SYM-AM-18-050

PROCEEDINGS
OF THE
FIFTEENTH ANNUAL ACQUISITION RESEARCH SYMPOSIUM

WEDNESDAY SESSIONS
VOLUME I

Acquisition Research:
Creating Synergy for Informed Change

May 9–10, 2018

Published April 30, 2018

Prepared for the Naval Postgraduate School, Monterey, CA 93943.

ACQUISITION RESEARCH PROGRAM
GRADUATE SCHOOL OF BUSINESS & PUBLIC POLICY
NAVAL POSTGRADUATE SCHOOL

Quantifying Annual Affordability Risk of Major Defense Programs

David M. Tate—joined the research staff of the Institute for Defense Analyses' (IDA) Cost Analysis and Research Division in 2000. Prior to that, he was an Assistant Professor of Industrial Engineering at the University of Pittsburgh and the Senior Operations Research Analyst (Telecom) for Decision-Science Applications, Inc. At IDA, he has worked on a wide variety of resource analysis and quantitative modeling projects related to national security. These include an independent cost estimate of Future Combat Systems development costs, investigation of apparent inequities in Veterans' Disability Benefit adjudications, and modeling and optimization of resource-constrained acquisition portfolios. Tate holds bachelor's degrees in philosophy and mathematical sciences from the Johns Hopkins University, and MS and PhD degrees in Operations Research from Cornell University. [dtate@ida.org]

Thomas J. Coonce—is an Adjunct Research Staff Member with the Institute for Defense Analyses (IDA). He performs cost and schedule research for various clients, including the Office of the Secretary of Defense (OSD), the Department of Homeland Security (DHS), and the Congressional Section 809 Panel. Prior to joining IDA, Coonce led NASA's cost analysis activities and served as a cost/schedule analyst for the OSD's CAIG, The MITRE Corporation, and the Government Accountability Office (GAO). [tcoonce@ida.org]

Michael R. Guggisberg—is a Research Staff Member in the Strategy, Forces, and Resources Division at the Institute for Defense Analyses. He earned his PhD in Economics at the University of California, Irvine, in 2017. Prior to that, he was a squad leader in the Marine Corps Reserve, with tours in Iraq and Afghanistan. [mguggisb@ida.org]

Abstract

To a first approximation, acquisition programs never spend what they originally said they would spend when they began. In fact, the error bars around an initial cost estimate are much larger than is generally understood, once program cancellations, restructurings, truncations, and block upgrades have been accounted for. Worse yet, all of this uncertainty arises in a context where programs must fit within annual budgets—it is not enough to only spend as much as you said you would; you must also spend it *when* you said you would, or problems ensue.

We have developed a methodology that uses historical program outcomes to characterize the year-by-year budget risk associated with a major acquisition program. This methodology can be applied to both development costs and procurement costs and can be extended to understand the aggregate affordability risk of portfolios of programs. The method allows Resource Managers to estimate annual budget risk levels, required contingency amounts to achieve a target probability of staying within a given budget, and many other relevant risk metrics for programs. It also allows policy makers to predict the impact on program affordability of proposed changes in how contingency funds are managed.

Planning Is About Not Being Surprised

A Hypothetical New Program

Suppose you are the Resource Manager for a portfolio of acquisition programs. A new helicopter program in your portfolio—call it "H-99"—has just received milestone approval. The H-99 Acquisition Program Baseline (APB) for development looks like Figure 1.

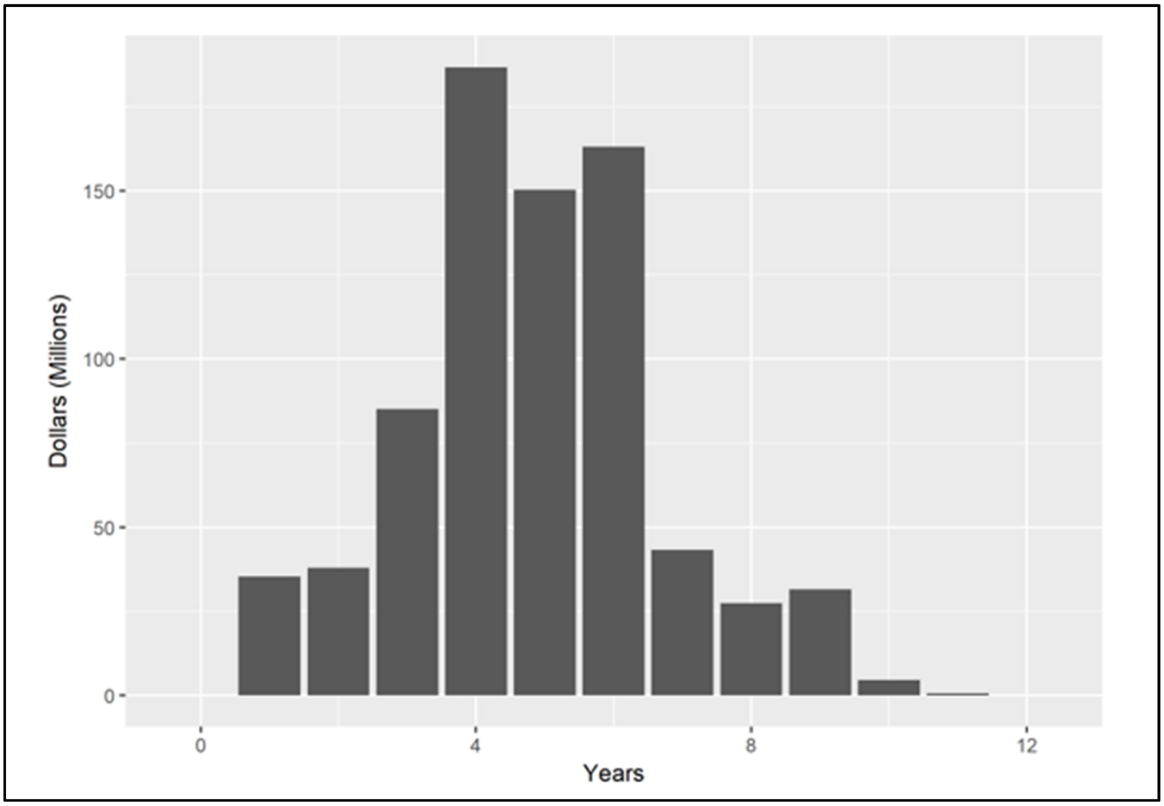

Figure 1. Proposed H-99 Development Costs

This is the official Service position about how many research, development, test, and evaluation (RDT&E) dollars this program will receive in each of the next dozen years—but it is not what will actually happen. What should you expect? Or, more precisely, what range of outcomes should you be prepared for, and how likely should you think those outcomes are? You don't want to be surprised.

Some Past Examples

Figure 2 shows the planned and actual development costs of the MH-60S Seahawk (a.k.a. "Knighthawk") helicopter program, as reported in the program's first Selected Acquisition Report (SAR) in 1998 and its final SAR in 2015. The original plan was to spend $71 million (constant 1998 dollars) over four years of RDT&E. The program actually spent more than $670 million over 19 years. The original requirement was for a ship-based cargo helicopter that could also support search-and-rescue and torpedo recovery. After the first 50 helicopters were built, the program evolved through several "block upgrades" to add airborne mine detection and countermeasures and shifted from an unarmed platform to an armed combat search and rescue (CSAR) and maritime interdiction platform—all within the same program of record.

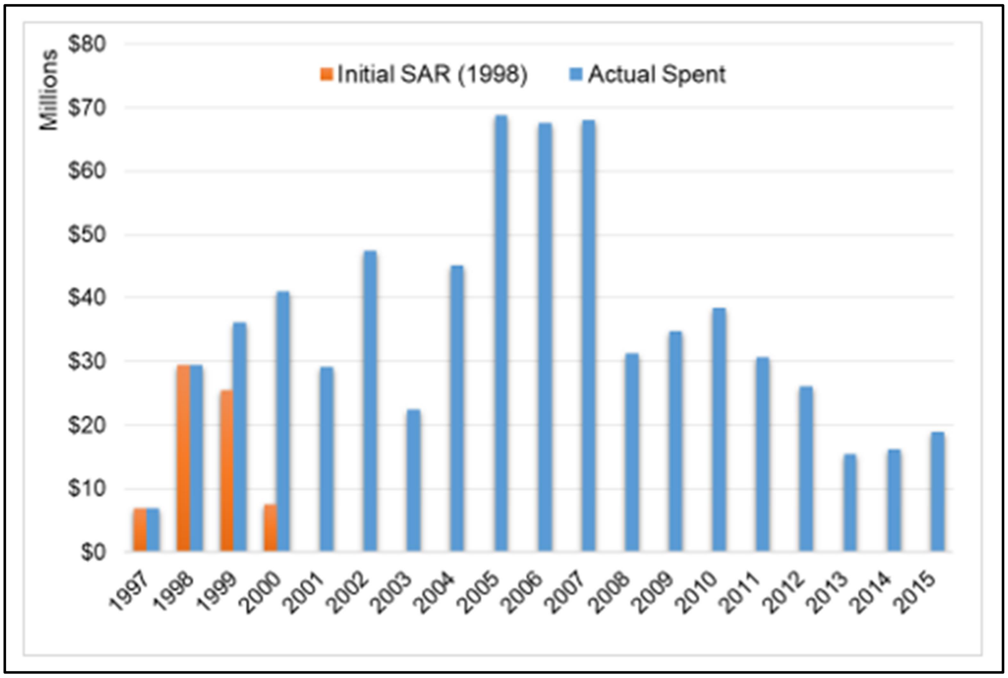

Figure 2. MH-60S Development Costs

Conversely, Figure 3 shows the planned versus actual procurement costs for the Armed Reconnaissance Helicopter (ARH) program. Here, the error is in the opposite direction; because the program was cancelled before procurement spending even began, none of the original estimated $3 billion (constant 2005 dollars) was ever spent.

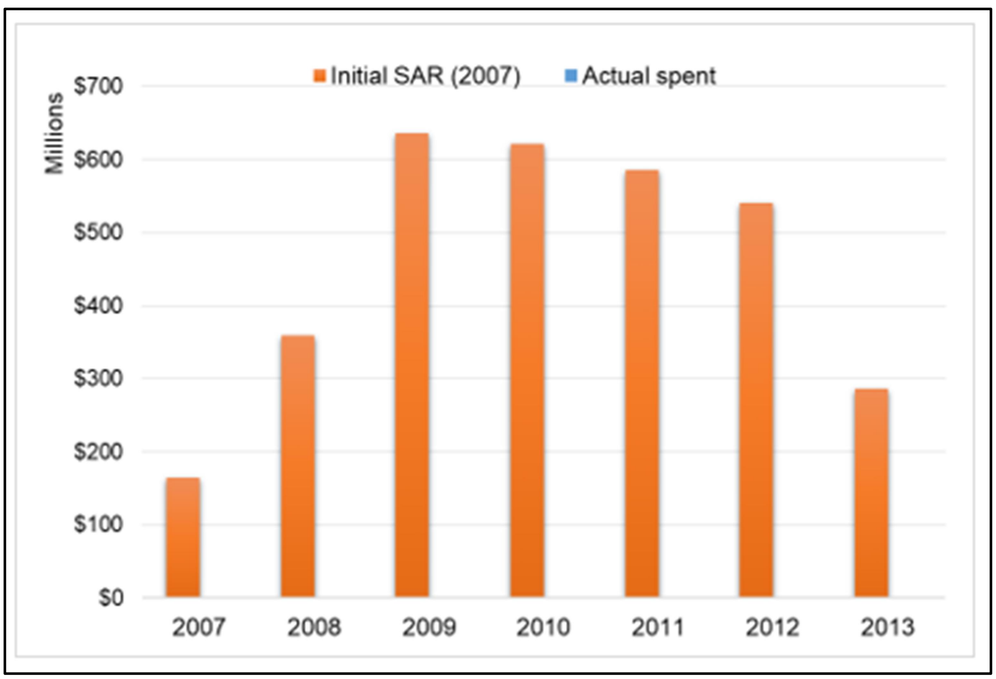

Figure 3. ARH Planned Procurement Costs (Zero Actually Spent)

Clearly, the dollars actually spent on development or procurement of a given program can vary wildly—either up or down—from what was originally planned. In order to manage acquisition portfolios sensibly, Resource Managers need to have some idea of just how much actual future resource usage might differ from its original estimate. We compiled original plans and actual outcomes from 115 historical major defense acquisition programs (MDAPs) in order to attempt to answer this question.

The Resource Manager's Challenge

We saw in the previous examples that what actually happens can be vastly different from what was predicted. Given the range of possible program outcomes, what should Resource Managers be prepared for? Are there tools that could help Resource Managers quantify the year-by-year affordability risk they face, based on actual historical outcomes and observable program characteristics, so that they can allocate contingency funds wisely?

Yes, there are tools—or at least there could be. That is the purpose of this research.

Affordability means having a high probability of being able to buy the thing you want with the funds you have designated for that purpose. We often think of this as a question of total cost or unit cost, but in practice it is a year-by-year question of whether there is enough money in this year's budget to cover this year's costs. The Resource Manager's goal is to balance the need for every program to have enough funds to execute efficiently this year against the demands of other programs and the overall budget. Planning sensibly for future years' demands is a major part of this. Programs that get stretched for lack of immediate funding in the budget year not only take longer to finish, but they also grow in total cost (due to inefficient use of labor and a higher proportion of fixed costs). On the other hand, providing enough contingency funding to make every program highly likely to stay within budget every year would be tremendously wasteful—especially since overfunded programs will tend to find something to spend that money on, rather than giving it back. Major differences between planned and actual demands for funding lead to budget instability, inefficient acquisition, and cancelled programs.

Existing Literature on Cost Growth Doesn't Help

What Question Are We Asking?

Most of the literature on cost growth is focused on trying to identify *causes* of cost growth in order to avoid them. That is a noble goal, but as a Resource Manager, you are probably aware that people have been trying to do that for decades, with limited success.

Similarly, most of the literature on cost growth looks at *unit* cost growth—either actual unit cost growth or unit cost growth adjusted to control for changes in the quantity procured (Asher & Maggelet, 1984; McNicol, 2017; McNicol et al., 2013). That makes sense if you're trying to understand the dynamics of why things cost more than expected. But as a Resource Manager, you don't care nearly as much about unit costs as you do about how many actual dollars will be needed.

In addition, nearly all cost growth studies describe cost growth (be it unit cost or total cost) in terms of a single program cost growth factor—the ratio of *final* total (or unit) cost to the originally estimated total (or unit) cost (Arena et al., 2006). To a Resource Manager, eventual total cost is not nearly as important as <u>annual</u> cost—that is, how many dollars will the program consume *this year*, and next year, and for each year in the Future Years Defense Plan (FYDP)?

Finally, nearly all past efforts are aimed at *predicting* costs, with those predictions expressed as an expected value or "50%" estimate (Anderson & Cherwonik, 1997; Bitten &

Hunt, 2017). This is in part because the literature has been driven by the needs of cost estimators, who are expected to produce point estimates that can be used for budgeting purposes. For cost estimators, the quality of the prediction is given by its *accuracy* (how much bias) and its *precision* (in terms of narrow error bars). Resource Managers prefer accurate (unbiased) forecasts, too, but do not really care about the mean value *per se.* Instead, they care greatly about being able to assess whether a program is likely to exceed yearly budgets and by how much, so that adjustments can be made before the program is forced to shed capabilities or be stretched out in order to “fit” within the budget profile. This is not a question of expected total cost or expected cost growth—it is a question of how the annual needs can vary and how to plan for those possibilities.

The authors are not aware of any past work that has addressed this need for annual resource requirement risk analyses. The Congressional Budget Office (CBO) uses commodity-specific cost growth factors when predicting future budget implications of current and predicted programs, but it does this in terms of average total program cost growth factors applied uniformly across all years of the planned program (CBO, 2017). As we shall see later, this is not an accurate estimate of *when* the cost growth in a given program would manifest itself. It is also just an average—it gives no information on the range of possible outcomes, and how relatively likely they are. This is the gap that our research aims to fill.

Desirable Outputs of a Model

Given a planned program (or set of programs—we’ll get to that later) and a budget, Resource Managers would very much like to answer questions such as

- What is the distribution of funding the program will receive in year N = 1, 2, …?
- What is the probability that the program will receive more funding in year N than is currently budgeted, for N = 1, 2, …?
- How many total contingency dollars would be enough to achieve a given percent certainty that the current budget plus the contingency is enough to fund the program over the FYDP?
- What is the probability that the program will use at least $X less than planned over the FYDP, for various values of X?
- …etc.

The goal of our research is to develop empirical models based on historical program attributes, environments, and outcomes that will allow us to answer questions like these.

What Tools Do We Need?

In this section, we describe the tools that are needed to attempt to answer the kinds of questions that Resource Managers care about. These tools are

1. a way to describe funding profiles mathematically;
2. a list of program attributes and environmental factors that help predict program outcomes;
3. a statistical model to estimate the probability distribution of final funding profile shapes given the initial funding profile, environmental factors, and other program attributes;
4. a mathematical characterization of how well the shape tends to fit actual data; and
5. historical data on program initial plans and final outcomes.

A detailed discussion of each of these tools follows.

Step 1: Modeling Funding Profiles

It is an essential (and bothersome) fact that the various years of an acquisition program are not independent. If the actual RDT&E obligation authority in year 5 of development turns out to be higher than originally planned, it is very likely that year 6 and year 7 (etc.) will also be higher than originally planned. In fact, we are not interested only in the distribution of outcomes for year N; we are interested in the *joint* distribution of outcomes in all years—including years that were not part of the original baseline plan at all. That means we can't just build separate risk models for year 1, year 2, and so forth, and then use that collection of models to understand the risk over the entire planning horizon. We have to account for the ways cost *profiles* change over time.

Instead of treating the year-by-year outcomes as having some complicated joint distribution, we will instead use the techniques of *Functional Regression* to treat the individual year-by-year outcomes as having been generated by some (noisy) underlying functional form and then think about probability distributions over the parameters of those generating functions.

Development Costs

For the development portion of program funding, there is already a history of fitting functional models to the yearly funding requirements. In particular, Weibull distributions (or Rayleigh distributions, which are a special case of the Weibull) have often been used to describe both the shape a development program ought to have and the observed actual funding profiles of historical programs. Brown, White, Ritschel, and Seibel (2015) provide a good summary of past approaches.

Weibull distributions can be parametrized in several ways. We base our model of development cost profiles on this version, which has two parameters, λ and α:

$$W(t|\alpha, \lambda) = \frac{\alpha}{\lambda}\left(\frac{t}{\lambda}\right)^{\alpha-1} \exp\left(-\left(\frac{t}{\lambda}\right)^{\alpha}\right) 1(t \geq 0) \tag{1}$$

In this parametrization, λ is a time-scaling parameter that determines how much the profile changes from year to year, while α is a shape parameter. The term $1(t \geq 0)$ is an indicator function that equals 1 if t is greater than or equal to 0 and 0 else. Figure 4 shows the flexibility of the Weibull for various (λ, α) pairs. The black $(\lambda = 1, \alpha = 2)$ and red $(\lambda = 2, \alpha = 2)$ profiles show how the scaling parameter λ affects how quickly the spending profile changes. The red profile evolves more slowly than the black one because it has a larger λ value. The black $(\lambda = 1, \alpha = 2)$, blue $(\lambda = 1, \alpha = 1.5)$, and teal $(\lambda = 1, \alpha = 4)$ profiles show how the shape parameter α affects the peakedness. The teal profile is the most peaked, because it has the largest α value. Values of α less than 1 (e.g., the green curve) are not appropriate for modeling spending profiles, since $\alpha < 1$ implies that peak spending occurs in year 0.

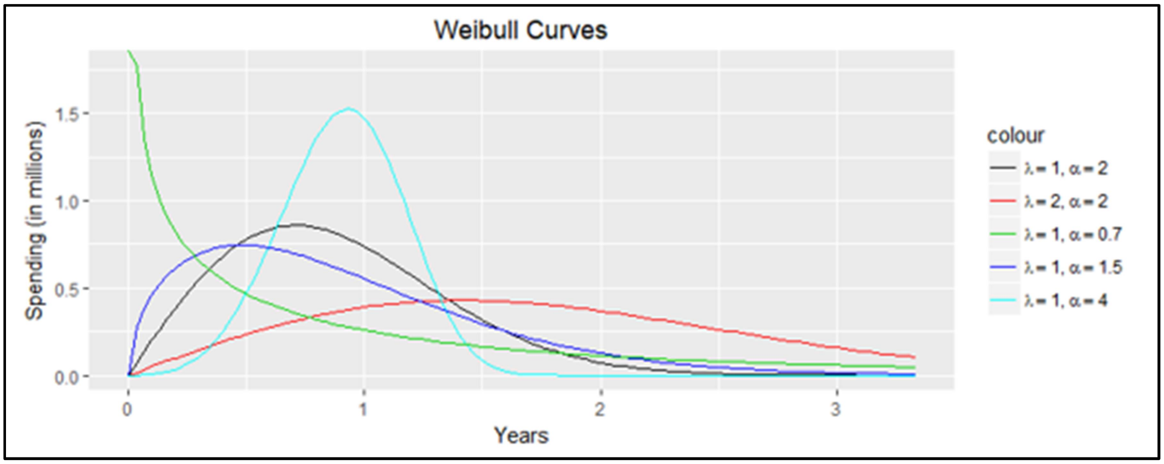

Figure 4. Weibull Distribution Shapes for Various Parameter Values

There are two major discrepancies between the Weibull distribution and the data we observe. First, the data we observe is annual, and therefore discrete, while the Weibull distribution is continuous. Second, some programs are cancelled during development, while the Weibull distribution assumes they are completed. To address these issues, we discretize and truncate the Weibull distribution using two additional parameters:

- the total development cost, denoted *C*
- the number of years of nonzero spending, denoted *T*

Let *C(t)* denote the costs in year t. The resulting functional form fit to the incurred costs in year t is

$$C(t) = K \cdot W(t|\alpha, \lambda) + \epsilon(t)\,, t = 1, \dots, T \tag{2}$$

where $\epsilon(t)$ is the independent random error in year *t* and the constant *K* is chosen such that $\sum_{t=1}^{T} C(t) = C$. This scaled, discretized truncation of the Weibull profile lets us accurately fit the final outcomes of programs cancelled in mid-development, as well as typical development profiles that taper off more slowly.

Procurement Costs

For procurement, there is no similarly traditional functional form to use as the basis of a functional regression. One possibility is to use a simple two-segment piecewise linear shape, consisting of an initial ramp-up, followed by a linear trend line. This distribution could be parametrized in various ways using four parameters. The choice of functional form here is purely descriptive—unlike the theory behind the application of Weibull distributions to development profiles, there is no underlying reason why procurement profiles should be roughly linear following their initial low-rate initial production (LRIP) ramp-up. For purposes of risk characterization, it is enough that historical programs do (approximately) follow this pattern.

For the remainder of this paper, we will ignore procurement costs and focus on modeling development cost profiles for simplicity of exposition.

Step 2: Identifying Potential Predictor Variables

Given choices for functional forms, the next challenge is to somehow characterize how the distribution of possible actual outcome profiles could be derived for a given initial plan. It seems obvious that different programs involve different levels of cost risk. There is a long literature attempting to identify specific factors that are correlated with program cost and schedule growth. Some factors that have been found by past researchers to be correlated with (unit) cost growth and/or total program cost growth risk include

- Commodity type (e.g., helicopter, satellite, MAIS, missile, or submarine) (Arena et al., 2006; Drezner et al., 1993; Tyson, Harmon, & Utech, 1994)
- Acquiring Service (Army, Navy, Air Force, Joint, Department of Energy [DoE]) (Drezner & Smith, 1990; Jessup & Williams, 2015; Light et al., 2017; McNicol, 2004)
- New design vs. modification of existing design (Arena et al., 2006; Coonce et al., 2010; Drezner et al., 1993; Jimenez et al., 2016; Marshall & Meckling, 1959)
- New build vs. remanufacture of existing units (Tyson et al., 1989)
- Budget climate at Milestone B (Asher & Maggelet, 1984; McNicol, 2017)
- Number of years of spending prior to Milestone B (Jimenez et al., 2016; Light et al., 2017)
- Schedule optimism (Arena et al., 2006; Asher & Maggelet, 1984; Glennan et al., 1993; Tate, 2016)
- Technology maturity of the program (Adoko, Mazzuchi, & Sarkani, 2015; GAO, 2006)
- Investment size (Bliss, 1991; Creedy, Skitmore, & Wong, 2010)

Because we are only trying to understand and characterize risk, on the assumption that the past is a reasonable guide to the future, we do not distinguish here between risks arising from discretionary choices, environmental factors, or intrinsic program features.

Step 3: Describing Changes in Cost and Schedule as Changes in Profile Functions

We saw previously that we can model development costs or production costs as being generated by an underlying functional form, with variation about that smooth curve treated as independent random noise. Mathematically,

$$C(t) = f(t|\theta) + \varepsilon(t) \qquad (3)$$

where *C(t)* is the cost in year *t, θ* is the vector of fitted parameter values for the family of curves being used, $f(\cdot)$ is the functional form we are assuming for the generating function, and $\varepsilon(t)$ is a random error whose distribution may depend on the year t. Let θ_0 denote the parameters that best fit the program's original profile and θ_1 denote the parameters that best fit the program's final profile. What we need to estimate is the conditional (joint) distribution of θ_1 given the appropriate program and environmental attributes and the fact that the program's original estimate was best fit by the curve $f(t|\theta_0)$.

There are several possible approaches to this and many choices of how to parametrize the family of curves being fit, but the general method will be the same in all cases. We estimate the distribution of θ_1 as a function of the best fit parameters θ_0 and the historical program characteristics *X*:

$$\log(\theta_1) = (X, \log(\theta_0))\beta + \eta, \qquad (4)$$

where X includes factors such as initial estimated cost, Service, budget climate, and so forth. The matrix element X_{jk} gives the value of predictor k for historical program j. The matrix θ_0 has one column for each parameter and one row for each historical program. $(X, \log(\theta_0))$ denotes the block matrix obtained by appending the componentwise natural logarithm of θ_0 as additional columns of X, one column per parameter.

This linear regression model implies a functional fit and distribution over the annual cost profile function $C(t)$. Rather than attempting to predict eventual actual cost as a function of initial estimated cost and other predictors, we instead attempt to predict the distribution of the parameters of a function that *generates* eventual cost, given program-specific attributes and the parameters that generate the initial estimate. Note that this is a multiple output regression—we are simultaneously estimating all of the best-fit parameters θ_1 and the covariance matrix that describes how those parameters are correlated.

We use a Bayesian estimation framework, starting with a weakly informative prior distribution $F_{prior}(\theta_1)$ and using Markov Chain Monte Carlo estimation to derive a posterior distribution $F_{posterior}(\theta_1)$, including the covariance matrix (Chib & Greenberg, 1995). We do this separately for development costs and procurement costs, using different families of profile-generating functions and treating their changes in shape and size as independent. Treating development and procurement jointly is a potential area for future research.

Step 4: Accounting for Noisy Curve Fits

The posterior distribution on θ_1 accounts for the uncertainty in the generating function parameters for the eventual profile of the program, but it does not capture the variability corresponding to the original error term $\varepsilon(t)$ when we fit truncated Weibull distributions to profiles. In order to capture all of the uncertainty in actual yearly costs, we need to also add in yearly random error terms $E(t)$. We derive the distribution of $E(t)$ from the observed $\varepsilon(t)$ in the best fits for actual costs of historical programs. In other words, we look at how much our curve fits to actual outcome profiles that tended to be off in each year, and we add corresponding random yearly error terms to any final profile generated from $F_{posterior}(\theta_1)$ in order to capture that additional source of uncertainty.

Step 5: Regression Data

The data for the regression are the initial estimate and final actual cost profiles for 155 completed historical MDAPs. The earliest program in the data set passed Milestone B in 1982. The data are taken from Selected Acquisition Reports (SARs), together with compiled attributes and environmental factors (as described previously in Step 2) for each program. In this paper, we will focus solely on characterization of development (RDT&E) cost risk. The methodology for procurement cost risk is similar, differing only in which predictor values are used and in the functional form of the generating function.

The specific predictor variables used in this paper are

- $\log(\alpha_0)$—natural logarithm of the shape parameter of the original estimate Weibull fit
- $\log(\lambda_0)$—natural log of the scale parameter of the original estimate Weibull fit
- $\log(C_0)$—natural log of the original total planned spending
- $\log(T_0)$—natural log of the original planned number non-zero spending years
- The Service overseeing the program (Navy, DoD, Air Force, Army, DoE)

- A commodity type (Air; Command, Control, Communications, Computers, Intelligence, Surveillance, and Reconnaissance (C4ISR); Ground; Ordnance; Sea; Space; other)[1]
- A measure of relative Service budget tightness compared to two years ago
- A measure of relative Service budget tightness over the last 10 years
- A measure of budget optimism—planned spending divided by the mean historical actual spending for this commodity type
- A measure of schedule optimism—planned duration divided by the mean historical actual duration for this commodity type
- Whether the program is based on a modification of a preexisting design (binary)

The measures of relative budget tightness were based on the year the program passed Milestone II/B.

Regression Methodology

Let $i = 1,2,\ldots,I$ index over the historical programs in our data set; let the subscript $l = 0$ denote an original profile estimate and $l = 1$ denote an actual realized profile. We compiled original estimates and actual outcomes for $I = 115$ historical programs. For each historical program i, we fit scaled, discretized Weibull distributions[2] to the original estimated and actual development cost profiles $C_{i0}(t)$ and $C_{i1}(t)$:

$$C_{il}(t) = K_{il}\, W(t|\alpha_{il}, \lambda_{il}) + \epsilon_{il}(t), t = 1, \ldots, T \quad (5)$$

As before, the constants K_{il} are chosen so that $\sum_{t=1}^{T} C_{il}(t) = C_{il}$, the total cost of the original/final profile for program i.

[1] More precise commodity categories (e.g., distinguishing helicopters from fixed-wing aircraft) might be useful, given enough data. We found that increasing the sample size in each category led to better results than increasing the precision of the categories.
[2] For historical profiles, we set T = estimated years to reach Milestone C and C = total development cost through year T. We added fictitious years of zero spending in year 0 and year T + 1, then fit a Weibull distribution to the scaled yearly costs. For actual profiles, we set C = actual total cost and T = actual years of spending. We then fit a Weibull to the scaled annual costs, holding the sum of costs through year T equal to C. We used the Levenberg-Marquardt algorithm to perform the nonlinear least-squares optimizations to find the best fits.

Let $\theta_{il} = (C_{il}, T_{il}, \alpha_{il}, \lambda_{il})$ be the parameters of those best-fit curves. Then θ_{i0} are the best fit parameters to the initial profiles and θ_{i1} are the best fit parameters to the actual outcomes. We model the distribution of θ_{i1} as a function of θ_{i0} and a set of predictor variables X_i simultaneously over all programs, where X includes the program-specific and environmental factors previously listed. Parametric linear models are simultaneously fit to obtain a predictive model for the final profile parameters θ_1. The parameters that uniquely identify a profile are $\theta_{i1} = (C_{i1}, T_{i1}, \alpha_{i1}, \lambda_{i1})$ where $C_{il} = \sum_{t=1}^{T_{it}} C_{il}(t)$ and the other parameters are as previously defined. The models are

$$\log(C_{il}) = (X; \log(\theta_0))\beta_C + \eta_C, \tag{6}$$

$$\log(T_{il}) = (X; \log(\theta_0))\beta_T + \eta_T, \tag{7}$$

$$\log(\alpha_{il}) = (X; \log(\theta_0))\beta_\alpha + \eta_\alpha, \tag{8}$$

$$\log(\lambda_{il}) = (X; \log(\theta_0))\beta_\lambda + \eta_\lambda\ , \tag{9}$$

where the error terms are assumed to be jointly normally distributed. The covariates X include information about previously finished programs that had initial planned spending profiles and actual final profiles. Using these historical data, the model is fit to predict final actual profiles using only information available from a program's Milestone B date. The parameters $\beta = (\beta_C, \beta_T, \beta_\alpha, \beta_\lambda)$ are jointly estimated using a Bayesian Seemingly Unrelated Regressions model with prior distributions on the parameters β and $Var[\log(\theta_{i1})\,|X] \equiv \Sigma$.

The prior for β has a multivariate normal distribution, calibrated such that prior belief is that there is no change in the profile from initial estimate to final actual profile and no other traits of the initial profile are predictive of the final actual profile. This prior belief is fairly strong in order to induce regularization. This prior choice balances the bias-versus-variance tradeoff to produce better predictions.

The prior for Σ has an inverse Wishart distribution, chosen to account for the errors in the initial profile fits to the Weibull curve. In addition, it includes central limit theorem estimates of the final length and size of actual programs. It also includes correlation between the final length, scale, and size parameters.

The joint posterior distribution of β and Σ incorporates the prior beliefs and the historical data to arrive at an updated posterior belief. The Bayesian machinery is especially useful for our purposes because it allows us to obtain random draws from the posterior distribution of β and Σ, which in turn allows us to generate random draws of a final profile distribution $\hat{\theta}_1$ for any program with known initial profile characterized by θ_0 and covariates $X = x$. This lets us estimate the complete (posterior) distribution of final profiles, rather than just a point estimate and variance measure.

Applying the Model to a Specific Program

Experimentation with the predictors listed previously and the 115 programs in our data set indicated that only the coefficients for $\log(C_{i0}), \log(T_{i0}), \log(\alpha_{i0}), \log(\lambda_{i0})$ and Service = Army were significant predictors of final RDT&E development profile shape and size.[3] However, all predictors were retained in the model and strongly regularized to improve predictive performance.

Returning to the hypothetical H-99 program we introduced early on, assume that the program passed Milestone B in 2014. The resulting *X* covariate values for estimating the distribution of eventual actual development expenditure profiles are

- Commodity = Aircraft
- $\alpha_0 = 3.3$
- $\lambda_0 = 5.3$
- C_0 = = $766.2 Million
- T_0 = 12
- Two-year tightness = -0.073
- 10-year tightness = 1.0
- Service = Army
- Commodity Size Optimism = 0.18
- Commodity Length Optimism = 1.11

Using these inputs, we compute the posterior distribution $F_{posterio}\ (\theta_1)$ of the parameter vector θ_1 that describes the best Weibull fit to the eventual actual development cost profile. In the Bayesian paradigm, this is not a point estimate for a best-fit curve, but instead an updated joint probability distribution that summarizes both our new beliefs about what those parameters are likely to be and our uncertainty about them. We can think of this posterior distribution as a probability distribution over funding profiles.

We also need to account for the fact that the best-fit curve isn't a perfect fit. We also estimate the year-by-year error distributions $E(t)$ using our historical data on how well Weibull distributions fit actual development profiles.

We now have all of the machinery we need to characterize year-by-year cost risk. We use Monte Carlo simulation to sample repeatedly from the $F_{posterior}(\theta_1)$ distribution, generating a large number of possible funding profiles. For each profile generated, we then perturb the annual values using offsets generated from the $E(t)$ distributions. We repeat this process tens of thousands of times, collecting year-by-year statistics on how frequently the required dollars in that year exceeded any given threshold. The result is a set of annual "dollars required" distributions that can be compared against both the original estimate and a hypothetical budget.

[3] Using type 2 sum of squared errors.

Monte Carlo Risk Analysis

Approach

Suppose that we have budgeted a program at some level, possibly different from its predicted cost profile. Let $B(t)$ be the budgeted funds in year t, and let $C_0(t)$ be the predicted cost that will be incurred in year *t.* There are many questions we might wish to ask about the program's affordability risk:

- In how many years will the program exceed the planned budget?
- How many total dollars over budget will the program spend?
- What is the probability of exceeding the budget at least once over the FYDP?
- How much contingency funding would be needed to achieve 90% confidence of staying within budget, assuming unspent contingency carries over to the next year?

These are all questions of potential interest to both Program Managers and Resource Managers. Using the posterior final profile distribution derived from the original profile C_0, we can perform many counterfactual Monte Carlo analyses to answer these kinds of questions. The general pattern for these analyses is as follows:

1. Given the initial development estimate for a program…
2. Define a yearly budget level $B(t)$, and a contingency fund size (if any).
3. Use the regression described previously to determine the posterior distribution on the parameters of the best fit to the final actual development profile for the program.
4. Define outcomes or events of interest.
5. For s = 1,…,S (indexing over iterations of the Monte Carlo algorithm):
    a. "Draw" random parameter vector $\theta_1^{(s)}$ from the posterior distribution.
    b. Compute the corresponding yearly values by evaluating the best fit curve at t = 1, …, $T_1^{(s)}$ and computing $K^{(s)}\,W\left(t|\theta_1^{(s)}\right)$..
    c. Add random noise $E(t)$ drawn from the estimated distribution of the historical errors in fitting curves to final development cost profiles to get a final spending profile $C_1^{(s)}(t), t = 1, \dots, T_1^{(s)}$.
    d. Evaluate and store any events or outcomes of interest.

Note that the value of $T_1^{(s)}$ used in step 5b is determined as part of $\theta^{(s)}$ in step 5a.

After S iterations, calculate the statistics of interest over the stored events or outcomes. For example, count the number of times N that $C_1^{(s)}(t) < B(t)$ for t = 1…5 and compute $N/_S$. This is the estimated probability of staying within budget for the first five years. The Monte Carlo framework can also allow comparison of different management policies. For example, one could compare the effect of pre-allocating contingency to specific program years versus maintaining a contingency fund to be spent down over time as needed.

In general, we would do this not only for development profiles, but also for procurement spending. In that case, policy makers might be interested in how much difference it would make to be able to manage both RDT&E and Procurement using a single combined budget and/or a single program contingency fund, rather than having to manage separate budgets and contingency amounts due to “color of money” prescriptions.

Example

Consider the H-99 program. Figure 5 shows the mean of the final estimated profile distribution, shown with the original (blue) estimated profile for comparison. Note that not only is the average predicted outcome longer and more expensive in total, but it also ramps up more slowly than the estimate, so that the year-by-year errors change sign over time.

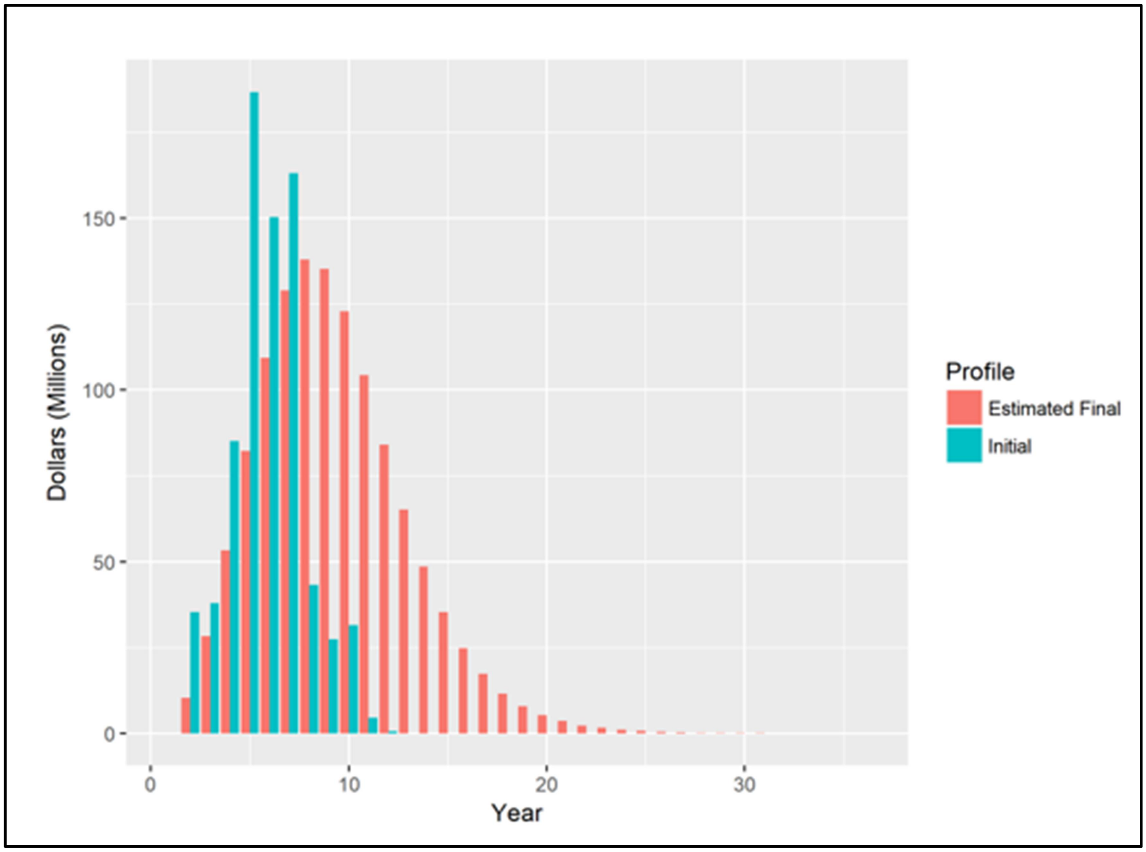

Figure 5. H-99 Original Estimate and Mean Estimated Final Profile

Figure 6 plots the initial planned profile (solid blue) against three different yearly quantile estimates. The dashed gray line shows the median predicted actual expense in each future year estimated from the Monte Carlo simulation. By “median profile,” we mean the year-by-year median spending levels $\mathrm{median}_{\theta_1}(C_1(t))$ over the distribution of all profiles; the shape of the dashed gray line does not correspond to any one set of θ_1 values. The solid black lines in Figure 6 show the original cost estimate profile against upper and lower 10% probability bands. The interpretation of these bands is that, in a given year, there is a 10% chance that the H-99 program will receive less funding than the lower band, and a 10% chance that the H-99 program will receive more funding than the upper band. Note that these probabilities are conditional on looking forward at Milestone II/B—if you already know that a program fell below the 10% band in year 4, this plot provides no information about what to expect in years 5 and beyond.

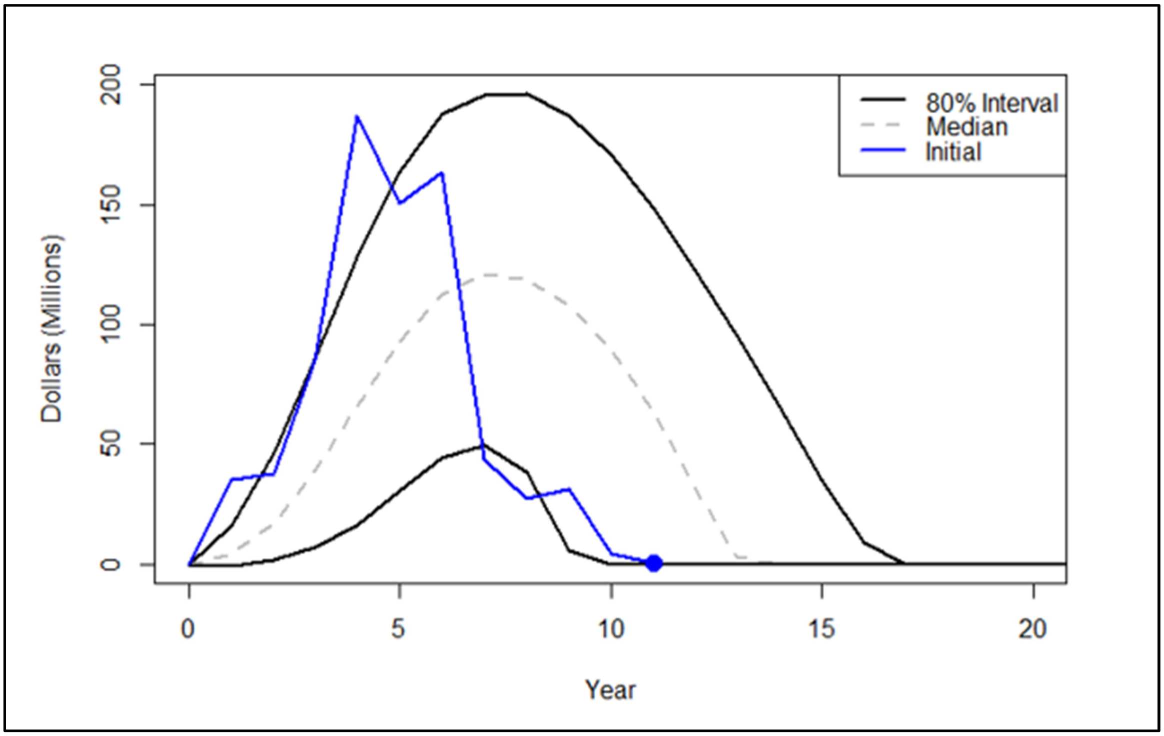

Figure 6. Quantiles of Predicted Annual H-99 Costs

Suppose you, as the Resource Manager for H-99, are interested in how the program will perform in its first FYDP, that is, the first five years of its profile. You wish to know how much additional contingency funding you should expect to require. Table 1 gives the expected overages for consecutive five-year periods.

Table 1. Expected Budget Overages in Five-Year Bins

Overage (Millions)	2.6	336.6	333.4	67.0	9.2	1.4
Years	1–5	6–10	11–15	16–20	21–25	26–30

The table shows that, on average, you would need $2.6 million more than was originally budgeted in years 1 through 5. In contrast, in years 6 through 10, you would need $336.6 million more than planned, on average. Note that even though this table goes out to 30 years, it is unlikely that spending would continue for 30 years. These averages include spending of $0 for outcomes where the program ends prior to that time frame.

It is important to remember that these single estimates of overage are summaries of distributions. Figure 7 shows a box-and-whisker plot that helps to visualize the distribution of the FYDP overages. In each five-year bin, there is a large point mass at zero overage and a highly-skewed distribution of nonzero overages. In all but the second and third bins, the median overage is zero.

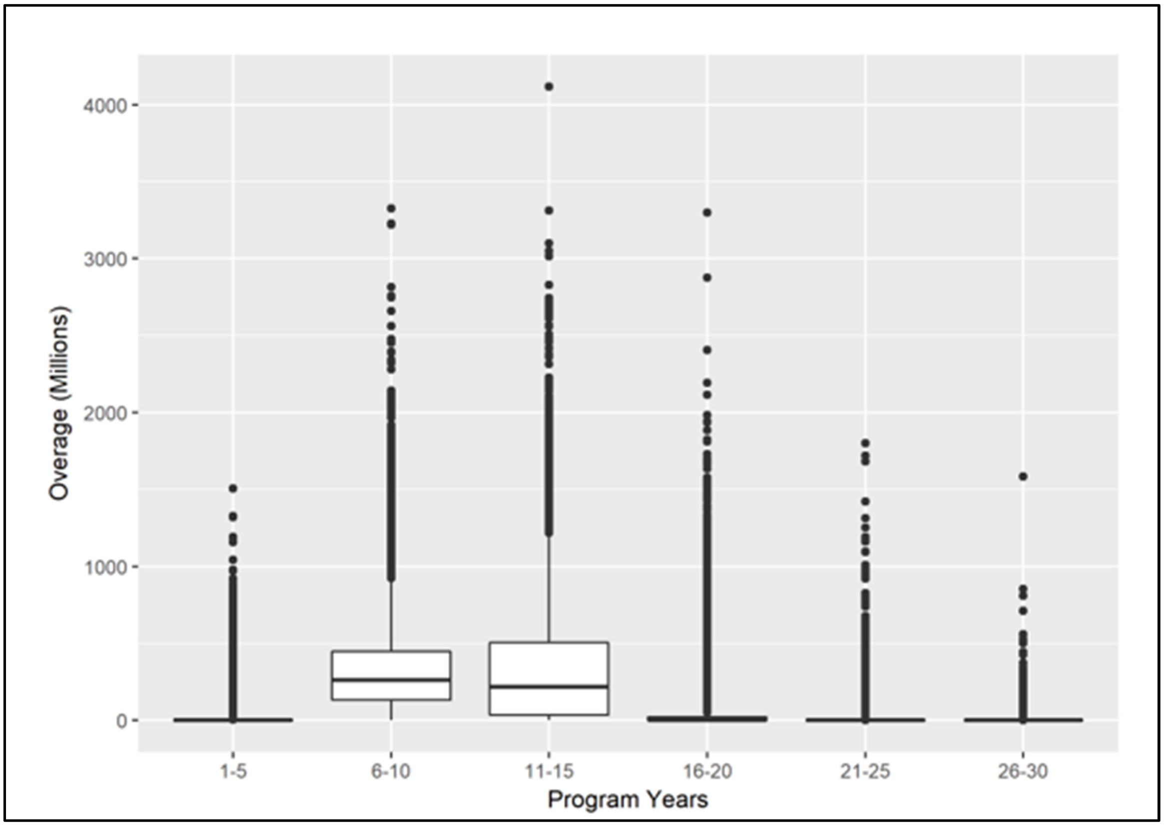

Figure 7. Distribution of Total Cost Overages in Five-Year Bins

Characterizing Mid-Life Programs

We have shown how our model can characterize the affordability risk of a new program's development budget. However, programs are only new once. For a program in the middle of its development, we would like to be able to take advantage of the program's history to date to make an updated, more precise characterization of the remaining cost risk.

A straightforward approach to this might be to add new predictive variables to the functional regression model, reflecting factors such as the age of the development program in years, the relative cost and schedule growth to date (compared to the original estimate), and the relative growth in the program's estimated cost at completion. These could be combined with a revised best-fit functional curve reflecting the program's actual history. In this approach, the same regression model would be used for all points in the development life cycle, with the original estimated profile being a special case with program age = 0 and cost growth factor = 1.

A second possible approach would be to use different regression models for programs at different points in their life cycle. This approach would have the advantage that different underlying functional forms could be used for profiles (or for remaining profiles), depending on the actual outcomes thus far. However, each prediction would then be based on a smaller historical data set, with corresponding loss of statistical power.

Using either of these approaches and the Monte Carlo framework, it would be possible to characterize the future year-by-year development cost risk of every program. Extending the technique to procurement cost risk is also (again) straightforward—the relevant predictive factors might be how many years the program has been in production, how actual unit costs compare to predicted unit costs, any change in planned quantity, etc. As we discuss below, this would allow defense resource analysts to completely characterize the collective behavior of the entire acquisition portfolio, or any subset of it.

Portfolios: More Than One Program at a Time

We have shown how our model can characterize the affordability risk of a single program's development budget. In practice, it would be even more useful to be able to characterize the affordability risk of a group of projects or programs being managed with a common contingency pool. If the outcomes of these programs are approximately independent, this is not much more complicated than the single-program case.

If we have estimated the $F_{posterior}(\theta_1)$ and $E(t)$ distributions for each of a set of programs, we can apply the same kind of Monte Carlo analysis to the sum of their annual costs, compared against a collective portfolio budget and contingency fund. This could be done separately for RDT&E and Procurement, each with its own budget, or it could be done using a combined investment budget. This would enable true affordability analysis of portfolios as envisioned by the Better Buying Power initiatives,[4] but with considerably more realism than current affordability analyses that are based on point-estimate cost profiles assuming fixed program content and quantities.

One potential use of such a model would be to quantify the benefits of portfolio-level contingency funding versus program-level contingency funding. It is well known in the project management world that allocating reserve funds to specific cost areas before you actually know where the cost growth is going to occur leads to less efficient use of those reserve funds. However, it is often politically impossible to protect funds that are not part of the base budget for some cost element. In the DoD, apart from a highly limited ability to reprogram funds from one program element or line item to another, there is currently no ability to reserve funds for contingency use outside of a specific program's budget.

Potential Criticisms of the Method

At this point, it would not be unreasonable to object that this modeling approach assumes that the DoD is incapable of learning to estimate costs more accurately or to contain cost growth more effectively. Worse yet, if Resource Managers were to actually use tools like these to manage portfolios of programs more efficiently, the resulting changes in program outcomes might invalidate the models.

There are perhaps two points of optimism:

- If use of these tools improves the efficiency of contingency funding of portfolios, and therefore reduces cost and schedule growth due to program stretches and funding instability, that will tend to cause users of the model to overestimate the required contingency. That isn't a bad thing; the opportunity cost of allocating a little too much contingency is far less than the marginal cost of allocating too little contingency. As noted previously, even small funding shortfalls lead to additional cost growth due to schedule stretches, increased fixed costs and overhead, and inefficient use of resources.

---

[4] DoD, *Better Buying Power*, http://bbp.dau.mil/index.html.

- Over time, as the outcomes of recent programs are fed back into the regression models, the posterior distributions produced by the model will correctly adjust to the new reality. If the changes are substantial, one could include “year of program start” as a predictor variable to capture the trend of improvement.

In the meantime, if the drivers of program outcomes are systemic—built into the incentive structure of defense acquisition—the models will continue to capture the likely results of those incentives.

A different potential criticism of this work is that it offers no insights into why costs and schedules deviate from their original estimates or how this could be “fixed.” That is correct—we take cost and schedule changes as given. We do not distinguish between “good” and “bad” causes. If a program is cancelled after seven years, that is just another source of negative cost and schedule growth. If a system turns out to be so useful that the original buy is tripled, or successive block upgrades continue for 30 years, that is just another source of positive cost and schedule growth. The difference between those is very important to warfighters, but it is not relevant to the question of how many dollars we can expect to spend within these programs over the next N years.

Finally, we note that this method explicitly models how much funding a program *will receive* in a given year—not how much it needs, or ought to receive, or would receive if there was more money to go around. As such, the model data incorporate the history of negotiations between the Services, the Office of the Secretary of Defense, and the Congress regarding how much to fund programs year by year, and when to cancel them. If there were to be a fundamental change in the dynamic of how those decisions are made, then that, too, might invalidate the link between historical outcomes and future program outcomes, at least until enough new data could be collected.

Conclusion

Quantifying Annual Resource Risks for a Program or Portfolio

To a first approximation, acquisition programs never spend what they said they would when they began. In fact, the error bars around an initial cost estimate are much larger than is generally understood once program cancellations, restructurings, truncations, and block upgrades have been accounted for. Worse yet, all of this uncertainty arises in a context where programs must fit within annual budgets—it is not enough to only spend as much as you said you would; you must also spend it when you said you would, or problems ensue.

We have developed a methodology to characterize the year-by-year budget risk associated with a major acquisition program. This methodology can be applied to both development costs and procurement costs and can be extended to understand the aggregate affordability risk of portfolios of programs. The method allows Resource Managers to estimate annual budget risk levels, required contingency amounts to achieve a specified probability of staying within a given budget, and a host of other relevant risk metrics for programs. It also allows policy makers to predict the impact on program affordability of proposed changes in how contingency funds are managed.

Future Research

This technique is currently in the prototype stage and is based on a relatively sparse set of historical program outcome data. There is still much work to be done on improved statistical techniques for the functional regressions, modeling of procurement profile risk, conditional modeling of procurement given development outcomes, and characterization of the distribution of residuals around the best-fit functional curve. There is also a great deal to be learned about how managers could best use the information provided by this method to manage actual programs and portfolios, and what the implications might be for recommending changes to acquisition law and regulations.

References

Adoko, M. T., Mazzuchi, T. A., & Sarkani, S. (2015). Developing a cost overrun predictive model for complex systems development projects. *Project Management Journal, 46*(6), 111–125. doi:10.1002/pmj21545

Anderson, T. P., & Cherwonik, J. (1997, Summer). Cost estimating risk and cost estimating uncertainty guidelines. *Acquisition Review Quarterly*, 339–348. Retrieved from www.dtic.mil/get-tr-doc/pdf?AD=ADA487532

Arena, M. V., Leonard, R. S., Murray, S. E., & Younossi, O. (2006). *Historical cost growth of completed weapon system programs* (TR-343-AF). Santa Monica, CA: RAND. Retrieved from https://www.rand.org/content/dam/rand/pubs/technical_reports/2006/RAND_TR343.pdf

Asher, N. J., & Maggelet, T. F. (1984). *On estimating the cost growth of weapon systems* (IDA Paper P-1494). Alexandria, VA: Institute for Defense Analyses.

Bitten, R. E., & Hunt, C. (2017). *Assessing the impact of confidence levels in funding and budgeting NASA science missions.* Presentation at the ICEAA Professional Development & Training Workshop, Portland, OR.

Bliss, G. (1991, June). *The accuracy of weapon systems cost estimates*. Presentation at the 59th Military Operations Research Symposium, U.S. Military Academy, West Point, NY.

Brown, G. E., White, E. D., Ritschel, J. D., & Seibel, M. J. (2015). Time phasing aircraft R&D using the Weibull and Beta distributions. *Journal of Cost Analysis and Parametrics, 8*(3), 150–164. doi: 10.1080/1941658X.2015.1096219

Chib, S., & Greenberg, E. (1995). Hierarchical analysis of SUR models with extensions to correlated serial errors time-varying parameter models. Journal of Econometrics, 68, 339–360. doi:10.1016/0304-4076(94)01653-H

Congressional Budget Office (CBO). (2017). *An analysis of the Obama administration's final future defense program* (CBO Report No. 52450). Retrieved from https://www.cbo.gov/system/files/115th-congress-2017-2018/reports/52450-fydp.pdf

Coonce, T., Bitten, R., Hamaker, J., & Hertzfeld, H. (2010). NASA productivity. *Journal of Cost Analysis and Parametrics, 3*(1), 59–78. doi:10.1080/1941658X.2010.10462228

Creedy, G. D., Skitmore, M., & Wong, J. K. W. (2010). Evaluation of risk factors leading to cost overrun in delivery of highway construction projects. *Journal of Construction Engineering and Management, 136*(5), 528–537. Retrieved from https://ascelibrary.org/doi/abs/10.1061/%28ASCE%29CO.1943-7862.0000160

Drezner, J. A., Jarvaise, J. M., Hess, R. W., Hough, P. G., & Norton, D. (1993*). An analysis of weapon system cost growth.* Santa Monica, CA: RAND. Retrieved from https://www.rand.org/content/dam/rand/pubs/monograph_reports/2006/MR291.pdf

Drezner, J. A., & Smith, G. K. (1990). *An analysis of weapon system acquisition schedules* (R-3937-ACQ). Santa Monica, CA: RAND. Retrieved from https://www.rand.org/pubs/reports/R3937.html

GAO. (2006). *Space acquisitions: DOD needs to take action to address unrealistic initial cost estimates of space systems* (GAO-07-96). Washington, DC: Author. Retrieved from https://www.gao.gov/products/GAO-07-96

Glennan, T. K., Bodilly, S. J., Camm, F., Mayer, K. R., & Webb, T. (1993). *Barriers to managing risk in large scale weapons system development programs* (MR-248-AF). Santa Monica, CA: RAND. Retrieved from https://www.rand.org/pubs/monograph_reports/MR248.html

Jessup, R., & Williams, J. (2015). *The cost of commonality: Assessing value in joint programs* (MBA professional report). Monterey, CA: Naval Postgraduate School.

Jimenez, C., 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. doi:10.1080/1941658X.2016.1201024

Light, T., Leonard, R. S., Pollak, J., Smith, M., & Wallace, A. (2017). *Quantifying cost and schedule uncertainty for major defense acquisition programs (MDAPs).* Santa Monica, CA: RAND. doi:10.7249/RR1723

Marshall, A. W., & Meckling, W. H. (1959). *Predictability of the costs, time, and success of development* (P-1821). Santa Monica, CA: RAND. Retrieved from https://www.rand.org/pubs/papers/P1821.html

McNicol, D. L. (2004). *Cost growth in major weapon procurement programs* (2nd ed.). Alexandria, VA: Institute for Defense Analyses.

McNicol, D. L. (2017). *Influences on the timing and frequency of cancellations and truncations of major defense acquisition programs* (IDA Paper P-8280). Alexandria, VA: Institute for Defense Analyses.

McNicol, D. L., Gould, B. R., Wu, L., Fuchs, E., & Bronson, P. (2013). *Evidence on the effectiveness of the Defense Acquisition Board process* (IDA Paper P-5011). Alexandria, VA: Institute for Defense Analyses.

Tate, D. M. (2016). *Acquisition cycle time: Defining the problem* (IDA Document NS D-5762). Alexandria, VA: Institute for Defense Analyses.

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

Tyson, K. W., Nelson, J. R., Om, N. I., & Palmer, P. R. (1989). *Acquiring major systems: Cost and schedule trends and acquisition initiative effectivenes*s (IDA Paper P-2201). Alexandria, VA: Institute for Defense Analyses.

Tyson, K. W., Nelson, J. R., & Utech, D. M. (1992). *A perspective on acquisition of NASA space systems* (IDA Document D-1224). Alexandria, VA: Institute for Defense Analyses.

Acknowledgments

This work was originally conceived and funded by the congressionally established Section 809 Panel (https://section809panel.org/), and in particular by Commissioners Dr. Allan Burman and David Ahern. Darren Harvey of the panel technical staff provided valuable guidance and feedback. Dan Cuda and James Bishop of the IDA were technical reviewers, and their observations and suggestions improved the paper significantly.

ACQUISITION RESEARCH PROGRAM
GRADUATE SCHOOL OF BUSINESS & PUBLIC POLICY
NAVAL POSTGRADUATE SCHOOL
555 DYER ROAD, INGERSOLL HALL
MONTEREY, CA 93943

www.acquisitionresearch.net