# Delusions of Success: Overcoming Optimism Bias in Schedule Forecasting

ICEAA 2022 Professional Development & Training Workshop; Pittsburgh, PA
iceaaonline.com/workshop
ICEAA

**Jeffrey M. Voth[1]; Maxwell C. Moseley, CCEA®; Ann E. Hawpe, CCEA®**

**Abstract**
Recent U.S. Government Accountability Office assessments have found that, in addition to substantial cost growth, major defense acquisition programs experience capability delays of more than two years. To improve schedule performance, stakeholders need to de-risk overly optimistic schedule estimates through data-driven approaches based on realized prior program histories. The authors evaluate the merits of taking an 'outside view' to mitigate optimism bias in schedule estimates through reference class forecasting and present an example leveraging empirical distributional information from 116 programs across six commodity classes to develop more realistic and reliable front-end schedule estimates.

**Keywords**
Risk; bias; uncertainty; statistics; scheduling; data-driven methods; megaprojects

## Introduction

***Background.*** Senate Armed Services Committee leadership publicly denounced the Pentagon's dismal schedule performance track record in a recent *Proceedings* magazine article, citing "absurd acquisition debacles that have set back the country tens of billions of dollars and delayed necessary weapon systems for years" (Inhofe & Reed, 2020). The U.S. Government Accountability Office (GAO) has reported similar critical capability delays in its annual assessments of major defense acquisition programs (MDAPs) for Congressional Committees, also noting that despite these lingering issues, "most MDAPs continue to forgo opportunities to improve cost and schedule outcomes" (GAO, 2021). To begin delivering capabilities at the speed and scale sought by Pentagon leadership, new MDAPs need to improve schedule performance. However, GAO's review of 42 future programs found that insufficient headway is being made to leverage information from past programs to make these improvements (GAO, 2021).

This paper responds to repeated calls from GAO (Schinasi, 2008; Francis, 2015) and emerging research by Professors Robert Mortlock and Nick Dew of the Naval Postgraduate School to curb excessive optimism in planning forecasts (Mortlock & Dew, 2021). GAO Managing Director Paul Francis explains, in written testimony for House Armed Services Committee leadership, how optimism is a dominant cognitive bias that negatively affects MDAP performance:

> *"Competition with other programs vying for defense dollars puts pressure on program sponsors to project unprecedented levels of performance (often by counting on unproven technologies) while promising low cost and short schedules. These incentives, coupled with a marketplace that is characterized by a single buyer (DOD), low volume, and limited number of major sources, create a culture in weapon system acquisition that encourages undue optimism."*
>
> Paul L. Francis, Managing Director
> Government Accountability Office (GAO)
> Acquisition and Sourcing Management
> 27 October 2015

Previous MDAP scheduling estimate efforts have attempted to regress schedule durations against programs' technical characteristics to develop parametric schedule estimating relationships (SERs). The high variability in program schedules, even for similar programs, precluded the development of any statistically meaningful SERs despite methodical and rigorous regression

[1] Corresponding author.
Herren Associates, Inc. email: jeff.voth@jlha.com

analyses (Jardine et al., 2019). While no trends for schedule durations exist across programs, technical characteristics, or timelines, these past program schedules, taken as a whole, can serve as analogies for future programs to provide a more data-driven approach to schedule estimating.

In this paper the authors discuss how to (a) identify an analogous reference class of past, similar programs; (b) establish a probability distribution from selected programs for the schedule duration being forecast; and (c) compare specific programs to the analogous reference class distribution to establish the most likely outcome for the specific program. Datasets of major milestone dates and schedule durations for multiple commodities, produced as part of SER development efforts, are used. These comprehensive, curated datasets and the derived descriptive statistics will ultimately provide insights regarding the implications of schedule dependencies. These datasets and descriptive statistics will be used to further refine parsimonious models to forecast schedule performance more accurately for programs integral to the U.S. Department of Defense.

***The effect of optimism bias.*** Despite efforts to improve faster to meet emerging operational needs, MDAPs continue to be undermined by the effects of optimism bias. In addition to threatening operational capabilities (Tyson, Harmon, & Utech, 1994), these schedule delays also come with a considerable price tag: The DoD reports that "a net stretch-out of development and procurement schedules" has contributed $2.93B in additional costs to programs captured in Selected Acquisition Reports (SARs) (DoD, 2018). Service Chiefs recognize the cost of these delays. "I don't mean to be dramatic," Admiral Mike Gilday, the 32nd Chief of Naval Operations (CNO), noted during the Surface Navy Association's 33rd National Symposium, "but I feel like if the Navy loses its head, if we go off course and we take our eyes off those things we need to focus on, I think we may not be able to recover in this century" (LaGrone, 2021).

In their groundbreaking studies on decision science (Shefrin & Statman, 2003), researchers Kahneman and Tversky (1979a) found that, under conditions of uncertainty, subjects exhibited "unwarranted optimism in the evaluation of the likelihood that a plan will succeed or that a project will be completed on time." This optimism bias is counterintuitive, as greater levels of uncertainty would logically foster greater levels of caution.

In reality, higher levels of uncertainty buoy tendencies towards unwarranted optimism. Examples range from the optimism bias financial analysts exhibit when forecasting earnings, resulting in costly mistakes (Galanti & Vaubourg, 2017), to the overconfidence small business owners place in their ability to succeed in the face of overwhelming odds (Patel & Tsionas, 2021). A poignant reminder of this phenomenon can also be found in response to the COVID-19 pandemic, where comprehensive cross-sectional and longitudinal analyses found that while "most individuals are aware of the risk caused by the pandemic to some extent, they typically underestimate their personal risk relative to others" (Wise et al., 2020).

The process of de-biasing stakeholder optimism requires concerted effort (Kahneman & Tversky, 1979b), as reflected in the results of research sponsored by the Defense Advanced Research Projects Agency (DARPA). Such an effort may include the need for stakeholders to systematically take an 'outside view' through Reference Class Forecasting (RCF), as advocated by Princeton Professor and Nobel laureate Daniel Kahneman (2011). This approach involves identifying and critically analyzing a set of similar programs and using empirical distributional information to de-bias estimates and calculate the likeliest outcome (Flyvbjerg, 2006).

In his seminal book challenging conventional views on decision-making, *Thinking, Fast and Slow*, Kahneman (2011) reviews Flyvbjerg's (2006) work, concluding "this may be the single most important piece of advice regarding how to increase accuracy in forecasting." Today, RCF has proven to be an effective intervention technique—widely referenced across Europe (United Kingdom, Denmark, Germany, Norway, Sweden, Switzerland, and the Netherlands)—in the assessment of large-scale programs (Park, 2021).

Building on Kahneman and Tversky, Flyvbjerg (2008; 2021) found that optimism bias and strategic misrepresentation are two fundamental areas that drive erroneous program forecasts. While a number of important elements—including "requirements, politics, economics, and the system's technological design" (Monaco & White, 2005) and the deployment of "immature technologies" (Blickstein, 2012)—may also contribute to schedule overruns during execution, Flyvbjerg's method to de-risk programs and improve forecasting accuracy focuses on "reducing uncertainty by getting a clear picture of the size and types of uncertainty that apply to

| Step | Action |
|---|---|
| **1** | Identify and describe the schedule or forecast to be evaluated |
| **2** | Establish a benchmark that represents the outside view, against which performance may be measured |
| **3** | Use the benchmark to evaluate performance in the forecast in question |
| **4** | Check the forecaster's track record from other, similar forecasts |
| **5** | Identify additional program risks |
| **6** | Establish the expected outcome |
| **7** | Solicit comments from the forecaster |
| **8** | Conclude whether the forecast is over- or underestimated and by how much |

*Table 1. Eight steps for performing due diligence through RCF (adapted from Flyvbjerg, 2012)*

specific decisions and forecasts" (Flyvbjerg, 2012). Table 1 lists Flyvbjerg's proposed series of sequential steps for taking an 'outside view' and performing due diligence through RCF.

Understanding that schedule delays pose a continued threat to our national security provides a compelling impetus to systematically reduce optimism bias, conduct meaningful risk and uncertainty analyses, and produce more accurate forecasts. Kahneman and Tversky's DARPA-sponsored research (1979b) cautions "the prevalent tendency to underweight, or ignore, distributional information is perhaps the major error of intuitive prediction", identifying scientists as a group of individuals who are "notoriously prone to underestimate the time to complete a project, even when they have considerable experience of past failures to live up to planned schedules." And while many readers may believe that a mindset grounded in analytical, logical, and rational thinking is the right prescription for rose-colored glasses, research suggests that the engineering community is not immune to optimism bias (Kidd, 1970; Buehler, 1994; Valerdi, 2009).

The following section explores the eight steps in Table 1 and highlights Flyvbjerg's pragmatic method to effectively de-bias program front-end estimates using all available distributional information. The authors will show how critically analyzing MDAP schedules using Flyvbjerg's method can serve as a reference to produce more accurate forecasts aligned with the future force design.

## Methods and Findings

***Shifting from theory to practice.*** The first step of RCF, can be implemented in support of the DoD's current $1.85T portfolio of major weapon systems acquisition programs (GAO, 2020). Engineers and analysts incorporate uncertainty and sensitivity analyses using Microsoft Excel and Oracle Crystal Ball across individual Work Breakdown Structure (WBS) elements. For example, analysts use Oracle Crystal Ball software to generate sensitivity analysis charts, which highlight the top factors contributing to variance and help to provide general assessments for each schedule driver.

To account for uncertainty pertinent to the early-stage development of new MDAPs, program sponsors often indicate that a contingency will be reserved above the 50% confidence level. This contingency for uncertainty is often offset by aggressive schedule development assumptions, synergies related to common engineering teams across disparate programs, and improvement curves on manufacturing labor and material. Unfortunately, these forecasting assumptions often do not take previous examples into account, creating what Kahneman and Lovallo (1993) characterize as the 'inside view'.

***Incorporating an outside view*.** The second step of RCF, creating a standard to measure the initial forecast, builds on Kahneman's 'outside view'. To apply this step and establish this standard for MDAPs, the authors of this paper consider a selection of ship; missile; command, control, communications, computers, combat systems, and intelligence (C5I); vehicle; fixed wing; and rotary programs from the U.S. Navy, Army, and Air Force. Detailed schedule data for these programs, initiated from 1962 to 2012, come primarily from congressionally mandated SARs—which contain the programs' realized milestone dates as well as their estimated milestone dates—and are supplemented with publicly available DoD-sponsored research where SARs are unavailable. Specific dates of interest in these datasets include the date of Milestone (MS) B, when a program enters its Engineering & Manufacturing Development (EMD) phase; the date of MS C, when a program enters its Production & Deployment (P&D) phase; and the date of Initial Operating Capability (IOC), along with the estimated date for IOC at the initiating milestone, either MS B or MS C.

An initial excursion using these datasets sought to regress programs' schedule durations with their technical characteristics—compiled from readily

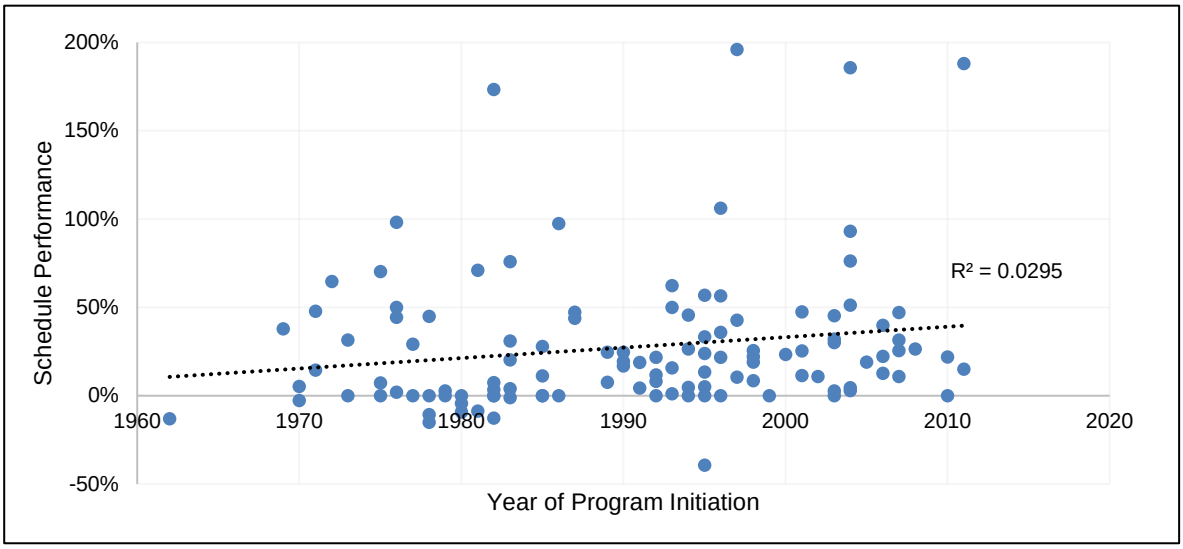


*Figure 1. Plot of MDAP schedule performance by year of program initiation (1962-2012)*

available, public-domain sources—to develop parametric SERs, similar to work by Jimenez et al. (2016). For example, Jardine et al. (2019) analyzed a dataset of 50 missile programs—including cruise missiles, rocket-propelled missiles, and smart munitions—and tested each missile's duration for nine scheduled events against its values for nine technical parameters.

Exhaustive bivariate and multivariate analyses were performed using JMP, a statistical analysis program from SAS, together with the application of standard statistical procedures. The analyses found no combination or transformation of schedule durations and technical parameters that yielded any statistically meaningful SERs. Statistical significance, in this case, was defined as having a *p*-value less than 0.05, a coefficient of determination ($R^2$) greater than 0.60, and a number of observations greater than 10. The program durations varied too drastically for any clear relationship to be discerned (Jardine et al., 2019), even after adjusting the data to account for schedule anomalies specific to each program. Such anomalies included those for missiles with similar technical characteristics, platforms, and genealogies.

Once technical parameters proved to be unsuitable predictors for program schedules, the authors turned to analogies as a method for better schedule forecasting. Where the missile SER effort by Jardine et al. included programs not recorded in SARs, the analogy effort for this paper uses only programs for which both estimated and realized milestone dates exist, resulting in a combined 116 MDAPs across six commodity classes. The inclusion of all available MDAPs—not only the most recent programs or programs perceived to be relevant—prevents the introduction of "judgmental and motivational biases" that arise when choosing to remove older programs from the dataset or programs that exceed an arbitrary threshold (Goodwin & Wright, 2010).

The authors analyzed MDAP schedule performance by year of program initiation. Figure 1 plots each program's year of initiation against its schedule performance, or the relative difference between its actual and estimated initiation-to-IOC durations. Though the initiations for the programs in the dataset span 50 years of changes in technology and acquisition policy, historical schedule performance shows no trend in time. This finding was similarly echoed by the 2020 *Performance of the Defense Acquisition System* series from the Office of the Under Secretary of Defense for Acquisition & Sustainment (DoD, 2020). This lack of any clear improvement or decline in programs' schedule performance from 1962 to 2012 is demonstrated by the poor fit of the trendline in Figure 1, which appears to show schedule overruns increasing over time but has an $R^2$ of 0.0295, falling far below the threshold of 0.60 used by Jardine et al. for statistical significance.

Of the 116 MDAPs involved in the study, roughly three in four programs experienced a schedule delay reaching IOC, with an average overrun of 37.6%, consistent with GAO's assessment (GAO, 2018). Figure 2 presents Beta distributions fitted to the estimated and actual MDAP schedule durations, in months, from program initiation to IOC. The graphic overlays the curves fitted to the frequency distributions of these schedule durations. Here, the dashed line reflects the estimated initiation-to-IOC duration, and the solid line reflects the actual realized initiation-to-IOC duration. The authors identified Beta

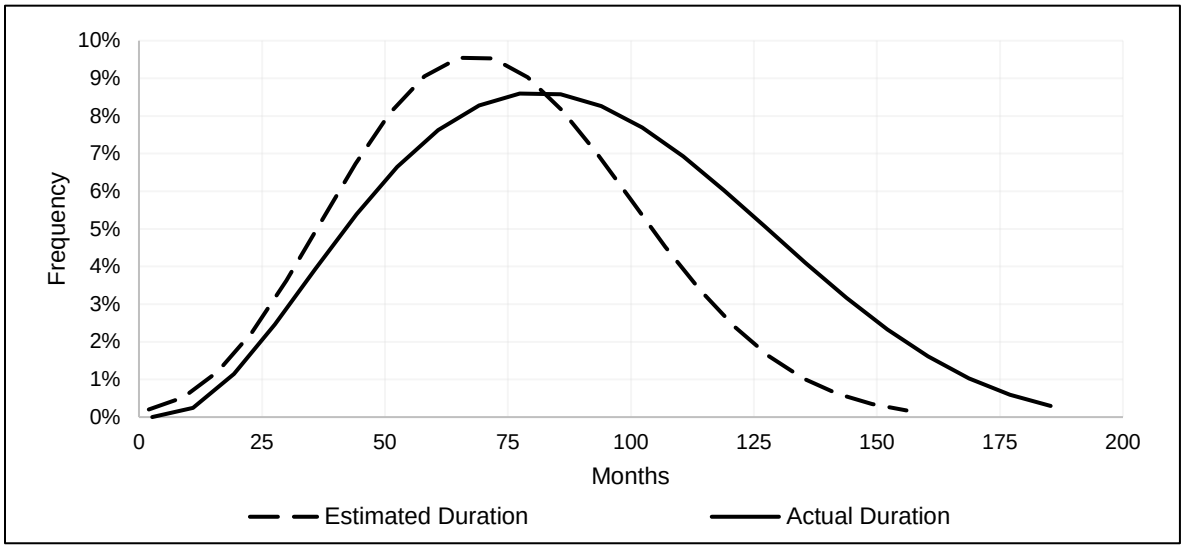


*Figure 2. Beta distributions fitted to the estimated and actual MDAP schedule durations, in months, from program initiation to IOC*

distributions as the best fit for both the estimated and actual duration data. This conclusion is in line with the historical assumption that the variability in an activity's time estimates generally follows the Beta distribution (Lee, 2005).

As illustrated in Figure 2, the estimated durations have a tighter distribution and higher peak than the actual durations, which have a wider, shorter distribution that indicates a greater spread or variability in the data. The distribution for the estimated durations is tighter and taller because programs tend to underestimate their durations, which is indicative of optimism bias.

While all MDAPs are not the same, similar patterns are observed across different commodity classes as depicted in Table 2, which displays some descriptive statistics for the planned and actual schedule durations from program initiation to IOC. The median and mean values for the actual durations exceed those for the estimated durations of each commodity class. Also, the interquartile range (IQR) for the actual durations is wider than the IQR for the estimated durations in all cases except C5I programs. A cursory review of the information in Table 2 indicates that internal program forecasts for schedules are often significantly underestimated from the outset. While the MDAP sample size is limited to the availability of program data collected by the U.S. Government in SARs, significantly underestimating schedule durations demonstrates that the risk of exceeding contingency reserves is extremely high.

Given the high risk of exceeding contingency reserves, program stakeholders would be well-advised to significantly increase contingency reserves for new programs. There is a large standard deviation for schedule performance as depicted in the critical evaluation of program forecasts presented in Table 2. The wide coefficients of variation (CVs) shown in Table 2 for the actual durations, ranging from 31% to 56% across the selected commodity classes, emphasize

| Commodity Class | | Median (months) | Mean (months) | Count (n) | IQR (months) | Standard Deviation (months) | CV | Min (months) | Max (months) |
|---|---|---|---|---|---|---|---|---|---|
| **Ship** | **Planned** | 86.0 | 78.8 | 17 | 47.5 | 26.8 | 0.34 | 42 | 126 |
| | **Actual** | 89.0 | 94.7 | 17 | 61.0 | 33.0 | 0.35 | 49 | 147 |
| **Missile** | **Planned** | 75.0 | 69.4 | 23 | 25.0 | 17.1 | 0.25 | 38 | 92 |
| | **Actual** | 88.0 | 92.1 | 23 | 43.0 | 28.1 | 0.31 | 46 | 158 |
| **C5I** | **Planned** | 68.0 | 68.3 | 29 | 55.0 | 31.9 | 0.47 | 16 | 140 |
| | **Actual** | 86.0 | 84.5 | 29 | 39.5 | 35.0 | 0.41 | 16 | 159 |
| **Vehicle** | **Planned** | 57.0 | 62.3 | 7 | 22.0 | 23.3 | 0.37 | 30 | 106 |
| | **Actual** | 82.0 | 75.6 | 7 | 55.0 | 26.4 | 0.35 | 37 | 106 |
| **Fixed Wing** | **Planned** | 63.5 | 66.6 | 26 | 61.8 | 35.3 | 0.53 | 16 | 149 |
| | **Actual** | 72.0 | 77.1 | 26 | 72.0 | 43.3 | 0.56 | 24 | 177 |
| **Rotary** | **Planned** | 72.0 | 65.4 | 14 | 26.5 | 25.4 | 0.39 | 9 | 106 |
| | **Actual** | 81.0 | 82.4 | 14 | 65.5 | 40.1 | 0.49 | 11 | 151 |

*Table 2. Descriptive statistics for estimated and actual MDAP schedule durations from program initiation to IOC*

risks and uncertainties inherent in complicated MDAPs. The variation in risks and uncertainties is produced from having captured the full range of issues from block upgrades to existing missile systems to launching the next generation of multi-role combat aircraft.

***Exercising due diligence – the case of SM-6.*** Step three in the RCF method seeks to de-bias programs and improve the quality of front-end estimates through due diligence. In this step, benchmark data from completed MDAPs are applied to programs' original schedule forecast. It is important to note that information regarding the amount of contingency reserved for program uncertainty was not made available for purposes of this analysis. This is due to the length of time between program initiation and initial approved budget estimate.

The SM-6 Block I missile program, a system developed to provide critical Fleet capabilities against emerging threats, is a good example for illustrating how programs can apply benchmark data to their schedule estimates. What is significant about the SM-6 Block I missile is the fact that it is "an evolutionary development that marries the propulsion, airframe, and warhead of the SM-2 Block IV missile with the active radar seeker of the AIM-120C-7 Advanced Medium-Range Air-to-Air Missile (AMRAAM) to provide an Extended Range Anti-Air Warfare (ER-AAW) capability over sea and land" (Scott, 2019).

The SM-6 Block I program was initiated at MS B in June 2004 and reached IOC in November 2013. This represents a 39-month delay from the originally projected IOC date of September 2010. Our analysis, drawing from the dataset assembled for this study, noted that there are 18 past missile MDAPs that reached IOC prior to June 2004, which is the date of SM-6 Block I's initiation at MS B. This means that the SM-6 Block I program could have referenced these past missile MDAPs when forecasting their schedule. Though programs tend to draw analogies to select individual past programs or a handful of past programs when planning new programs, larger reference classes "lend themselves to statistical analysis, so that judgmental biases can be avoided in the estimation task" (Goodwin & Wright, 2010).

The 18 past missile programs in SM-6 Block I's example reference class are shown in Table 3. Further depicted is each program's estimated and actual duration from initiation to IOC and the difference between the two durations.

In this sample reference class, it is important to note that four in five missile MDAPs in the past 30 years have experienced a schedule overrun. Out of the total of 18 referenced programs, only two have finished on time since 1971, and just one was completed ahead of schedule. On average the 18 programs experienced a 37.2% delay in the delivery of their operational capabilities.

The SM-6 Block I program could have taken cues from the experience of the 18 referenced programs. Had the program referenced these past programs and acknowledged the reality that their program would likely run longer than initially projected, the SM-6 Block I program could have added a growth factor to their planned 76-month duration and extended their estimate accordingly. Applying a Lognormal distribution to the percent growth of past missile programs as recommended by Smart (2021) to represent project risk, the SM-6 Block I program could have compared their planned schedule against the confidence levels

| Program | Year of Initiating Milestone | Initiation to Est. IOC (months) | Initiation to Actual IOC (months) | Delta (months) | Percent Overrun |
|---|---|---|---|---|---|
| Program A | 1982 | 49 | 134 | 85 | 173.5% |
| Program B | 1976 | 58 | 115 | 57 | 98.3% |
| Program C | 1986 | 80 | 158 | 78 | 97.5% |
| Program D | 1972 | 65 | 107 | 42 | 64.6% |
| Program E | 1978 | 40 | 58 | 18 | 45.0% |
| Program F | 1976 | 88 | 127 | 39 | 44.3% |
| Program G | 1973 | 38 | 50 | 12 | 31.6% |
| Program H | 1977 | 65 | 84 | 19 | 29.2% |
| Program I | 1998 | 47 | 59 | 12 | 25.5% |
| Program J | 1996 | 69 | 84 | 15 | 21.7% |
| Program K | 1990 | 77 | 92 | 15 | 19.5% |
| Program L | 1971 | 83 | 95 | 12 | 14.5% |
| Program M | 1992 | 74 | 80 | 6 | 8.1% |
| Program N | 1989 | 79 | 85 | 6 | 7.6% |
| Program O | 1983 | 75 | 78 | 3 | 4.0% |
| Program P | 1994 | 92 | 92 | 0 | 0.0% |
| Program Q | 1979 | 88 | 88 | 0 | 0.0% |
| Program R | 1978 | 80 | 68 | -12 | -15.0% |

*Table 3. Schedule performance data from the 18 missile MDAPs available for reference at the time of SM-6 Block I's initiation*

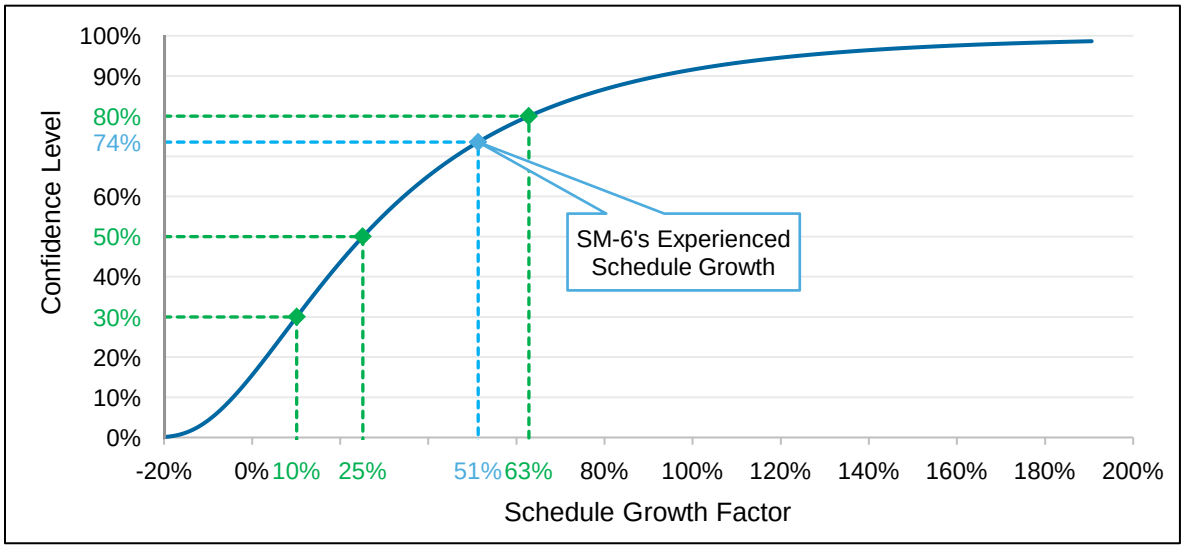


*Figure 3. Cumulative Lognormal distribution fitted to the schedule overruns of the 18 missile MDAPs in SM-6 Block I's potential reference set, with the 30%, 50%, and 80% confidence levels marked*

provided by the distribution shown in Figure 3 and made necessary adjustments.

Flyvbjerg (2018) highlights the 30%, 50%, and 80% confidence levels of the uncertainty curve as important for major programs. The 30% confidence level represents the "ambitious/optimistic estimate". Meanwhile, the 50% level represents the "most likely estimate" and the 80% level represents the "conservative estimate".

Scheduling to the optimistic 30% confidence level for schedule growth would produce a planned duration of about 84 months, which is 8 months longer than SM-6 Block I's planned duration of 76 months. Forecasting below the 30% confidence level signals optimism bias in SM-6 Block I's initial schedule estimate. Had the SM-6 Block I program planned to the 50% or 80% confidence level, they would have arrived at revised schedule estimates of about 95 months and 124 months, respectively. In fact, a forecast at the 80% confidence level would have provided the conservative schedule float to fully support the program's actual experienced duration of 115 months.

Comparing the SM-6 Block I program to other MDAPs in Figure 4, illustrates the SM-6 Block I's schedule growth from estimated to actual duration is in range with other MDAPs.

The U.S. Navy recently announced plans to fast-track the development and fielding of a new variant of the SM-6 missile, integrating a new government-developed rocket motor (Geurts, 2020). For the new SM-6 missile variant, RCF could be used to set more realistic schedule expectations than those set for its predecessors.

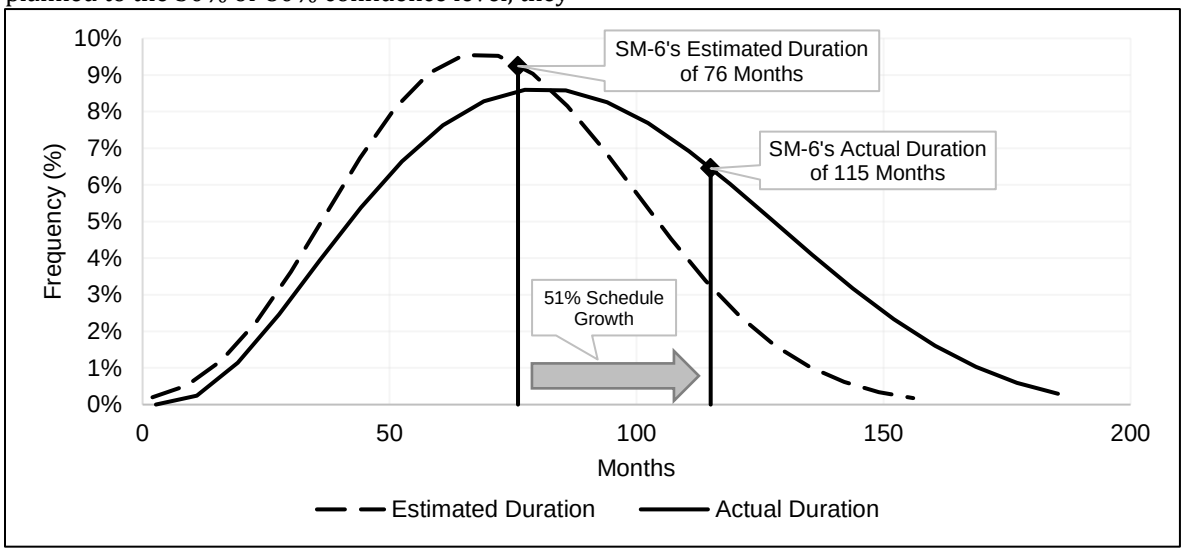


*Figure 4. Beta distributions fitted to the estimated and actual MDAP schedule durations, in months, from program initiation to IOC, with SM-6 Block I's estimated and actual schedule durations marked*

***Improving accountability.*** The fourth stage of RCF involves an examination of forecasting past performance and the credibility of previous forecasts in the continuing effort to de-bias schedule estimates. The step is needed since optimism bias, which has plagued previous programs, may well apply to most, if not all, engineering and independent forecasting organizations.

It should be noted that meaningful risk and uncertainty analyses generally result in sound, rigorous, and objective forecasts that capture the full range of outcomes for a major program. However, engineering and program teams often do not have the expertise to adequately determine the confidence interval in which the new schedule estimate will fall or to determine the schedule contingencies required to increase the probability of program success.

For example, McKinsey & Company published an assessment of U.S. DoD MDAPs that concluded that significant reforms are required to improve the accuracy of government schedule forecasts for the 96 programs representing nearly $2T in capital outlays. The report found that "equipment acquisition is notoriously difficult, too often characterized by cost growth and year-long delays" and that "the gestation time from program start to initial operating capability for major weapon systems has grown persistently" (Dowdy & Niehaus, 2010).

***Assessing risk.*** The fifth step of the RCF process aims to analyze cost and schedule risks beyond those captured at program initiation. This step necessitates an understanding of critical technologies and interdependencies across various program areas. Referring to the SM-6 Block I example, it is recognized that a development delay in the program's ability to deliver new extended range capability in direct response to Joint, Fleet, and U.S. Navy Urgent Operational Needs (UON) could negatively affect the flexibility Combatant Commanders need to increase surface force capabilities. In this example, forecasters can better identify and assess the program areas that are most sensitive to external factors by investigating a program's impact on other programs and vice versa. Similarly, forecasters can identify and assess the areas that could have the greatest effect on the program's cost per unit and integration schedule.

***Developing realistic estimates.*** Schedules have been underestimated by taking an 'inside view' to forecasting, as reflected in the review of the risks and uncertainties inherent in new MDAPs. The sixth step of RCF brings together the previous five steps to establish the program's most likely development with the goal of reducing the error in schedule estimates. While this more realistic estimate does not necessarily provide defense capabilities faster, it does allow dependent programs to forecast their own schedules more accurately. It also invites further scrutiny on previously unrecognized sources of schedule delays that could affect future programs. Given the strong levels of uncertainty that characterize weapon system development, each new program would benefit significantly from reviewing and integrating the distributional information from the 116 previously completed MDAPs. Programs would also benefit by taking an overall 'outside view' that helps to mitigate optimism bias through upward adjustments to schedule forecasts.

***Incorporating feedback.*** The goal of the seventh step is to solicit feedback from forecasting organizations after presenting the findings listed above. Here, all models created—which include the analysis of benchmark data of the 116 completed programs and any assumptions, inputs, calculations, outputs, graphs, and tables created—should be made available for comment. All forecast ground rules and assumptions should be clearly identified to eliminate any discrepancies in the interpretation of cost and/or schedule risks and uncertainty elements. In instances where assumptions were made related to sensitivity, a separate analysis could be conducted to resolve differences.

***Reducing optimism bias.*** The eighth and final step of formulating a true 'outside view' to de-bias the forecast is intended to allow the external stakeholders to expose the optimism bias inherent in MDAP schedule objectives by "using benchmark data to evaluate performance in the forecast in question" (Flyvbjerg, 2012). As highlighted throughout this Methods and Findings section, MDAP schedules have been significantly underestimated across all commodity classes—from missile systems to ships to the latest C5I programs. Many issues can be mitigated by incorporating knowledge-based best practices (GAO, 2021), yet it is clear that other MDAP risks will continue to be ignored or moderated for "fear of cancellation" (Emmons et al., 2018).

Stakeholders can leverage readily accessible benchmark schedule data to evaluate future program forecasts. This will guide weapon system programs to establish contingency reserves—or optimism bias uplifts—to meet the required confidence level and effectively de-bias the initial forecast to reflect the uncertain nature of new programs. With the abundance of information available today, robust analyses are required to determine whether the schedules established for

new platforms and associated warfighting capabilities align with the speed and scale envisioned for the future force.

## Conclusions and Future Research

Schedule performance is one of the most common reasons for MDAP failure. In fact, schedule performance contributes to large cost overruns and the delay of key capabilities. For the past several decades, we have seen a number of initial plans to field new warfighting capabilities not materializing due to systemic optimism bias. When the data on previously completed programs are reviewed and analyzed, the results consistently show that costs for new programs are underestimated, schedules do not fully reflect the innate risks and uncertainties of emerging programs, and overall warfighting capability benefits are often not able to meet the needs of a changing threat landscape.

In view of these challenges involving the forecasting of program costs, schedules and capabilities, the authors examined the schedule performance of 116 MDAPs in detail and in the process built a comprehensive information repository. This paper presented a research study that sought to empirically investigate the important attributes of RCF across a wide variety of MDAPs. Similarly, the study statistically analyzed the ability of RCF to develop more realistic and reliable front-end schedule estimates.

Constrained by the information contained within congressionally mandated SARs and other publicly available research, the dataset developed for this study was limited to program-level schedule performance, as data was not available at the individual WBS element level for MDAPs. Meanwhile, the limitations and insights of this study offer areas for future research.

First, research extending beyond program-level schedule performance is warranted to refine parsimonious models for shipboard systems and subsystems, given the need to keep pace with technological developments and changing strategic contexts. Second, there are some temporal gaps in the dataset due to difficulties obtaining access to estimated and actual schedule durations for every MDAP managed across the DoD. Examining additional MDAPs not contained within the dataset would help supplement and expand upon current findings. Third, the dataset in this study was limited to MDAPs, and research into international programs, particularly megaprojects managed by the Government of the United Kingdom where RCF has been employed, would deepen insights into this understudied topic and contribute to the generalizability of the findings.

This study supports the practical relevance of applying RCF to substantially de-risk schedule estimates and improve MDAP performance. Our findings confirm extant research on optimism bias and provide more in-depth data showing that the problem of underestimation is embedded in schedule estimates for MDAPs across all commodity classes with no signs of improvement over time. The findings indicate that RCF provides a more accurate method to mitigate optimism bias by integrating the successes and failures of previously completed programs.

Taking an ‘outside view’ of weapon system program forecasting enables forecasters to uncover the optimism bias inherent in MDAP schedule objectives. This view leads to more realistic schedule forecasts and offers a means to successfully manage the increasing complexity of large-scale programs so that time, effort, and costs are minimized, while relationships with prime contractors are streamlined. By following in Flyvbjerg’s footsteps and implementing the theories, practical tools, and due diligence of RCF based on an ‘outside view’, stakeholders can systematically de-bias estimates, conduct more meaningful risk and uncertainty analyses, and produce more accurate forecasts aligned with the design of the U.S. Department of Defense’s future force.

## References

Blickstein, I., Drezner, J. A., McInnis, B., McKernan, M. P., Nemfakos, C., Sollinger, J., & Wong, C. (2012). Methodologies for analyzing the root causes of Nunn-McCurdy breaches. Santa Monica, CA: RAND Corporation.

Buehler, R., Griffin, D., & Ross, M. (1994). Exploring the "planning fallacy": Why people underestimate their task completion times. *Journal of Personality and Social Psychology*, *67*, 366–381.

Dowdy, J., & Niehaus, J. (2010). Improving US equipment acquisition. *McKinsey on Government*, *5*, 14-21.

Emmons, D., Mazzuchi, T., Sarkani, S., & Larsen, C. (2018). Mitigating cognitive biases in risk identification: Practitioner checklist for the aerospace sector. *Defense Acquisition Research Journal*, *25*(1), 52-93.

Flyvbjerg, B. (2008). Curbing optimism bias and strategic misrepresentation in planning: Reference class forecasting in practice. *European Planning Studies*, *16*(1), 3-21.

Flyvbjerg, B. (2012). Quality control and due diligence in project management: Getting decisions right by taking the outside view. *International Journal of Project Management*, *31*(5), 760-774.

Flyvbjerg, B. (2018). How to de-risk major programmes. Presented at the Oxford Major Programme Management Conference, Technology and Major Programmes: Master Scaling, Oxford, England, UK.

Flyvbjerg, B. (2021). Top ten behavioral biases in project management: An overview. *Project Management Journal*, *52*(6), 531-546.

Francis, P. (2015). Defense acquisitions: Joint action needed by DOD and Congress to improve outcomes (GAO-16-187T). Testimony before the House Armed Services Committee, Washington, D.C.: U.S. GAO.

Galanti, S., & Vaubourg, A. (2017). Optimism bias in financial analysts' earnings forecasts: Do commissions sharing agreements reduce conflicts of interest? *Economic Modelling*, *67*, 325-337.

Geurts, J. (2020). Testimony before the Subcommittee on SeaPower and Projection Forces of the House Armed Services Committee. 4 March 2020.

Goodwin, P., & Wright, G. (2010). The limits of forecasting methods in anticipating rare events. *Technological Forecasting and Social Change*, *77*(3), 355-368.

Inhofe, J., & Reed, J. (2020). The Navy needs a course correction: Prototyping with purpose. *U.S. Naval Institute Proceedings*, *146*(6), 27-31.

Jardine, S., Moseley, M., Moul, J., & Trapp, D. (2019). Schedule estimating relationship (SER) development using missile and radar datasets. Paper presented at the 2019 International Cost Estimating & Analysis Association Professional Development & Training Workshop, Tampa, FL.

Jimenez, C. A., White, E. D., Brown, G. E., Ritschel, J. D., Lucas, B. M., & Seibel, M. J. (2016). Using pre-milestone B data to predict schedule duration for defense acquisition programs. *Journal of Cost Analysis and Parametrics*, *9*(2), 112-126.

Kahneman, D. (2011). *Thinking, fast and slow*. New York City, NY: Farrar, Straus and Giroux.

Kahneman, D., & Lovallo, D. (1993). Timid choices and bold forecasts: A cognitive perspective on risk taking. *Management Science*, *39*(1), 17-31.

Kahneman, D., & Tversky, A. (1979a). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263-292.

Kahneman, D., & Tversky, A. (1979b). Intuitive prediction: Biases and corrective procedures. *TIMS Studies in Management Science*, *12*, 313-327.

Kidd, J. B. (1970). The utilization of subjective probabilities in production planning. *Acta Psychologica*, *34*, 338-347.

LaGrone, S. (2021). CNO Gilday's new guidance sets clear goals to bring lethality to the surface force. USNI News. U.S. Naval Institute. Retrieved from https://news.usni.org/2021/01/11/cno-gildays-new-guidance-sets-clear-goals-to-bring-lethality-to-surface-force

Lee, D. E. (2005). Probability of project completion using stochastic project scheduling simulation. *Journal of Construction Engineering and Management*, *131*(3), 310-318.

Monaco, J. V., & White III, E. D. (2005). Investigating schedule slippage (ADA441770). Alexandria, VA: Defense Acquisition University.

Mortlock, R., & Dew, N. (2021). Behavioral biases within defense acquisition. Proceedings of the Eighteenth Annual Acquisition Research Symposium (SYM-AM-21-049). Monterey, CA: Naval Postgraduate School.

Park, J. (2021). Schedule delays of major projects: What should we do about it? *Transport Reviews*, *41*(6), 814-832.

Patel, P., & Tsionas, M. (2021). Macroeconomic uncertainty and risk: Collective optimism of small-business owners. *Entrepreneurship Theory and Practice*.

Schinasi, K. V. (2008). Defense acquisitions: Better weapon program outcomes require discipline, accountability, and fundamental changes in the acquisition environment (GAO-08-782T). Testimony before the Committee on Armed Services, U.S. Senate, Washington, D.C.: U.S. GAO.

Scott, R. (2019). US Navy reveals plan for extended range SM-6 missile. Jane's Missiles & Rockets. Retrieved from https://www.janes.com/article/87434/us-navy-reveals-plan-for-extended-range-sm-6-missile

Shefrin, H., & Statman, M. (2003). The contributions of Daniel Kahneman and Amos Tversky. *The Journal of Behavioral Finance*, *4*(2), 54-58.

Smart, C. B. (2021). *Solving for project risk management: Understanding the critical role of uncertainty in project management*. New York City, NY: McGraw-Hill.

Tyson, K. W., Harmon, B. R., & Utech, D. M. (1994). Understanding cost and schedule growth in acquisition programs (P-2967). Alexandria, VA: Institute for Defense Analyses.

U.S. Department of Defense. (2018). Comprehensive selected acquisition reports (SARs) for the December 31, 2017 reporting requirement as updated by the President's FY 2019 budget. Washington, D.C.: U.S. DoD.

U.S. Department of Defense. (2020). Performance of the defense acquisition system: 2020 annual report. Washington, DC: Office of the Under Secretary of Defense for Acquisition and Sustainment.

U.S. Government Accountability Office. (2018). Weapon systems annual assessment: Knowledge gaps pose risks to sustaining recent positive trends (GAO-18-360SP). Washington, D.C.: U.S. GAO.

U.S. Government Accountability Office. (2020). Defense acquisitions annual assessment: Drive to deliver capabilities faster increases importance of program knowledge and consistent data for oversight (GAO-20-439). Washington, D.C.: U.S. GAO.

U.S. Government Accountability Office. (2021). Weapon systems annual assessment: Updated program oversight approach needed (GAO-21-222). Washington, D.C.: U.S. GAO.

Valerdi, R. (2009). Optimizing optimism in systems engineers. Proceedings from the INCOSE Conference on Decision Analysis and Its Applications to Systems Engineering. Newport News, VA: INCOSE.

Wise, T., Zbozinek, T. D., Michelini, G., Hagan, C. C., & Mobbs, D. (2020). Changes in risk perception and self-reported protective behaviour during the first week of the COVID-19 pandemic in the United States. *Royal Society Open Science*, *7*(9), 200742.

## Authors' Biography

***Jeffrey M. Voth*** *is the President of Herren Associates, an engineering and management consulting firm headquartered in Washington, D.C. He studied at St. Catherine's College, University of Oxford, obtaining his MSc degree with distinction and specializing in the economic impact and risk of global megaprojects. Prior to this, he received an MBA from Georgetown University and was awarded a bachelor's degree from the University of Massachusetts, Amherst.*

***Maxwell C. Moseley*** *is a CCEA® at Herren Associates in Washington, D.C., where he performs cost and statistical analyses for several defense agencies. He graduated summa cum laude from Mississippi State University with his B.S. in Industrial Engineering and B.A. in Communication. Prior to joining Herren, Mr. Moseley engaged in operations research with homeland security applications.*

***Ann E. Hawpe*** *is a co-founding member of Herren's Program Analysis and Cost Engineering practice. She has established and managed the firm's business with the US Government for over 19 years in Washington, D.C. Ms. Hawpe continues to identify trends across the cost estimating industry and deliver advanced training throughout the cost community. She holds a B.S. in Industrial and Systems Engineering from Virginia Tech and a M.S. in Systems Engineering from The George Washington University.*