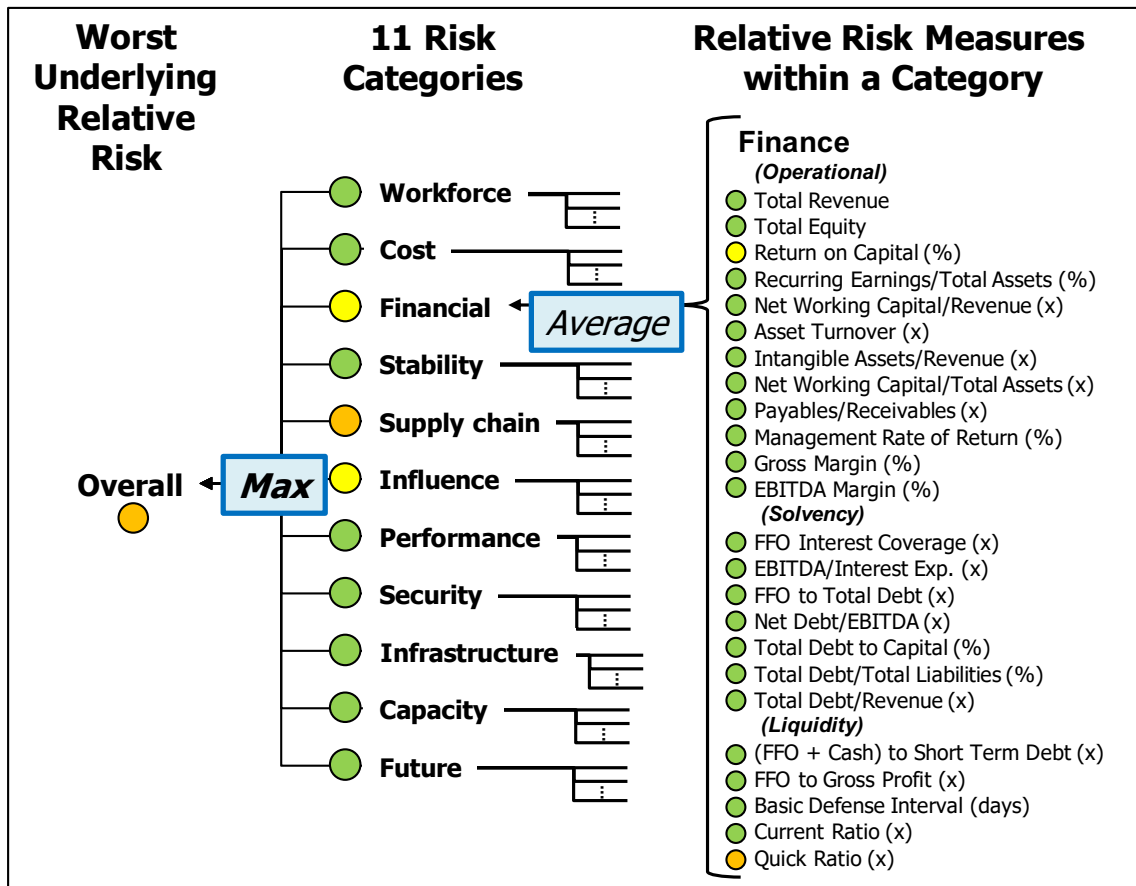
<i>Relative score = 1 / Fraction</i>			
		<i>Fraction</i>	<i>Score</i>
G	$x < 1\sigma$	84.2%	1.2
Y	$1\sigma \leq x < 2\sigma$	13.6%	7.4
O	$2\sigma \leq x < 3\sigma$	2.1%	47.6
R	$3\sigma \leq x$	0.1%	1000

Combining Scores Within a Category

Now that we have converted an individual indicator to a numerical score, we can use simple functions to combine multiple measures with a category and across categories to get summary risk scores.

Within a risk category, the preponderance of multiple indicators should reflect the strength of the potential concerns and the confidence that there may be a concern worthy of further examination. In other words, if multiple measures in the same category indicate that there may be a concern, then the confidence should be higher. Thus, we took the average (arithmetic mean) of the available indicators within the category. In some cases, we have only one measure, so the category score is simply that single measure score. In other cases, we have many measures, so the average helps to understand how strong the indication is when we look at it in multiple ways within a category. The right side of Figure 3.7 illustrates this concept for the large number of financial health measures available. In this case, we have 24 measures across operational, solvency, and liquidity categories. All are green except one orange and one yellow rating. Here the average is yellow because we have a nonlinear value function for combining ratings. We do not lose the fact that there are two measures worse than green, but the preponderance of measures indicates that these are localized to just two of 24 measures and thus somewhat better than, say, the highest indication (orange).

Figure 3.7. Combining Indicators Within and Across Categories to Reveal Worst Potential Risks



NOTES: Each of the 11 risk categories have underlying risk measures. Here we show the 24 financial risk measures that are averaged to obtain the overall rating for the Financial Category. Nonlinear weights are used to emphasize any negative measures in the averaging. The maximum ("Max") of the 11 risk category ratings provides the Worst Overall Risk rating on the left. Green ratings are not outliers compared with the company's peers. Yellow, orange, or red ratings are 1, 2, or 3 standard deviations from the population mean, respectively, on the worse side of the distribution, or cross-set threshold values for increasingly distant outliers from the population. EBITDA = earnings before interest, taxes, depreciation, and amortization; FFO = funds from operations.

Once we have a combined score within a category, we simply applied the same breakpoints in the original scoring scheme to assign a color to the combined score (see Figure 3.8). Anything at 1.2 (the lowest assigned value) is green. Anything above that and up to the nominal yellow score of 7.4 is yellow, and so on.

Figure 3.8. Relative Risk Rating Based on Combined Scores for Multiple Measures

Combined score	Rating
1.2	G
1.3–7.4	Y
7.5–47.6	O
≥47.7	R

Other functions could be used to combine scores within a category depending on how many risks are triggered and how many the user wants revealed. For example, one might prefer to use a maximum function so that any risk (the worse risk) is propagated up from the category level. If the maximum were taken among the relative risk measures in Figure 3.7, this would mean that the Financial category value would be orange instead of yellow. Therefore, as with the scoring functions, the choice of a combination function has a significant effect on the revelation of potential risks and thus the user's perception of potential issues. Further research and user inputs are needed to understand what the best function might be, or whether a set of selectable functions should be made available through the user interface to allow the user to tailor the output based on his or her understanding of a program's or portfolio's specifics. Ideally, we would understand both the relative importance of different measures within a category and how those measures interrelate (either generically or for the particular situation in a specific program or contractor), but that level of understanding is beyond the scope of this initial stage of exploratory work.

In addition to taking the average of the indicator scores, we can weigh the relative importance of the available measures. In the case of financial health measures, we implemented weights to prioritize a few that our financial expert on the team deemed the most important (see Anton et al., 2022). Such weightings could conceptually be set by the user based on what risks she or he is most concerned about.

Combining Scores Across Categories

To combine indicators across categories, we used the maximum function because we wanted the user to see the worse indicator. This initial choice was driven in part because the data we used initially had few indicators above green and we did not want the indicators in a single category to be washed out by other categories. Of course, other functions could be easily implemented (e.g., using the average, as we did within each category, or perhaps a simple summation function).

The left side of Figure 3.7 illustrates how the maximum function reveals the worse indicators. In this notional example, the Supply Chain is a relative risk of orange, so the overall indicator is shown as orange.

Although we use colors to rate the indicators and make it easy for the user to quickly see higher potential concerns, the system retains and displays the underlying scores. Thus, a

combination yellow score that is supported by two yellow scores has a higher risk than one that is supported by a single yellow score.

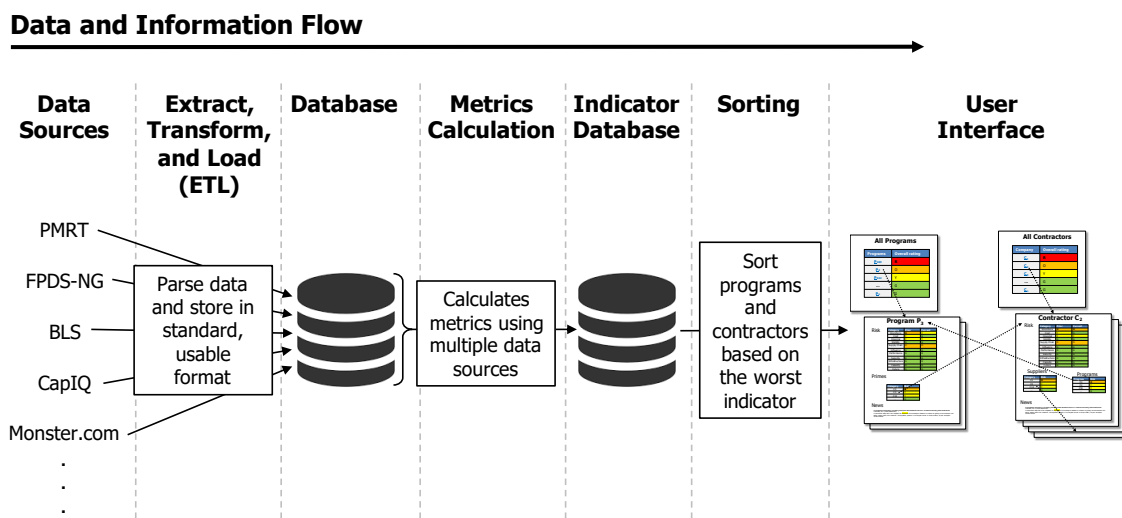
Observations

In summary, the indicators discussed in this report combine available data to indicate when a company exhibits an increased concern when compared with companies with different values. Although the approach usually cannot definitively identify that a risk indicator has a significant likelihood and consequence or whether an execution problem definitively exists, those measures are objectively applied to all companies and constitute a way to reduce the number of issues for further analysis or investigation (due diligence). In other words, we are not predicting (with probabilities and identified consequences) specific risks, but this approach raises risk and performance *indicators* of potential concern for further investigation by PMs and others.

Architecture Overview

The prototype uses data from a variety of sources, including restricted government sources (e.g., PMRT), open government sources (e.g., FPDS-NG, the BLS website, and USAspending.gov), and open or for-fee commercial sources (e.g., job posting websites and aggregators of financial data). Each of these sources feeds a data and information pipeline in the PIERCE architecture (see Figure 3.9). The pipeline has two primary components: data parsers and processors.

Figure 3.9. The Prototype's Architecture



Extract, Transform, and Load (ETL). As a first step, parsers specific to each data source clean and normalize the data into standard formats. This serves several purposes. First, it removes data that appear to be missing critical fields or that fail other validity checks. Second, it associates data that originate from different sources to facilitate shared concepts or data constructs. For example, correctly identifying that data from different sources describe the same company or DAF program can be a challenge when (as is often the case) the input data lack a unique identifier to confirm the match. Here, parsers look for near-matches between the input data and known companies or programs to determine whether the input data represent something new or can be matched to a known entity. Finally, cleaning and normalizing the data allows PIERCE to add, remove, or change data sources that provide similar types of information without the need to alter or rewrite any part of PIERCE other than the data parser specific to the data source in question.

Metrics calculation. Once the data have been parsed and cleaned, a different set of software code examines the data to assess all indicators that can be calculated based on the available data. These indicators use data from one or more data sources, depending on the specific nature of the metric. Data pipelines are not run continuously; they run only when new data are received. This varies based on the nature of the data. For example, financial data are typically updated quarterly (when SEC requirements require corporations to update their investors and potential investors as to the state of their business). Other data, such as the posting of open job positions, could be updated on a monthly or even daily basis if so desired. Once complete, the indications generated by these pipelines are stored in a MySQL database that other portions of the application can reference at any time.

Sorting. As a final step, once all of the indicator modules have completed and any newly identified potential concerns have been stored, a final software job calculates the performance indicators (described earlier in this report) and the relative risks of all known DAF contractors within each of the 11 risk categories (described in Anton et al., 2022) and an overall indicator rating for each contracting company.

User Interface. Each of these data pipelines and risk calculators operate independently from the final element of PIERCE: the user interface. Users can interact with PIERCE through a standard web browser with a CAC reader. The research prototype's user interface has been built using the Model-View-Controller paradigm²⁴ and can be deployed on a standard Microsoft Internet Information Services web server.

²⁴ See, for example, Trygve Reenskaug, "THING-MODEL-VIEW-EDITOR: An Example from a Planning System," May 12, 1979a; Trygve Reenskaug "MODELS – VIEWS – CONTROLLERS," December 10, 1979b; Trygve Reenskaug, "The Model-View-Controller (MVC): Its Past and Present," August 20, 2003; and Steve Burbeck, "Applications Programming in Smalltalk-80(TM): How to use Model-View-Controller (MVC)," *Smalltalk-80*, 1987.

4. Insights, Conclusions, and Next Steps

Insights on Data Utility and Availability

The intent of this project was to combine traditional and nontraditional²⁵ data in a way to provide acquisition professionals earlier indications of relative contractor performance risks and execution concerns. Although there are several traditional ways that program information is shared (e.g., MARs, SARs, DAES), supplementing these methods with other data sources—government, public, and commercial—is useful. There are some data that could potentially make program execution indicators more robust but are not available to the public and require specific permissions from the requisite authorities. Although some of these data fall into the Sensitive Data category as authorized for release to federally funded research and development centers under Section 235 of the National Defense Authorization Act for Fiscal Year 2017,²⁶ there are additional useful data sources that remain unavailable.

Sensitive Data Access and Use

We were able to access and incorporate data from government, public, and commercial data sources. Much of this is not identified as sensitive; however, some portions of it are.

DCMA is the “independent eyes and ears of DOD and its partners . . . providing relevant acquisition insight supporting affordability and readiness” (DCMA, undated). Therefore, DCMA surveils hundreds of thousands of contracts at more than 15,000 contractor locations. DCMA has a system called eTools, which contains several kinds of data that could be used by this prototype.

DCMA’s monthly PARs are an example of these kinds of data. These are assessments of contractor performance by government personnel either collocated at the contractor facilities or who are part of a regional team that regularly visits these facilities with far-reaching access. The PAR consists of a contract performance assessment, management assessment, and production

²⁵ Again, these data are nontraditional for government PMs. These data are used by a great many people for other purposes.

²⁶ Public Law 114-328, 2016, defines *sensitive information* as

confidential commercial, financial, or proprietary information, technical data, contract performance, contract performance evaluation, management, and administration data, or other privileged information owned by other contractors of the Department of Defense that is exempt from public disclosure under section 552(b)(4) of title 5, United States Code, or which would otherwise be prohibited from disclosure under section 1832 or 1905 of title 18, United States Code.

This law authorized a pilot program permitting the release of these kinds of information to federally funded research and development centers. Our project was a recipient of these types of information.

assessment. These are targeted at acquisition professionals, providing them with contract execution status and emerging issues that might merit additional management attention (DCMA Instruction 205, 2013).

However, eTools and the data contained therein are not readily accessed and were generally not made available to us, even under the auspices of Pub. L. 114-328's Section 235 data pilot. Having these surveillance reports incorporated into PIERCE, combined with the other previously described data sources, would provide acquisition professionals additional insight into contractor performance and program execution.

This was but one example of sensitive data. Others that would have been useful include contractor performance assessments, the Air Force Cost Analysis Agency's service cost positions (SCPs), winning contractor proposals (e.g., staffing plans, technical approaches), actual contracts, and cybersecurity compliance reporting.

Conclusions

Our research shows that it is possible to access and combine traditional and nontraditional data sources to highlight areas for additional management attention. It would then be up to the government acquisition professionals to use their expertise and program insights to determine whether this focus area was relevant or significant to the point where risk was no longer tolerable.

By adding more data, including those described as sensitive, the acquisition professional will have a more complete picture at his or her fingertips. These data and the addition of more metrics in the 11 risk categories will provide a more balanced assessment for the acquisition professional's use. For this to be successful, data availability and accessibility are key. Also, to reduce errors induced by manual manipulation of data, an application programming interface or machine-to-machine access is preferred to manual extraction of data for inclusion in the tool's data set. This makes it easier to handle large data sets more efficiently.

With larger data sets, the algorithms can be further "tuned." Longer periods of data and more-diverse data sources related to the same underlying risk condition provide increasing amounts of context to examine a particular risk condition. With this done, more fine-grained filtering of potential contractor risks could be accomplished.

Additionally, it is important to understand the types of questions the user communities will want to answer. A PM is typically focused on one program with one contract team. Consequently, a PM might focus on the Program page. A PEO or an MDA have a portfolio to manage and might find the All Programs page more useful, with the ability to drill down into specifics about a program's contractor team as needed. Other officials with oversight responsibilities, i.e., performance of the defense industrial base, might find the All Contractors view more interesting as this could provide insights into issues different segments of the industrial base are having.

Next Steps

As previously mentioned, we created a prototype tool, which we believe will be useful to acquisition professionals regardless of position—in execution through oversight. The algorithms and assumptions could be refined, tuned, and validated with increased data access, implementation of additional measures, and more government beta testing.

Gaining Access to Other Data

There are other data sets that could be immediately useful. Access to all government contracts through Electronic Data Access, a system managed by the Defense Logistics Agency for OSD, would allow for comparisons across contracting types, incentive schemes, implementation of clauses, etc., to understand their impacts. This would have acquisition and contracting policy implications.

Accessing the Contractor Performance Assessment Reporting System (CPARS) and integrating its data into the tool would make the assessment of contractor performance more robust. It would also make this valuable data source more available to the acquisition community and could expand its use.

DCMA is the organization charged with monitoring contractor execution performance and, therefore, has compiled data for many programs of interest. The PARs are one such instance of the available data. Incorporating other DCMA data into this prototype would also enhance the insight of acquisition professionals and help with earlier notification of contractor performance risks.

Within the current cybersecurity environment, it would be helpful for acquisition professionals to have an increased understanding of their contractor teams' security posture. This directly affects program security and safety. Properly making this more available would therefore benefit programs.

Implementing Other Possible Risk Measures Identified

There are multiple measures we discussed at varying levels of detail throughout this report with some being closer to implementation than others. These include the Rayleigh Function (discussed earlier in the report and later in Appendix B), additional sentiment analyses, and, if provided access, assessing CPARS, SCPs, and winning contractor proposals. Next, we highlight how some of the possible measures could be part of next steps.

Additional Sentiment Analyses

As we found with the MARs, there is a significant amount of text in formal program management reporting. Two additional specific instances immediately come to mind: SARs and DAES. These are well-defined reports that the program offices perform on a recurring basis for Congress and OSD, respectively. As we found with the MARs, text could be mined to identify

potential indicators of execution issues. Although these tend to be trailing or limited indicators, artificial intelligence (e.g., machine learning or neural networks) and traditional statistical analysis tools may be able to identify new leading indicators. Although the frequency of these reports is not high (annual for SARs and quarterly for DAES), they are kept for years and are good potential sources for longitudinal analyses or potentially algorithm and metric validation.

We also note that other research on the potential use of text, sentiment analysis, and machine learning to identify early predictors of final performance have had mixed results. At the program level, using text in the SARs is an option; however, Joseph and Sconion, 2020, reported that only 12 percent of the SARs' executive summaries in their sample produced strong correlations with future program-level cost growth. Also, preliminary research in Davis et al., 2020, was unable to ascertain whether machine learning algorithms could predict program quantity and cost performance but asserted that future research is warranted. Davis and colleagues also noted that such challenges as political externalities, high uncertainties, and small numbers of programs likely aggravate efforts to predict major program outcomes; this is in contrast to the contract-level work discussed in our report.

CPARs

CPARs are formal assessments by the government of a contractor's performance on a specific contract during a specific period. This assessment is recorded via a web-based application overseen by Defense Pricing and Contracting. It should be objective and fact-based (CPARS, 2020). Therefore, these data, much like DCMA's PARs, could be used to highlight areas of interest regarding a specific contractor's previous performance on similar programs (contracts).

Contracts

Actual contracts contain a significant amount of information; however, much of it is text saved in portable document format (PDF). These are stored in the PEE in the Electronic Data Access application. The actual contracts contain information on incentives, clauses, and initial costs; all of which could be examined to associate text with program performance. Text analysis tools can improve search capabilities to locate known topical indicators of problems in free text through preprocessing of information and facilitating searches by keyword, context, and, in some cases, sentiment and tone.

Contractor Proposals

Access to the winning contractor's proposal would allow insight into how the contractor plans to execute the contract. One item of interest would be the contractor's staffing plan. For example, with this information, PIERCE could compare advertised vacancies with the staffing plan's identified critical positions and help the government acquisition professional assess the

level of risk. Ideally, the contractor would be sharing these issues with the government; however, this might allow the PM to get ahead of potential issues.

Some of this information might be available via DCMA's eTools suite, potentially within the Contract Business Analysis Repository.

Other EV-Based Predictive Measures

Finally, further work could be conducted to review the range of EV-based predictive measures of final cost and schedule. For example, Lipke, 2004, and GAO, 2020a, review the reliability of different IEACs and treatments of CPI at different stages in a contract (e.g., early, mid-term, or late in terms of percentage completed). Further refinements in the prototype to adjust which measures to use at different stages in the contract could improve the reliability of applying these approaches.

Appendix A. Sentiment Modeling and Analysis and Modeling of Free-Text Assessments and Associated Ratings

This appendix provides detailed descriptions of our approach to modeling the association between the unconstrained natural language reports in a MAR that summarize the status of a program and the reports' overall summary scores of green, yellow, or red.²⁷ Our general approach is to employ sentiment analysis to determine the general attitude (e.g., positive, neutral, or negative) of the text in a program's MAR and compare it with the human assessor's summary score for the program in the same MAR. The objective is to use this sentiment text analysis to detect cases where the text (the sentiment) appears different than the summary score (i.e., that the text indicates a sentiment that is more negative than the summary score indicates). Such differences may detect problems earlier than the scores report and may predict future score changes (e.g., that a currently green score might turn yellow in the future).

Approach

To begin, we reviewed and assessed a variety of text mining methods for this application. For instance, RAND has an in-house suite of lexical analysis tools (collectively called RAND-Lex) that can be used to examine text through four different modules: lexical analysis, stance analysis, topic modeling, and autocoding. The stance analysis module in RAND-Lex helps researchers understand how the world is being represented in a text. We also explored other tools, such as Word2Vec, which employs a machine learning algorithm to assess word similarity. It identifies the unique words in a text corpus of vectors so that each unique word is mathematically related to all of the other unique words in the corpus. That is to say, the model assigns a list of numbers to each unique word that is the length of the total number of unique words in the corpus. Each entry in a word's list is that word's numeric relation to another word in the corpus. The model can then compute relationships between and within words. This helps researchers identify relation and similarity between words within a text. In the end, we decided which method to use after exploring the data and processing initial results, as explained next.

In each MAR, there are various textual statements that summarize the status of the program from different perspectives. For instance, there is the general text description of the program, the program's Acquisition Category (ACAT), the program type, a set of Program Issues and Accomplishments texts, and a textual status report from each of the PM, PEO, and 3-Letter (or

²⁷ These colors reflect the subjective expert judgments of the human assessor on the status of the program. Their meanings are discussed in a footnote in Chapter 1.

other).²⁸ Each MAR has seven summary scores—one each for PM, PEO, 3-Letter/other, cost, schedule, performance, and funding (see Figure A.1)—where the color summarizes the status of the program.

Figure A.1. Example MAR Scores and Assessments (anonymized)

Program XXX

ACAT: II | Program Type: Acquisition Program | MAR Aug yyyy

Description: The ... upgrade includes ... aircraft equipped with all variants of the ... improves Reliability, Maintainability & Supportability (RM&S), and substantially improves operational effectiveness and combat lethality.

PEO: Brig Gen Smith, Jane | PM: Doe, John, Col.

Assessments

PEO | 3-Ltr/Other | PM | Cost | Schedule | Performance | Funding

Contractors

#

Top Issues

Area

1

Issue: Reduction to Funding for FY19 decreased the possibility to procure all ...

Impact: Budget cuts have driven the POR to ...

Way Ahead: None.

Issue: FY19 PB reduced Program funding in FY...

Cost, Funding

2

Impact: Based on ... , thus prohibiting the

program from executing final ...

Way Ahead: program submitted POM request ...

Working with ACC to finalize Way Ahead COAs for ...

Funding

Accomplishments

- Completed... Val/Ver - Completed... - Total ... installs now at ...

PEO:

Concur with PM Assessment. Should the recent SASC mark hold, the <program> will be unexecutable. (August)

3-Ltr/Other:

Concur with PM Assessment. Should the recent SASC mark hold, the <program> will be unexecutable. (August)

PM:

PM Overall rating is green. The <program> has been successfully executing and delivering ...

Minor issues with the delay to the start of modifying...

"overlapping" to "nose-to-tail", so as to help their aircraft availability. This stretches the ... install schedule out to a May 2020 completion. Mitigation efforts are in work to recover.

With delta ASP approval for ... , the RFP was released for ... depending on

negotiations. Further, with SASC proposed reduction of ... will cripple this program and RFP will be cancelled. These subsequent cuts will eliminate our ability to procure and

install ...

SOURCE: Adapted from data provided by DAF.

The data are initially put into a dataframe with 25 columns that provide information on each program. Each program is listed in a row in the dataframe. The first four columns contain the four textual statements that are extracted from the MAR reports:

1. combined Program Issues and Accomplishments
2. PM's status report
3. PEO's status report
4. 3-Letter/Other's status report

²⁸ As discussed in footnote 5 in Chapter 1, the *3-Letter* is a staff organization that provides assessments of the program independent of the PM's and PEO's assessments.

The first column (Program Issues and Accomplishments) includes technical discussion along with what is supposed to be a list of issues encountered by the program. The second column is the status report provided by the PM. In our sample, we found that the last two columns frequently either are blank or consist entirely of “Concur with the PM” or the like, but occasionally they express disagreement with what was written previously by the PM.

The next six columns in the dataframe are binary variables indicating the program’s type. There are six different program types that appear in our data. Table A.1 shows the counts of each in the 3,530 programs in our sample.

Table A.1. Program Types and Sample Size

Program Type	Count
Acquisition Program	2,191
Support Program	57
Operating and Support Program	197
Investment Activity	263
Foreign Military Sales Program	594
Operations and Maintenance Program	228

A binary indicator variable is *1* if and only if the characteristic matches the row of data. For example, if the program type of a data row is “Acquisition Program,” then the variable *PT_Acquisition_Program* will be *1*, while all other program-type variables will be *0*.

The next eight columns in the dataframe are binary indicator variables for ACAT levels. Table A.2 lists the eight ACAT levels and the number of occurrences of each in our data sample (counting “blank” as one of the eight levels).

Table A.2. Number of Programs by ACAT Level in MAR Sample

ACAT Levels	Count
I	22
IB	58
IC	269
ID	53
IAC	26
III	1,348
II	415
unspecified ^a	1,339

SOURCE: PMRT MARs.

NOTE: ACAT levels are defined in DoDI 5000.02. Generally, ACAT I (and IXX) programs are large major programs, while ACAT II and III are progressively smaller in dollar size.

^a The data source (PMRT) did not specify the ACAT level for these 1,339 MARs.

As with program types, one and only one of the eight columns contain a *1*, while the rest contain a *0*. Thus, for example, the column for variable *AL_IC* will be equal to *1* in exactly 269 data rows and equal to *0* in the 3,261 other rows.

The last seven columns in the dataframe contain the assessment ratings. These are the seven different ratings that were listed in the report. Because these are rated as colors in the report, we converted the colors to numbers, with *0* corresponding to green, *1* to yellow, and *2* to red. Sometimes the colors were either blank or listed as N/A; in such cases, the rating was ignored. In the 3,530 program rows in our data set, Table A.3 lists the assessment categories and the corresponding number of rows in which it has a proper color rating of green, yellow, or red.

Table A.3. Color Rating Assessment Categories in MAR Sample

Assessments	Count
PEO	2,714
3-Letter/other	3,295
PM	3,527
Cost	3,530
Schedule	3,530
Performance	3,530
Funding	3,530

SOURCE: PMRT MARs.

The next step was converting the text columns into meaningful numeric values. Such an approach as simply assigning a number to each word and creating a binary variable is not sufficient because this effort ignores the context of where words appear in relationship to one another. Instead, we employed two different open-source methods—a sentiment intensity analyzer and a text encoding. The latter produced better results, but we will discuss both methods because the sentiment intensity analyzer approach is both conceptually easier to understand and is less of a black box (i.e., it produces results with a clear explanation of how it reaches its conclusions).

Sentiment Intensity Analyzer Approach

As mentioned earlier, sentiment analysis is the use of natural language processing to determine whether a corpus of text is positive, negative, or neutral. Sentiment analysis tries to express the voice of the text as quantitative numbers. To accomplish this, we use Python’s Natural Language Toolkit (NLTK)—an open-source library that has a built-in Sentiment Intensity Analyzer (NLTK, 2020; Bird, Loper, and Klein, 2009).

Each call to the Sentiment Intensity Analyzer returns four numbers, which are its ratings of positive, negative, neutral, and compound. The compound score is an aggregate of the three other scores. The non-compound ratings are between 0 and 1, while the compound score provides ratings between -1 and 1. By taking each of the four text writings and computing these four scores for each, we generate 16 numbers for each program. These 16 numbers are then inserted into our dataframe, and the four columns representing the text are removed. This turns the original dataframe with 25 columns into a new one with 37 columns containing all numeric values. Having only numeric values allows for insertion into a supervised learning algorithm, as described below.

Text Encoding Approach

To improve these results, we used the Google Universal Sentence Encoder (Cer et al., 2018; TensorFlow, 2020). This is an open-source software library that encodes text into high-dimensional vectors that can be used for text classification, semantic similarity, clustering, and other natural language tasks. This encoder was chosen because of its recent success with other natural language processing tasks and because it was trained using a large number of natural language tasks that required modeling the meaning of sentences—not just the meanings of words.

The Universal Sentence Encoder takes text in standard paragraph form and encodes it into a vector of dimension 512. We can think of this as turning text into 512 numbers that represent it. Unlike sentiment analysis tools, a human looking at the numbers would have no way to infer what each number means or why it was encoded in a certain way, but the idea is that two excerpts of text that are very similar will be mapped very closely in distance in the 512-dimension vector space, while different components of the vector account for differences in the

meaning of the text. There is no guarantee that this will work effectively for larger sets of data, but initial results from the testing indicate that this model is performing well in encoding the underlying textual content, so this appears to be useful for our purposes.

In this approach, each of the four pieces of text are effectively transformed into 512 different numbers (a 512-dimensional vector), which turn our initial dataframe from 25 columns, including four columns of text, into 2,069 columns, now entirely numeric.²⁹ Fifteen columns have entries that are entirely binary integers, the 2,048 text encoding columns consist of floating-point numbers mostly between -1 and 1 , and the assessment columns have entries that are 0 , 1 , or 2 , unless they are blank, in which case they are ignored.

Learning the Relationship Between Text and Green, Yellow, and Red Scores

Next, we essentially treat each of the seven assessment (green, yellow, red) scores (variables) as a different supervised learning problem. For each of the seven scores, the 2,062 nonassessment variables create a feature vector, and the assessment variable is the output. We removed a random sample of 20 percent of the data from the sample to be our testing set. We trained the model on the remaining 80 percent, and then used the resulting model to predict values for the testing set. Examination of its performance on this testing set (the reserved 20 percent) is what allows us to determine its effectiveness.

For our preliminary research, we used the free, open-source machine learning system called Light Gradient Boosted Machine (LightGBM) (Ke et al., 2017; Microsoft, 2020). LightGBM is a gradient boosting framework that uses tree-based learning. Basically, it creates many tiny decision trees, building and combining on good results as they are found, until a full decision process is created. LightGBM was chosen because of its reported effectiveness on a wide variety of problems and because it is optimized to use less memory and perform at a faster speed than other algorithms (Ke et al., 2017; Titov, 2020; Titov et al., 2021). Also, LightGBM takes many parameters as inputs; each parameter stands to influence the model and change the predictions that it makes. We tried many different sets of parameters to determine which ones performed best on the relatively small amount of available data, but further research might determine whether we are using the best ones.

We also used the LightGBM regressor instead of its classifier. Had we used the classifier, it would return values of 0 , 1 , or 2 , as in its training set, based on what it decides is the most likely. Instead, we chose to use the regressor, which treats 0 , 1 , and 2 as continuous values rather than categories. The regressor then returns a decimal number generally between 0 and 2 (and sometimes less than 0 or greater than 2), which we then use to make a prediction. For example, if the regressor predicts a value of 1.7 , then we count that as a prediction of 2 , while a value of 0.1

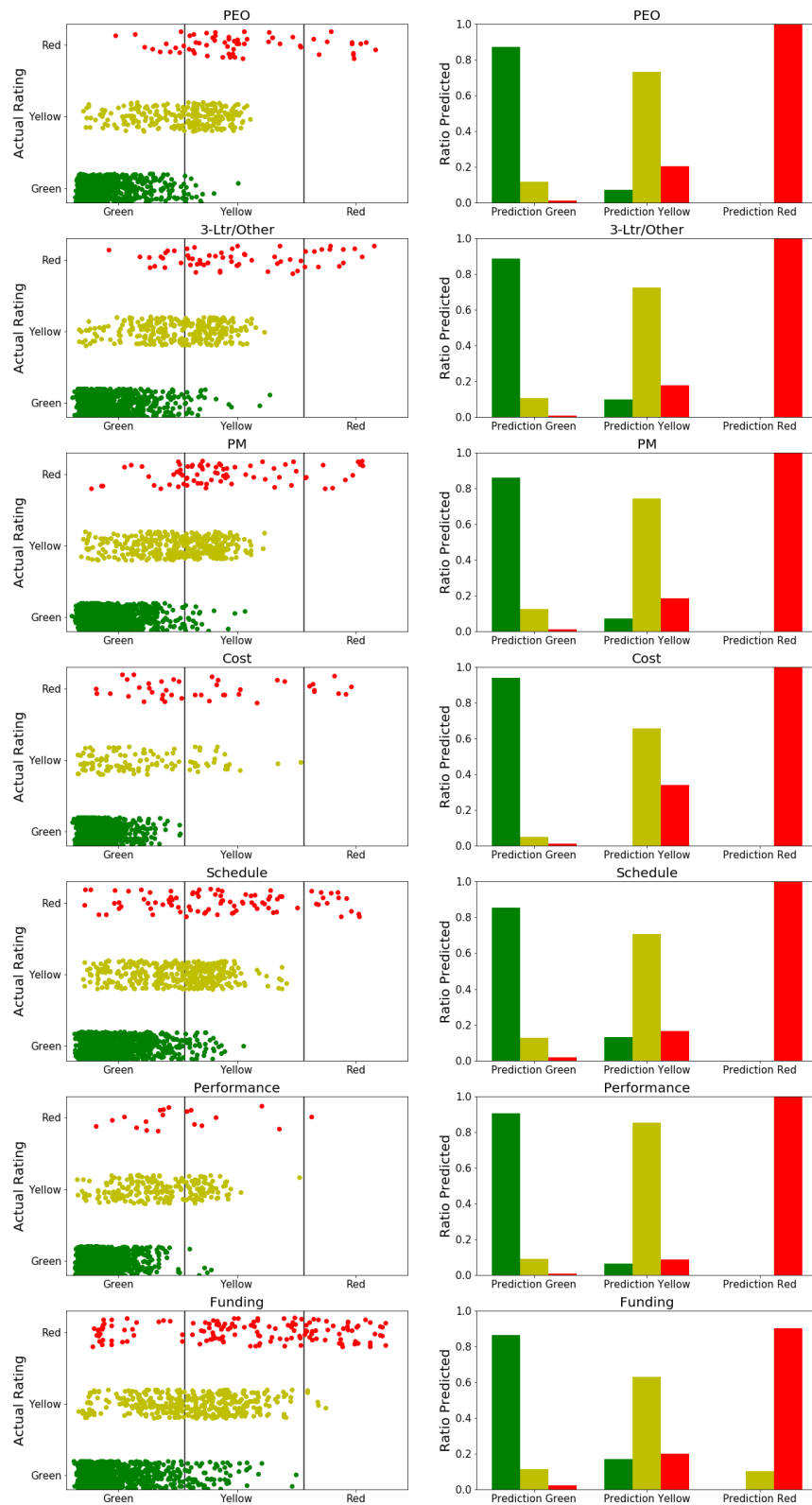
²⁹ For example, $4 \times 512 + (25 - 4) = 2,069$.

is a prediction of 0. In this preliminary research, we set the thresholds at 0.7 and 1.5, but further sensitivity analysis may be warranted.

Results

Figure A.2 shows our results. Each of the seven variables (PM, PEO, 3-Letter/Other, cost, schedule, performance, and funding) for the seven scorings has its own set of two adjacent plots. The plots on the left compare the model's prediction with the human's score for each program in the training set. Most of the time, the model's prediction equaled that of the human's score. For example, on the top-left plot for the PEO data, there is a large concentration of green in the bottom left, showing that in many cases the algorithm predicted green scores (low numbers) when the human assessment was green (0); it also shows a few cases where the model predicted yellow when the human scored the program as green. These latter cases may be instances of errors in the model, or it may indicate cases where the human has included discussion in the textual summaries that indicate concerns, but the human did not elevate the score to yellow. These are the cases that we highlight in PIERCE for further investigation by the use to see whether these are early indicators of performance issues not reflected in the color score. In other words, for these cases, the model predicts that the text provides indicators of performance problems earlier than by just relying on the color ratings.

Figure A.2. Semantic Text Assessment Results (using LightGBM)



The bar plots on the right side show the comparative results for each prediction if we use 0.7 and 1.5 as the thresholds between color codings. So, for example, when the algorithm predicts green for PEO, it matched the human’s score about 85 percent of the time but predicted yellow about 15 percent of the time. Also, whenever the model predicted red, the human also gave a red score 100 percent of the time; thus, the red bars are all at 100 percent (except for the last plot). Examining all bar plots shows similar results. Yellow predictions are more variable; the largest differences are in the Cost assessment, wherein the algorithm predicts yellow when the human gave a red score much more often than green. This is likely a good result for our purposes because we are more interested in PIERCE flagging cases where the human’s score is lower than indicated by the text, not higher. Of note, in this preliminary analysis, we are training the model on a set of data that may lack ground truth. The human scores in the training set may contain the same kind of scoring bias that we are trying to detect in the test cases (i.e., where the text indicates some problems that are not [yet] reflected in the color score). Further research could develop a cleaned data set where subject-matter experts recategorize such cases to help the machine learning system better model the correlation between issues in the text and the color rating. Still, it is interesting that the model appears to be able to learn the relationship even with noncurated data—perhaps because the frequency of these instances is low.

Finally, LightGBM also provides a potentially useful measurement called Feature Importance. Essentially, this measure tells how importantly the algorithm viewed each variable in its calculation. The algorithm’s views varied among the seven variables, but we were quickly able to see that although the model appears to have learned the sentiment relationship, it pays almost no attention to the Program Type and the ACAT Levels. For example, Table A.4 presents the average feature importance for the Performance Assessment scoring.

Table A.4. Average Feature Importance for the Four Text and Two Category Variables

Assessments	Value
Program Issues and Accomplishments	2.798
PM	2.65
PEO	0.456
3-Letter/Other	1.76
ACAT Level	0.05
Program Type	0.03

Here, the average importance it gave each of the 512 values in Program Issues and Accomplishments was almost 2.8, while ACAT level instead received only 0.05. Also, the PM

text was regarded as much more important when it came to the schedule and funding assessments than in the other five variables.

Ultimately, this method has created a model that appears to be fairly accurate at predicting the color assessment score given the free-text assessments provided in the MARs. Further research (especially with a curated training set) could refine the model and help researchers understand the false positive and false negative rates when employing such a model to detect early problems buried in textual reports.

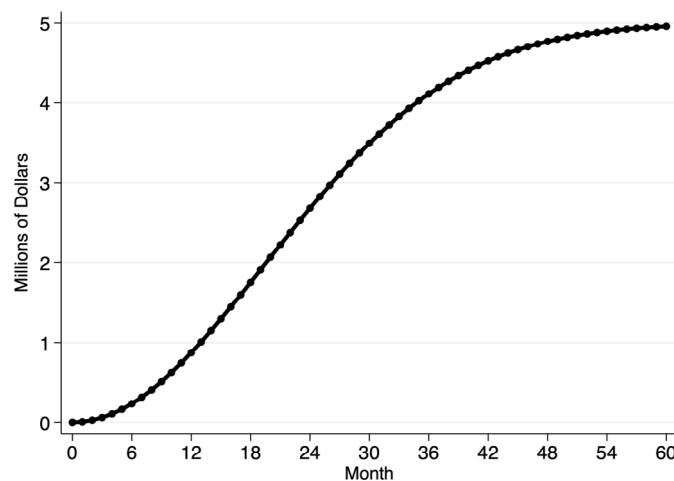
Appendix B. Assessing Final Cost and Schedule Using a Rayleigh Function

This appendix provides technical details of the Rayleigh model approach discussed briefly in Chapter 2. Here, we discuss the structure of the Rayleigh model and the steps taken to define a heuristic rule to handle when significant work content is added to a contract. The following section includes the results obtained from our preliminary development and analysis of a Rayleigh model.

The Rayleigh Model

The model is defined by the Norden-Rayleigh Function. This function is derived from the Rayleigh distribution. This cumulative distribution function is characterized by an S-shaped curve (see Figure B.1). Peter Norden at IBM studied the relationship between R&D project staffing and schedule over time and Rayleigh curves, which led to the definition of the Norden-Rayleigh Function (Norden, 1970).

Figure B.1. Example Rayleigh Cumulative Plot



NOTE: This plot displays a Norden-Rayleigh Function where the parameters α and d are set to $1/750$ and 5 , respectively.

Over the years, there have been multiple studies that have looked at how the Rayleigh curve—and particularly the Norden-Rayleigh Function—can be used to analyze and predict the cost and schedule for contracts in the defense sector. (For example, see Watkins, 1982; Davis, Christle, and Abba, 2009; Tracy and White, 2011; Brown et al., 2015; and Yang et al., 2019.)

Some of these publications have shown that the cumulative cost for these contracts, especially those focused on R&D contracts, tend to follow closely the curve defined by the Norden-Rayleigh Function:

$$c(t) = d (1 - e^{-\alpha t^2}) .$$

The Norden-Rayleigh Function is defined by two key parameters: the scale parameter d , and the shape parameter α . The d parameter can be used to predict the total cost of the project. In general terms, the relationship is defined as the total contract cost being a fraction of the scale parameter value.³⁰ The tolerance for this fraction has been defined at different points in past publications, but, for the purposes of this research, it was set at 0.99 in the following formula:

$$c_f = 0.99 d .$$

This fraction tolerance is also used to predict the contract duration in terms of α . Through algebraic derivation, it can be shown that if the final cost of the project is a fraction of d , then the final duration of the project can be expressed as

$$t_f = \sqrt{-\ln (.01)/\alpha} .$$

These equations—summarizing the relationship between the Rayleigh function parameters and the cost and duration of contracts—are used throughout this research to convert parameter estimations from curve fitting into contract metrics that help predict final cost and schedule given progress reports to date.

Bounding Parameters for Curve Fitting

Davis, Christle, and Abba, 2009, discuss the advantages of introducing bounds for the parameters when curve fitting. In their report, the authors compare the results of using unconstrained curve fitting with the ones obtained with parameter bounds. Their results showed that if there is complete data for a contract then the unconstrained approach obtains a better fit to the data. However, at the earlier stages of the contract, when there are not enough data points on the cost curve, the parameter bounds give a more-reliable structure to the curve and reduce the noise in the predictions. Based on their results, we decided to incorporate bounds for our parameter estimations.

³⁰ This is because the scale parameter itself is an asymptotic value.

To decide on a proper value for the parameter bounds, we did a sensitivity analysis on a sample of six contract tasks. Because of the correlation between the two parameters and how they influence the shape and scale of the curve, we decided to fix the α parameter bounds and evaluate the sensitivity of the d parameter bounds in relation. The α parameter's bounds were set so that the duration prediction ($\hat{t}$) would be constrained between $t_f \leq \hat{t} \leq 1.5t_f$, where t_f represents the contract estimated completion date reported by the PM or contractor (i.e., here we assume that the duration will be at least as long as currently scheduled). Further analysis may determine whether the lower bounds should be a large fraction of t_f to help estimate whether the contract may finish earlier than anticipated.

With the α parameter bounds defined, we evaluated the sensitivity to d parameter bounds by testing various sets of values. We defined six different bounds, ranging from constraints on the d value to between half the EAC and two times that value. For each set of bounds, we iteratively fitted the Rayleigh curve defined by the Norden-Rayleigh Function to the sample contracts, increasing the sample size for the fit on each iteration.³¹ The Non-Linear Least-Squares Minimization and Curve-Fitting (LMFIT) library package in Python was used to complete the curve fitting procedures (Newville et al., 2014; LMFIT, undated). For each resulting curve fit, we first calculate the predictions for cost and duration and then determine the percentage difference between the reported estimations and the predictions using the following formula:

$$\% \text{ Difference} = \frac{\text{Rayleigh Prediction} - \text{Contract Estimate}}{\text{Contract Estimate}} \times 100 .$$

The average percentage difference and standard deviations from each set of runs are shown in Table B.1. These results showed that Run 2 obtained percentage difference below 20 percent for both cost and duration predictions with relatively low standard deviations. Run 2 used bounds for d that correspond to the cost predictions to be within the interval $0.75 \text{ EAC} \leq \hat{C} \leq 1.5 \text{ EAC}$. We used these bounds throughout this analysis. Thus, we assume that the final costs will be at least 75 percent of the current estimates and as high as 50 percent higher than current estimates. PMs should be updating their EACs as work content changes, overruns are realized, and other data indicate that costs are higher than originally expected. Of course, the EAC could be so low (not properly maintained by managers) that the actual estimate could be higher than 150 percent of EAC, but this initial analysis and our experience indicate that this should be uncommon. Although this preliminary analysis found the best solutions from using these bounds, additional sensitivity analysis should be conducted to determine whether these bounds are too constrained.

³¹ We used an iterative procedure here to mimic how the model would be used in practice. When applied to an ongoing contract, the Rayleigh estimates are updated as each new month of data becomes available. Therefore, we replicated that procedure in our tests (by increasing the sample size for a given contract) to examine how sensitive our prediction profiles over the life of a contract were to alternative bounds for the d parameter.

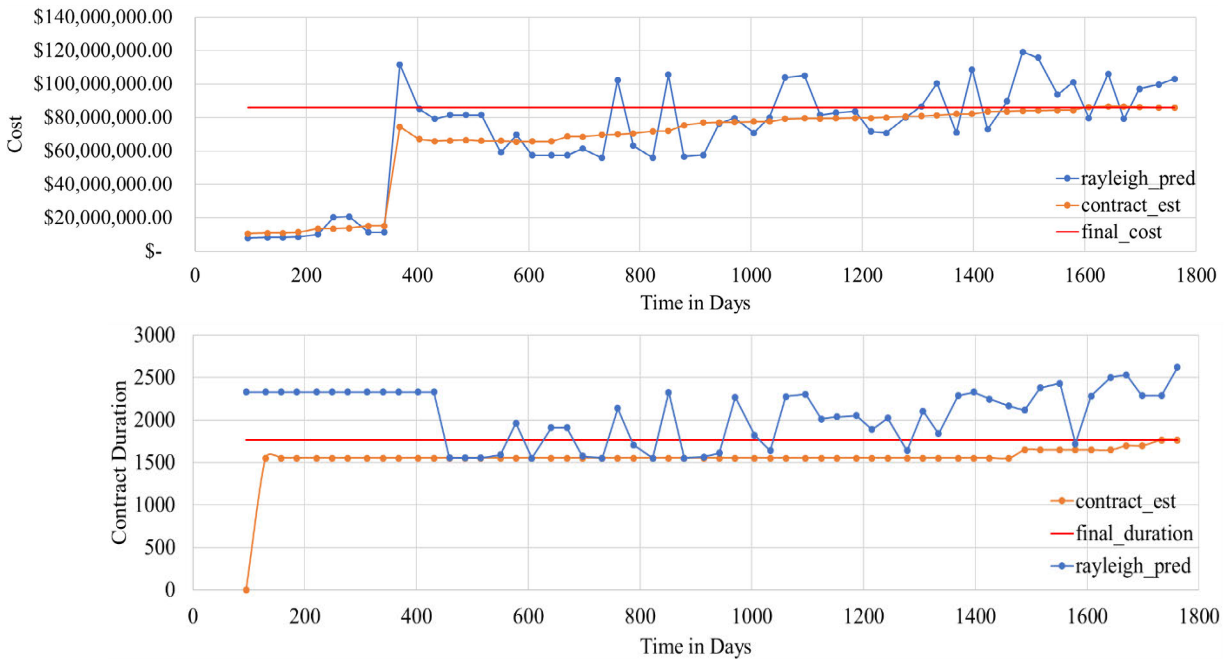
Table B.1. Percentage Difference Between Rayleigh Predictions and Contract Estimates

Run	Cost		Duration	
	Average % Difference	Standard Deviation (% points)	Average % Difference	Standard Deviation (% points)
1	27.1	5.6	18.0	4.3
2	17.8	1.8	19.8	3.5
3	19.2	4.3	21.6	6.3
4	17.4	3.0	20.0	9.1
5	16.7	4.8	22.6	8.0
6	16.8	6.0	23.4	8.3
7	33.5	40.5	20.2	5.2

Contract Selection Heuristic

With the parameter bounds described above, we fit curves to the cumulative EV data of sample contracts. Using the same sample of six contracts used for the bounding sensitivity analysis, we generated curve fits using the Rayleigh model, starting with the initial three data points reported and increasing the fit sample size until all the data points were included (i.e., simulating the subsequent application of the Rayleigh model over time, starting with the receipt of three reports). For each contract (or contract task, if reported separately), we plotted graphs comparing the Rayleigh predictions for contract cost and duration, the reported estimates, and the final reported values. Figure B.2 shows examples of these prediction plots for cost and duration of a sample contract.

Figure B.2. Example Rayleigh Cost and Duration Predictions



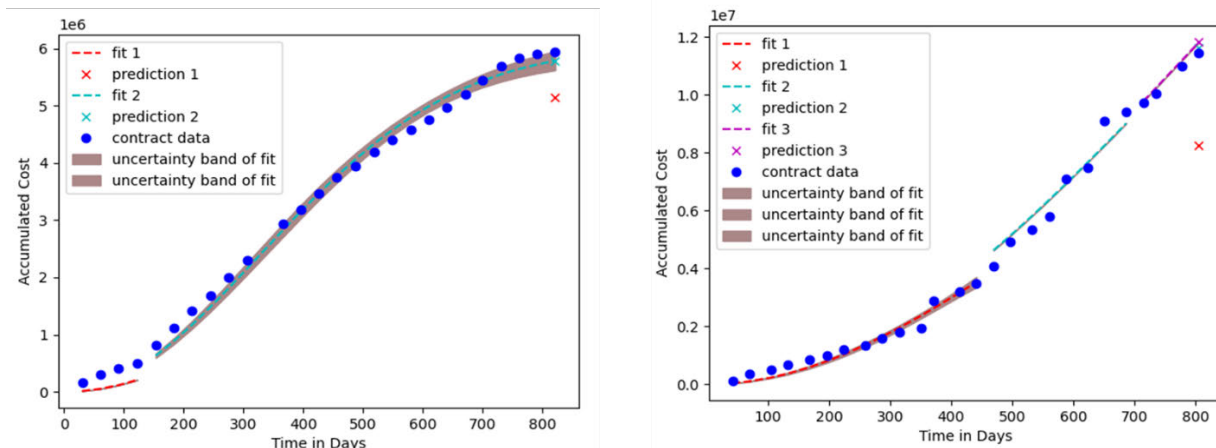
Adjusting for Work Added to a Contract

Taking a deeper look into the reported data for the sample contracts, we saw that increases in EAC and CBB were common throughout a contract. For example, the EAC cost (the lower orange line in the first plot of Figure B.2) had step increases, starting around day 1,480. Analysis of the data in these reports found that the CBB for these contracts also showed significant increases. This confirmed that these contracts had significant work added to them in the course of the contract, causing the sudden increase in EAC. Because these increases in EAC are not obvious by only looking at the cumulative cost of work performed on a contract, we wanted to explore an additional method to identify these changes and adjust the fitting method accordingly.

Because a Rayleigh Function conceptually reflects the usual shape of cumulative ACWP over time for a fixed set of work, a significant increase in CBB would be expected to present an issue when using prior cumulative costs for the prior CBB. Therefore, we investigated using a rule that checks for significant increases in CBB between consecutive contract reports to determine whether a new Rayleigh Function is needed at these points. For this research, we defined significant increases to be a percentage change of 20 percent or more in consecutive

CBB reports.³² Whenever a change of this magnitude was identified, a new curve was fitted using only data points from that point forward. Figure B.3 shows two examples of this approach. Here, the accumulated costs for the contract are shown in blue dots, and the multiple curves fitted are shown in dashed lines. For each curve fitting, a prediction for the last reported date is shown by a red “X.”

Figure B.3. Multiple Curves Fitting for Work-Added Adjustments



Selecting Appropriate Contracts for the Rayleigh Model

We compared the average R^2 obtained from fitting a single Rayleigh curve with the average R^2 obtained using the multiple curve approach on a sample of 116 completed Air Force contracts or contract tasks.³³ The average R^2 doubled in value when changing from the single to the multiple curve approach in cases of significant work added to the contract. However, the R^2 was still relatively low and close to zero. On closer inspection of the cost curves, it became apparent that not all contracts closely follow the S-curve pattern of the Rayleigh Function. This could be for a variety of reasons, ranging from the type of system the contract is associated with, the type of task within the contract, or other significant changes in the project because of renegotiations or large-scale budget adjustments. Some of the publications reviewed on this topic only research

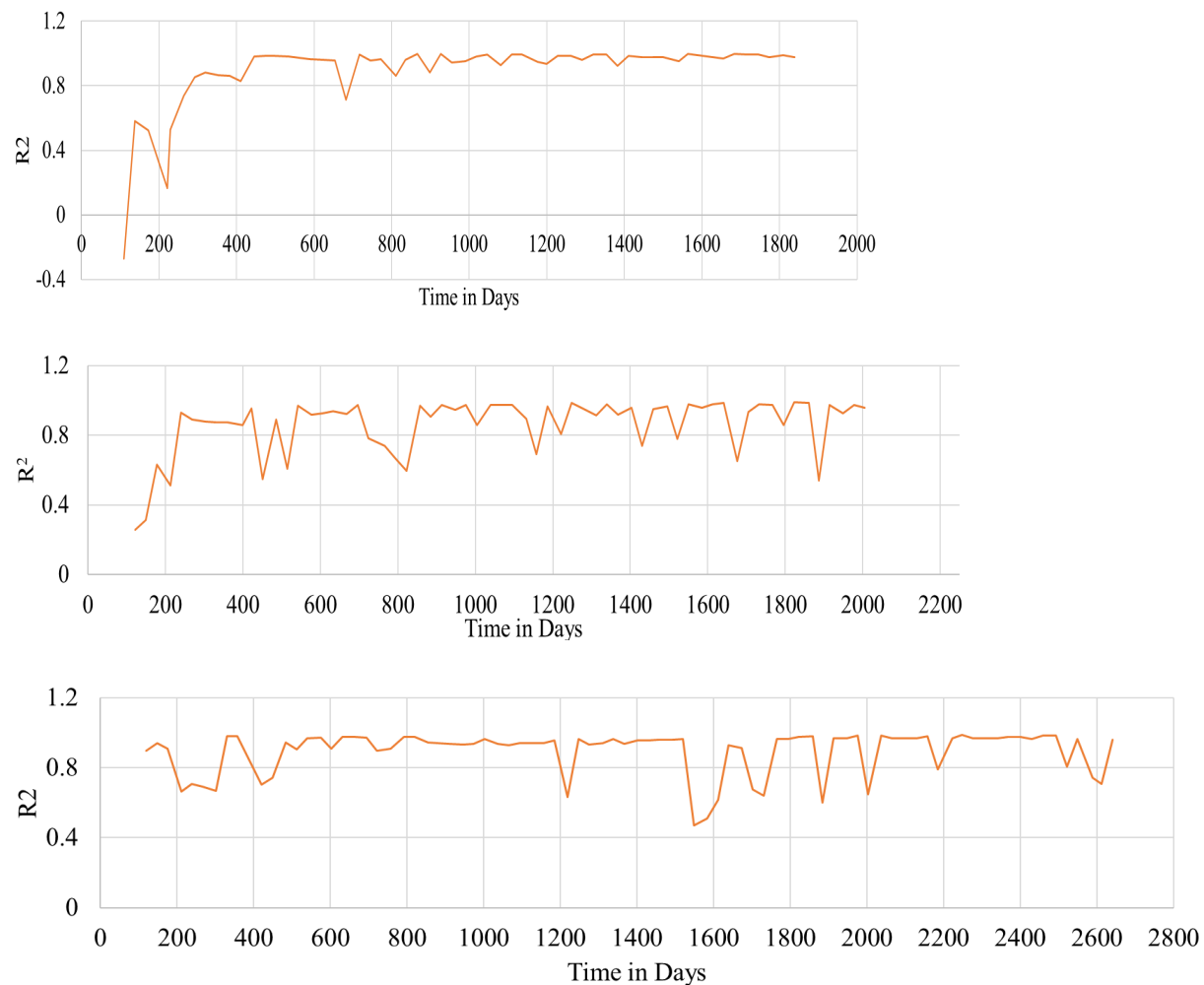
³² There are numerous other ways to identify structural breaks in the data. Doing so is challenging in this context because of the use of a non-linear model on relatively short time series where the goal is to update forecasts quickly as new data become available. We believe substantial change in the CBB is an indicator of a structural break and chose a relatively conservative threshold of a 20-percent change to signify a break in the curve. However, there may be other indicators of a structural break. Future research could further explore the factors that govern structural breaks; this knowledge could help improve the predictive accuracy of the model.

³³ R^2 is a commonly used measure for evaluating goodness of fit. See, for example, Davis, Christle, and Abba, 2009.

projects, such as aircraft (Brown et al., 2015) or information technology (Yang et al., 2019) development.

In response, we developed a preliminary heuristic rule using an R^2 threshold test to select which contracts behave in a way that is more appropriate for predictions using the Rayleigh model. First, we manually collected a subset of six contracts from our sample that showed cumulative costs that closely followed an S-curve pattern. We then fitted Rayleigh curves to these contracts over their reported time period, calculating the R^2 of each resulting curve fit. Figure B.4 shows a sample of R^2 plots generated from curve fitting this subset.

Figure B.4. R^2 for Example Contracts That Display an S-Curve Shape Over Time



From these graphs, we noticed a trend of R^2 improving in the initial days of the project and then leveling out at or above a value of 0.80 near 300 days (ten months). The plots also show that there is still some noise in the R^2 values after the initial 300 days, but the graphs seem to continue to oscillate around an average value of over 0.80. Because the sample of contracts used

to create these plots appeared to visually follow a Rayleigh curve, we defined this behavior as the heuristic rule to decide whether a particular contract task should be included in the set of contracts to generate Rayleigh predictions. In other words, if the Rayleigh Function explains nearly all of the variation in the data (i.e., 80 percent, on average), then applying a Rayleigh Function to predict final cost and schedule should be a reasonable approach. Further research may determine whether 80 percent is too high a criterion given that, in our experience, an R^2 of 80 percent in acquisition is a very good fit.

Therefore, we developed a preliminary Rayleigh metric procedure as follows. First, we collected the initial data points for each contract task, up to ten months. If work added is identified within this period, then the count to ten months is restarted from the report with increased CBB. A Rayleigh curve is then fitted using the identified data points and the R^2 is calculated. If a contract task obtains an R^2 of at least 0.8, then it is marked as an acceptable contract to use for the Rayleigh model prediction analysis. Further research could refine the approach to form a Rayleigh prediction in less than ten months but with higher uncertainty bounds for the prediction (e.g., see Davis, Christle, and Abba, 2009). In the following section, we discuss the results obtained from testing this rule out on a larger sample of contracts and the final assessment output proposed for PIERCE.

Implementation and Assessment Rating

Table B.2 describes the data used for our analysis. From the sample of 35 completed Air Force contracts, we identified a total of 116 contract tasks with reported cumulative cost in the EVM database. We used the heuristic selection rule on these contract tasks; 44 percent of the 116 tasks were over the threshold with a R^2 above 0.80. The contract tasks that passed the selection rule represented 65 percent of the 35 contracts selected. We calculated the percentage difference between the Rayleigh predictions for the tasks that passed the test and their reported estimates of cost and duration. Overall, the absolute value average percentage difference for both contract cost and duration was 30 percent. This is a significant result because a 2009 work also reported an estimated discrepancy between Rayleigh predictions and contract estimates of 30 percent (Davis, Christle, and Abba, 2009). This supports our assessment that the heuristic selection rule is a simple and effective way of selecting a subset of contracts that should provide effective predictions when the Rayleigh model is applied. We implemented this rule on the larger EVM data set. From the contract tasks that passed the rule, we developed an assessment structure to create an output that could be included in future versions of PIERCE.

Table B.2. EVM Data Summary Statistics

	Cost (\$, millions)					Duration (months)				
	Mean	Median	σ	IQR	<i>N</i>	Mean	Median	σ	IQR	<i>N</i>
All	824.021	103.74	5,410.026	381.72	1,110	58.866	52.73	38.496	37.78	957
Sample of completed contracts	1,257.780	52.08	11,755.620	92.58	115	56.030	47.36	37.845	50.20	106

NOTES: The EVM data span 2007 to 2020. *N* refers to the number of unique contract tasks with observations for each respective variable. The entire sample covers 204 programs, 596 contracts, and 1,161 contract tasks. Of the 1,161 contract tasks, 12.49 percent correspond to U.S. Army contracts, 41.69 percent correspond to Navy contracts, 27.3 percent correspond to Air Force contracts, 9.22 percent correspond to DoD contracts, and 9.3 percent correspond to Missile Defense Agency contracts. For contracts not yet completed, cost and duration reflect estimates as of the most recent month of reporting. Out of 1,161 total contract tasks, 1,110 contained EAC information and 957 contained duration information. The sample of completed contracts consists exclusively of DAF contracts. It covers 20 programs, 35 contracts, and 116 contract tasks. Cost is expressed in millions of U.S. dollars, and duration is expressed in months. IQR = interquartile range; σ = standard deviation. While the table shows that the contracts have a wide (skewed) range of magnitudes, our analysis is *not* of the relative magnitudes of cost and duration across the sample but rather the degree to which individual contracts in the sample (regardless of absolute magnitude) exhibit an S-shaped curve with regard to cumulative cost or schedule over time.

Assessment Rating for Prototype Output

The larger EVM data set included a total of 1,161 contract tasks. After using the selection rule, 40 percent of the tested tasks passed the 0.80 R^2 threshold. For the resulting 465 contract tasks, we fitted Rayleigh curves iteratively, starting with the initial three reported data points and increasing the sample size in every iteration until every data point was included in the curve. For each curve-fitting iteration, we calculated the Rayleigh predictions of final cost and duration and compared them with the reported estimates. A green, yellow, or red rating could be assigned based on the percentage difference between the predictions and report estimates. Possible decision criteria were discussed in Chapter 2 and illustrated in Table 2.13.

Using these criteria, the final assessment of the EVM data showed that 24 percent of the contract tasks evaluated had a red rating, while 58 percent were marked with a green rating. The goal of this rating is to serve as an indicator for the project manager and other relevant stakeholders to examine contracts with a red or yellow rating to identify any issues and make any necessary corrections.

Future Research

The proposed assessment criteria use only a simplified version of the output obtained from the Rayleigh analysis. A potential area for future research is to improve the decision criteria by incorporating additional information from the Rayleigh curve fit (e.g., confidence intervals for the parameters). For our preliminary analysis, we calculated confidence intervals for both the *d*

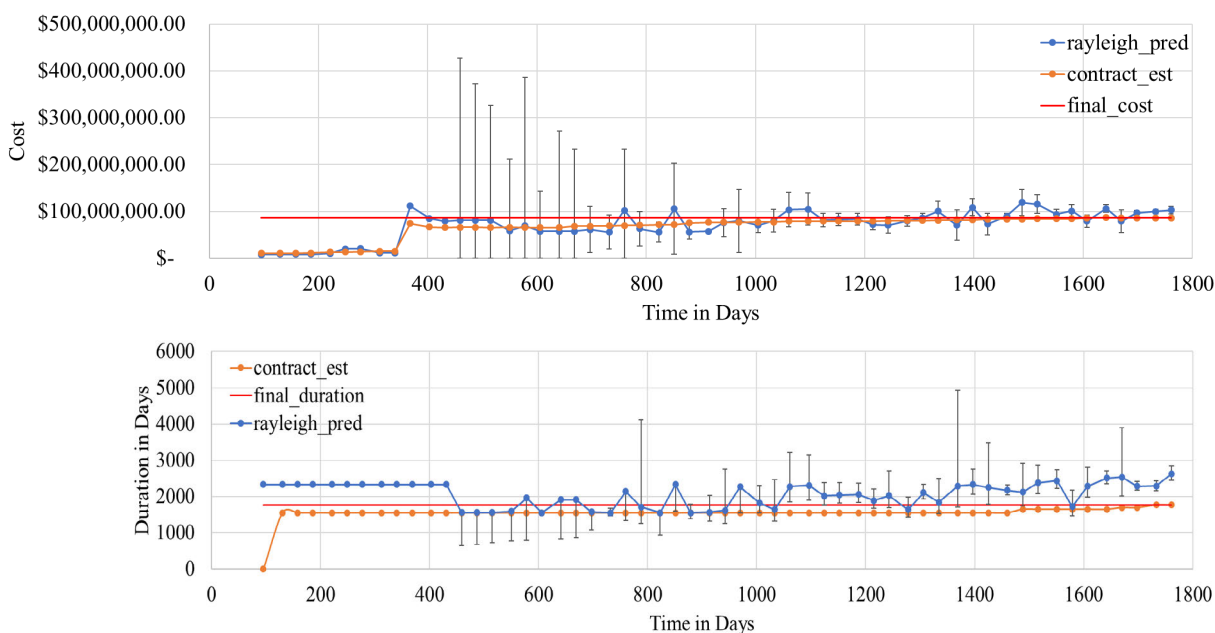
and α parameters through assuming that the error terms and estimators were normally distributed. This allowed us to apply the following formula to calculate 95-percent confidence intervals for both parameters:

$$\beta_i \pm 1.96 SE(\beta_i) ,$$

where β_i is the parameter estimate obtained from the Rayleigh curve fit and $SE(.)$ is the associated standard error. However, oftentimes—particularly relatively early on in a contract—the parameters are estimated on a fairly small sample, meaning leveraging the asymptotic properties of estimators may not be appropriate. Future research should explore alternative means for constructing confidence regions.³⁴

Figure B.5 shows examples of the confidence intervals when shown on prediction plots for both cost and duration. Including prediction plots, such as these, in future versions of PIERCE could improve the assessment of contracts because they would provide a better picture of the behavior of the contracts over time and can be recognized visually at a glance.

Figure B.5. Prediction Plots with Confidence Intervals



Another area of potential future research is to improve on how to handle cases where major work content is added to contracts. For this research, we used a 20-percent increase in CBB as

³⁴ See Davis, Christle, and Abba, 2009, for another approach to estimating confidence regions.

the threshold to start a new curve fit. However, a more robust threshold rule might be developed that better reflects how increases in CBB affect the outcome of contracts. Also, more-complex methods of incorporating multiple curve fittings could be developed to create predictions for the contracts.

Abbreviations

ACAT	Acquisition Category
ACWP	Actual Cost of Work Performed
AD	actual duration
BAC	Budget at Completion
BCWP	Budgeted Cost for Work Performed
BCWS	Budgeted Cost for Work Scheduled
BLS	U.S. Bureau of Labor Statistics
BI	business intelligence
CAC	common access card
CBB	Contract Budget Base
COTG	cost-over-target growth
CPI	cost performance index
CPARS	Contractor Performance Assessment Reporting System
DAES	Defense Acquisition Executive Summary
DAF	Department of the Air Force
DAMIR	Defense Acquisition Management Information Retrieval
DAU	Defense Acquisition University
DAVE	Defense Acquisition Visibility Environment
DCMA	Defense Contract Management Agency
DoD	U.S. Department of Defense
EAC	estimate at completion
ECD _{CTR}	contractor's estimated completion date
ES	earned schedule
ETL	Extract, Transform, and Load
eTools	Electronic Tools (Defense Contract Management Agency)
EV	earned value

EVM	earned-value management
EVM-CR	Earned-Value Management—Central Repository
FPDS-NG	Federal Procurement Data System—Next Generation
FSRS	Federal Funding Accountability and Transparency Act Subaward Reporting System
GAO	U.S. Government Accountability Office
IEAC	independent (cost) estimate at completion
IECD _{es}	independent estimated completion date based on earned schedule
LightGBM	Light Gradient Boosted Machine
LMFIT	Non-Linear Least-Squares Minimization and Curve-Fitting
LRE	Latest Revised Estimate
M&As	mergers and acquisitions
MAR	(DAF) Monthly Acquisition Report
MDA	milestone decision authority
MDAP	major defense acquisition program
NASA	National Aeronautics and Space Administration
NCC	Negotiated Contract Cost
NDIA	National Defense Industrial Association
OBS	Organizational Breakdown Structure
OSD	Office of the Secretary of Defense
OTB	Over Target Baseline
PAF	Project AIR FORCE
PAR	Program Assessment Report
PDWR	planned duration of work remaining
PEO	program executive officer
PIEE	Procurement Integrated Enterprise Environment
PIERCE	Performance Indicators and Execution Risks of Contractors Early On
PM	program manager
PMRT	Project Management Resource Tools

R&D	research and development
RD	remaining duration
RDT&E	research, development, test, and evaluation
SAR	Selected Acquisition Report
SCP	service cost position
SEC	U.S. Securities and Exchange Commission
SPI	Schedule Performance Index
SPI_t	Time-Based Schedule Performance Index
TCPI	To-Complete Performance Index
$TCPI_{EAC}$	To-Complete Performance Index to Meet the Estimate at Completion
$TSPI_{ed}$	To-complete Schedule Performance Index to Meet the Estimated Duration
USD(AT&L)	Under Secretary of Defense for Acquisition, Technology, and Logistics
WBS	Work Breakdown Structure
WCG	work content growth

References

- Air Force Instruction 63-101, *Acquisition and Sustainment Life Cycle Management*, Washington, D.C.: Department of the Air Force, August 3, 2011. As of September 11, 2020:
<https://www.acq.osd.mil/dpap/ccap/cc/jcchb/Files/FormsPubsRegs/Pubs/AFI63-101.pdf>
- Alteryx, homepage, undated. As of December 27, 2019:
<https://www.alteryx.com>
- Anton, Philip S., William Shelton, James Ryseff, Stephen B. Joplin, Megan McKernan, Chad J. R. Ohlandt, and Samantha Cohen, *Relative Contractor Risks: A Data-Analytic Approach to Early Identification*, Santa Monica, Calif.: RAND Corporation, RR-A433-1, 2022. As of April 25, 2022:
https://www.rand.org/pubs/research_reports/RRA433-1.html
- Bird, Steven, Edward Loper, and Ewan Klein, *Natural Language Processing with Python*, Boston, Mass.: O'Reilly Media Inc., 2009.
- Bounds, Thomas, Andrew Clark, Todd Henry, John Nierwinski, Suzanne Singleton, and Brian Wilder, *Army Independent Risk Assessment Guidebook*, Aberdeen Proving Ground, Md.: U.S. Army Materiel Systems Analysis Activity, TR-2014-19, April 2014. As of July 31, 2019:
<https://apps.dtic.mil/dtic/tr/fulltext/u2/a599874.pdf>
- Brown, Gregory E., Edward D. White, Jonathan D. Ritschel, and Michael J. Seibel, "Time Phasing Aircraft R&D Using the Weibull and Beta Distributions," *Journal of Cost Analysis and Parametrics*, Vol. 8, No. 3, December 2015, pp. 150–164.
- Burbeck, Steve, "Applications Programming in Smalltalk-80™: How to Use Model-View-Controller (MVC)," *Smalltalk-80*, 1987. As of September 1, 2020:
<https://web.archive.org/web/20120729161926/http://st-www.cs.illinois.edu/users/smarch/st-docs/mvc.html>
- Cer, Daniel, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil, "Universal Sentence Encoder," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Brussels, Belgium: Association for Computational Linguistics, November 2018, pp. 169–174.
- Christensen, David S., and Scott R. Heise, "Cost Performance Index Stability," *National Contract Management Journal*, Vol. 25, No. 1, 1993, pp. 7–15.

- Christensen, David S., and K. I. Payne, “Cost Performance Index Stability—Fact or Fiction?” *Journal of Parametrics*, Vol. 12, No. 1, 1992, pp. 27–40.
- Christensen, David, and Carl Templin, “EAC Evaluation Methods: Do They Still Work?” *Acquisition Review Quarterly*, Vol. 9, No. 1, 2002, pp. 105–116.
- Clayson, Denis S., Alfred E. Thal, Jr., and Edward D. White III, “Cost Performance Index Stability: Insights from Environmental Remediation Projects,” *Journal of Defense Analytics and Logistics*, Vol. 2, No. 2, 2018, pp. 94–109.
- Coleman, Richard L., Megan E. Dameron, Heather F. Chelson, Jessica R. Summerville, and Steve L. Van Drew, “Predicting Final CPI,” briefing, 4th Joint Annual ISPA/SCEA International Conference, Orlando, Fla., June 2003. As of December 11, 2020: <http://www.iceaaonline.com/ready/wp-content/uploads/2017/11/Predicting-Final-CPI-Coleman.pdf>
- Contractor Performance Assessment Reporting System, *Guidance for the Contractor Performance Assessment Reporting System (CPARS)*, Portsmouth, N.H.: U.S. Department of Defense, June 2020. As of September 10, 2020: <https://www.cpars.gov/documents/CPARS-Guidance.pdf>
- CPARS—See Contractor Performance Assessment Reporting System.
- Davis, Dan, Gary Christle, and Wayne Abba, *Using the Rayleigh Model to Assess Future Acquisition Contract Performance and Overall Contract Risk*, Arlington, Va.: Center for Naval Analysis, CRM D0019289.A2, January 2009.
- Davis, Gregory A., Travis L. DePriest, Brian G. Gladstone, Laura A. Hildreth, and Miranda G. Seitz-McLeese, *Evaluating and Predicting Contract Performance Using Machine Learning: A Feasibility Study*, Alexandria, Va.: Institute for Defense Analyses, D-14275, September 2020. As of December 16, 2020: <https://apps.dtic.mil/sti/pdfs/AD1121596.pdf>
- DCMA—See Defense Contract Management Agency.
- Defense Acquisition University, *Earned Value Management “Gold Card,”* Fort Belvoir, Va.: U.S. Department of Defense, September 4, 2018. As of August 17, 2020: <https://www.acqnotes.com/Attachments/DAU%20EVM%20Gold%20Card.pdf>
- Defense Contract Management Agency, “About the Agency,” webpage, undated. As of September 9, 2020: <https://www.dcm.mil/About-Us/>
- Defense Contract Management Agency Instruction 205, *Major Program Support*, Washington, D.C.: U.S. Department of Defense, Engineering and Analysis Directorate, December 4, 2013.

As of September 9, 2020:

<https://www.dema.mil/Portals/31/Documents/Policy/DCMA-INST-205.pdf>

Department of Defense Instruction 5000.02, *Operation of the Adaptive Framework*, Washington, D.C.: U.S. Department of Defense, January 23, 2020. As of January 11, 2021:

<https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/500002p.pdf?ver=2020-01-23-144114-093>

Department of Defense Instruction 8510.01, *Risk Management Framework (RMF) for DoD Information Technology (IT), Incorporating Change 2*, Washington, D.C.: U.S. Department of Defense, March 12, 2014, effective July 28, 2017. As of July 31, 2019:

<https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/851001p.pdf?ver=2019-02-26-101520-300>

Deputy Assistant Secretary of Defense for Systems Engineering, *Department of Defense Risk, Issue, and Opportunity Management Guide for Defense Acquisition Programs*, Washington, D.C.: U.S. Department of Defense, 2017.

Defense Federal Acquisition Regulation Supplement, Subpart 234.2, “Earned Value Management System,” revised December 7, 2011. As of December 14, 2020:

https://www.acq.osd.mil/dpap/dars/dfars/html/current/234_2.htm

DoD—*See* U.S. Department of Defense.

DoDI—*See* Department of Defense Instruction.

Domo, homepage, undated. As of December 27, 2019:

<https://www.domo.com/>

Dukovich, John, Scott Houser, and David A. Lee, *The Rayleigh Analyzer*, Volume 1: *Theory and Applications*, Bethesda, Md.: Logistics Management Institute, October 10, 1999. As of October 21, 2021:

<https://apps.dtic.mil/sti/citations/ADA371360>

Dundas BI, homepage, undated. As of September 10, 2020:

<https://www.dundas.com>

Federal Procurement Data System—Next Generation, “FPDS-NG FAQ,” website, undated. As of December 12, 2020:

https://www.fpds.gov/wiki/index.php/FPDS-NG_FAQ

FPDS-NG—*See* Federal Procurement Data System—Next Generation.

GAO—*See* U.S. Government Accountability Office.

Henderson, Kym, “Further Developments in Earned Schedule,” *Earnedschedule.com*, Spring 2004. As of December 11, 2020:

<https://www.earnedschedule.com/Docs/Further%20Developments%20in%20Earned%20Schedule.pdf>

———, “Earned Schedule in Action,” Earnedschedule.com, Spring 2005. As of December 11, 2020:

<https://www.earnedschedule.com/Docs/Earned%20Schedule%20in%20Action.pdf>

———, “Earned Schedule: A Breakthrough Extension to Earned Value Theory? A Retrospective Analysis of Real Project Data,” *Proceedings of PMI Global Congress 2007—Asia Pacific*, Hong Kong: Project Management Institute, January 2007, pp. 13–23. As of December 11, 2020:

<https://www.earnedschedule.com/Docs/Earned%20Schedule%20-%20A%20Breakthrough%20Extension%20to%20EVM.pdf>

Henderson, Kym, and Ofer Zwikael, “Does Project Performance Stability Exist? A Re-Examination of CPI and Evaluation of SPI(t) Stability,” *CrossTalk: The Journal of Defense Software Engineering*, Vol. 21, No. 4, April 2008. As of December 11, 2020:

https://www.researchgate.net/publication/240460093_Does_Project_Performance_Stability_Exist_A_Re-examination_of_CPI_and_Evaluation_of_SPIt_Stability

iDashboards, homepage, undated. As of December 27, 2019:

<https://www.idashboards.com>

InsightSquared, homepage, undated. As of December 27, 2019:

<https://www.insightsquared.com/>

International Standard 21508, *Earned Value Management in Project and Programme Management*, Geneva, Switzerland: International Standard Organization, ISO 21508:2018(E), April 2018.

Joint Staff J-8, *JCIDS Manual: Manual for the Operation of the Joint Capabilities Integration and Development System*, Washington, D.C., August 31, 2018.

Joseph, Brian B., and Darris Sconion, “Using Natural Language Processing, Sentiment Analysis, and Text Mining to Determine if Text in Selected Acquisition Report Executive Summaries Are Highly Correlated with Major Defense Acquisition Program (MDAP) Unit Costs and Can Be Used as a Variable to Predict Future MDAP Costs,” *Proceedings of the Seventeenth Annual Acquisition Research Symposium*, SYM-AM-20-061, Monterey, Calif.: Acquisition Research Program, Naval Postgraduate School, April 13, 2020. As of December 16, 2020: https://www.acq.osd.mil/aap/assets/docs/SYM-AM-20-061_Panel11_Joseph_Paper_4-13-2020.pdf

Ke, Guolin, Qi Meng, Thomas Finely, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu, *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*, Redmond,

- Wash.: Microsoft Research, December 2017. As of September 11, 2020:
<https://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree>
- Lee, David A., Michael R. Hogue, and Mark A. Gallagher, *Determining a Budget Profile from a Development Cost Estimate*, Washington, D.C.: U.S. Department of Defense, Director Program Analysis And Evaluation, Office of the Secretary of Defense, September 1, 1993. As of August 14, 2020:
<https://apps.dtic.mil/sti/citations/ADA275864>
- Lipke, Walt, “Schedule Is Different,” Earnedschedule.com, March 2003. As of August 18, 2020:
<https://www.earnedschedule.com/Docs/Schedule%20is%20Different.pdf>
- , “Independent Estimates at Completion: Another Method,” *CrossTalk: The Journal of Defense Software Engineering*, October 2004, pp. 26–30. As of December 13, 2020:
https://www.researchgate.net/publication/289791379_Independent_estimates_at_completion_-_Another_method
- , “Project Duration Forecasting . . . A Comparison of Earned Value Management Methods to Earned Schedule,” Earnedschedule.com, Spring 2009. As of December 11, 2020:
<https://www.earnedschedule.com/Docs/Project%20Duration%20Forecasting.pdf>
- , “Introduction to Earned Schedule,” *PM World Journal*, Vol. 3, No. 11, November 2014. As of August 18, 2020:
<https://www.earnedschedule.com/Docs/Lipke-introduction-to-earned-schedule-utd.pdf>
- LMFIT, “Non-Linear Least-Squares Minimization and Curve-Fitting for Python,” webpage, undated. As of August 21, 2020:
<https://lmfit.github.io/lmfit-py/>
- Marine Corps Systems Command, *MARCORSYSCOM Acquisition Guidebook (MAG)*, Quantico, Va., October 30, 2020.
- McKernan, Megan, Nancy Young Moore, Kathryn Connor, Mary E. Chenoweth, Jeffrey A. Drezner, James Dryden, Clifford A. Grammich, Judith D. Mele, Walter T. Nelson, Rebeca Orrie, Douglas Shontz, and Anita Szafran, *Issues with Access to Acquisition Data and Information in the Department of Defense: Doing Data Right in Weapon System Acquisition*, Santa Monica, Calif.: RAND Corporation, RR-1534-OSD, 2017. As of December 26, 2019:
https://www.rand.org/pubs/research_reports/RR1534.html
- Microsoft, “LightGBM,” homepage, undated. As of September 11, 2020:
<https://lightgbm.readthedocs.io/>
- NASA—See National Aeronautics and Space Administration.

- National Aeronautics and Space Administration, *Earned Value Management Reference Guide for Project-Control Account Managers*, Washington, D.C., NASA/SP-3708, November 1, 2019. As of August 17, 2020:
https://www.nasa.gov/sites/default/files/atoms/files/reference_guide_for_project_cams_sp-3708_-_11_01_2019.docx
- , *NASA Schedule Management Handbook, Revision 1*, Washington, D.C., NASA/SP-2010-3403, January 30, 2020. As of August 18, 2020:
https://www.nasa.gov/sites/default/files/atoms/files/nasa_schedule_management_hand_book.pdf
- National Defense Industrial Association, *A Guide to Managing Programs Using Predictive Measures, Revision 2*, Arlington, Va.: Integrated Program Management Division, September 12, 2017. As of August 14, 2020:
<https://www.ndia.org/-/media/sites/ndia/divisions/ipmd/ndia-ipmd-predictivemeasuresguide-rev-2-july312017.ashx?la=en>
- Natural Language Toolkit, homepage, last updated April 13, 2020. As of September 11, 2020:
<https://www.nltk.org>
- Naval Air Systems Command, *NAVAIR Acquisition Guide 2016/2017*, Patuxent River, Md., September 22, 2015.
- NDIA—See National Defense Industrial Association.
- Newville, Matthew, Till Stensitzki, Daniel B. Allen, and Antonino Ingargiola, “LMFIT: Non-Linear Least-Square Minimization and Curve-Fitting for Python,” webpage, September 21, 2014. As of September 9, 2020:
<https://zenodo.org/record/11813#.X1j5Y3lKhE>
- NLTK—See Natural Language Toolkit.
- Norden, Peter V., “Useful Tools for Project Management,” in Starr M. K., ed., *Management of Production*, Baltimore, Md.: Penguin Books, 1970, pp. 71–101.
- Oracle Solutions Business Intelligence, homepage, 2019. As of December 27, 2019:
<https://www.oracle.com/solutions/business-analytics/business-intelligence/>
- Orad, Amir, “Periscope Data Is Now Sisense for Cloud Data Teams,” *Sisense*, blog post, January 7, 2020. As of September 10, 2020:
<https://www.sisense.com/blog/periscope-data-is-now-sisense-for-cloud-data-teams/>
- Parker, William, *Defense Acquisition University Program Managers Tool Kit*, 16th ed., Fort Belvoir, Va.: Defense Acquisition University, January 2011.

- Petter, Jacob L., *An Analysis of Stability Properties in Earned Value Management's Cost Performance Index and Earned Schedule's Schedule Performance Index*, Air Force Institute of Technology, Graduate School of Engineering and Management, theses and dissertations, 719, March 14, 2014. As of December 13, 2020:
<https://scholar.afit.edu/etd/719>
- Public Law 114-328, National Defense Authorization Act for Fiscal Year 2017, Section 235, Pilot Program on Disclosure of Certain Sensitive Information to Federally Funded Research and Development Centers, (g) Sensitive Information Defined, December 23, 2016. As of August 9, 2021:
<https://www.congress.gov/114/plaws/publ328/PLAW-114publ328.pdf>
- Reenskaug, Trygve, “THING-MODEL-VIEW-EDITOR: An Example from a Planning System,” May 12, 1979a. As of September 1, 2020:
<http://heim.ifi.uio.no/~trygver/1979/mvc-1/1979-05-MVC.pdf>
- , “MODELS – VIEWS – CONTROLLERS,” December 10, 1979b. As of September 1, 2020:
<http://heim.ifi.uio.no/~trygver/1979/mvc-2/1979-12-MVC.pdf>
- , “The Model-View-Controller (MVC): Its Past and Present,” August 20, 2003. As of September 1, 2020:
http://heim.ifi.uio.no/~trygver/2003/javazone-jao0/MVC_pattern.pdf
- SAP Crystal Reports, homepage, undated. As of December 27, 2019:
<https://www.sap.com/products/crystal-reports.html?infl=05f03c7f-c3f8-4f62-8266-8f79fec692f2>
- Schwartz, Moshe, and Heidi M. Peters, *Acquisition Reform in the FY2016–FY2018 National Defense Authorization Acts (NDAAs)*, Washington, D.C.: Congressional Research Service, R45068, January 19, 2018. As of December 23, 2020:
<https://crsreports.congress.gov/product/pdf/R/R45068/6>
- S&P Global, “S&P Capital IQ Platform,” webpage, undated. As of March 24, 2020:
<https://www.spglobal.com/marketintelligence/en/solutions/sp-capital-iq-platform>
- TensorFlow, “universal-sentence-encoder,” webpage, last updated August 25, 2020. As of September 11, 2020:
<https://tfhub.dev/google/universal-sentence-encoder/4>
- Titov, Nikita, “Machine Learning Challenge Winning Solutions: LightGBM Is Used in Many Winning Solutions,” blog, last updated October 30, 2020. As of August 3, 2021:
<https://github.com/microsoft/LightGBM/blob/master/examples/README.md#machine-learning-challenge-winning-solutions>

Titov, Nikita, Guolin Ke, Darío Hereñú, G. Young, Zach Kurtz, James Lamb, Shihab Shahriar Khan, and Andrew Ziem, “Experiments,” blog, last updated May 4, 2021. As of August 22, 2021:

<https://github.com/microsoft/LightGBM/blob/master/docs/Experiments.rst#comparison-experiment>

Tracy, Steven P., and Edward D. White, “Estimating the Final Cost of a DoD Acquisition Contract,” *Journal of Public Procurement*, Vol. 11, No. 2, 2011, pp. 190–205.

Under Secretary of Defense for Acquisition, Technology, and Logistics, *Performance of the Defense Acquisition System: 2013 Annual Report*, Washington, D.C.: U.S. Department of Defense, 2013. As of March 9, 2018:

<http://www.dtic.mil/docs/citations/ADA587235>

———, *Performance of the Defense Acquisition System: 2014 Annual Report*, Washington, D.C.: U.S. Department of Defense, 2014. As of March 9, 2018:

<http://www.dtic.mil/docs/citations/ADA603782>

———, *Performance of the Defense Acquisition System: 2015 Annual Report*, Washington, D.C.: U.S. Department of Defense, 2015. As of January 2, 2020:

<http://www.dtic.mil/docs/citations/ADA621941>

———, *Performance of the Defense Acquisition System: 2016 Annual Report*, Washington, D.C.: U.S. Department of Defense, October 2016. As of January 2, 2020:

<http://www.dtic.mil/docs/citations/AD1019605>

U.S.C.—See U.S. Code.

U.S. Code, Title 10, Section 2431a, Acquisition Strategy, January 3, 2016.

U.S. Code, Title 10, Section 2433, Unit Cost Reports, January 24, 2020.

USD(AT&L)—See Under Secretary of Defense for Acquisition, Technology, and Logistics.

U.S. Department of Defense, *The DoD Contractor Risk Assessment Guide*, Washington, D.C., October 1988.

———, *Earned Value Management Implementation Guide (EVMIG)*, Washington, D.C., January 2019a. As of August 9, 2021:

<https://www.acq.osd.mil/asda/ae/ada/ipm/docs/DOD%20EVMIG-01-18-2019.pdf>

———, *Earned Value Management System Interpretation Guide (EVMSIG)*, Washington, D.C., March 14, 2019b. As of August 17, 2020:

https://www.acq.osd.mil/asda/ae/ada/ipm/docs/DoD_EVMSIG_14MAR2019.pdf

U.S. Government Accountability Office, *Weapon Systems Annual Assessment: Limited Use of Knowledge-Based Practices Continues to Undercut DOD’s Investments*, Washington, D.C.,

GAO-19-336SP, May 7, 2019. As of December 23, 2020:

<https://www.gao.gov/products/GAO-19-336SP>

———, *GAO Cost Estimating and Assessment Guide: Best Practices for Developing and Managing Capital Program Costs*, Washington, D.C., GAO-20-195G, March 2020a. As of December 13, 2020:

<https://www.gao.gov/assets/710/705312.pdf>

———, *Defense Acquisitions Annual Assessment: Drive to Deliver Capabilities Faster Increases Importance of Program Knowledge and Consistent Data for Oversight*, Washington, D.C., GAO-20-439, June 3, 2020b. As of December 13, 2020:

<https://www.gao.gov/products/GAO-20-439>

Watkins, Harry, *An Application of Rayleigh Curve Theory to Contract Cost Estimation and Control*, Monterey, Calif.: Naval Postgraduate School, 1982.

Yang, Rebekah, Justin Coe, Benjamin Dryer, Matthew Cumpian, Maeve Budi, Matt Prebble, Mary (Jocelyn) Cox, and Shaelynn Hales, “Application of the Rayleigh Model to Predict Information Technology Program Cost and Schedule Performance,” in Stephen Adams, Peter A. Beling, James H. Lambert, William T. Scherer, and Cody H. Felimg, eds., *Systems Engineering in Context*, New York: Springer, 2019, pp. 183–193.



etting early indication of potential contractor performance risks and contract execution issues is critical for proactive acquisition management. When contractors are in danger of not meeting contractual performance goals, Department of the Air Force (DAF) acquisition management may not be fully aware of the shortfall until, for example, a schedule deadline is missed, government testing indicates poor system's technical performance, or costs exceed expectations.

Concerns continue to be raised about cost and schedule growth in acquisition and experts postulate about a lack of knowledge about the status of acquisition programs. In this report, the authors focus on metrics to identify emerging execution problems earlier than traditional acquisition oversight systems to enable more-proactive risk and performance management. They summarize their findings, which include a taxonomy of contractor relative risks, leading indicators of performance, relevant data sources, risk measures and equations, and a prototype that implements some of these findings using real data sources. This research should be of interest to acquisition professionals and leadership who are searching for ways to improve acquisition performance through early identification of potential relative contractor risks and execution problems to inform active program management and mitigation of risks. The prototype should be of interest to acquisition officials (from program managers to milestone decision authorities) to help them access more data in an easy-to-understand way so they can focus their limited time on areas that require increased management attention. This approach should be useful during any phase of the acquisition process.

\$41.00

ISBN-10 1-9774-0954-7
ISBN-13 978-1-9774-0954-6



www.rand.org