

# CHARACTERIZING THE ACCURACY OF DoD OPERATING AND SUPPORT COST ESTIMATES

*Erin T. Ryan, David R. Jacques, Jonathan D. Ritschel, Christine M. Schubert  
Air Force Institute of Technology, Wright-Patterson Air Force Base*

**Abstract.** For decades, the DoD has employed numerous reporting and monitoring tools for characterizing the acquisition cost estimates of its programs. These tools have led to dozens of studies thoroughly documenting the magnitude and extent of DoD acquisition cost growth. However, little attention has been paid to the behavior of the other main cost component of a system's life cycle cost: Operating and Support (O&S) costs. Consequently, the DoD has little knowledge regarding the accuracy of O&S cost estimates or how that accuracy changes over time. In a previous paper, the authors describe an analytical methodology for remedying this deficiency via a study to characterize the historical accuracy of O&S cost estimates. The results are presented here, and indicate there tend to be large errors in DoD O&S cost estimates, and that the accuracy of the estimates improves little over time.

## INTRODUCTION AND MOTIVATION

The DoD's definition of Life Cycle Cost (LCC) is the total cost to the government spanning all phases of the program's life: development, procurement, operation, sustainment, and disposal (DoD 1992). Note that this definition includes some costs accrued before a system formally enters the acquisition phase (e.g., concept refinement and technology development) as well as certain costs accrued as the system transitions out of sustainment (e.g., demilitarization and disposal). These initial and final costs—though sometimes sizeable from an absolute perspective—tend to be negligible when compared to the costs incurred during the program's acquisition phase and its Operations and Support (O&S) phase (DoD CAIG 2007; DAU 2012). Consequently, one can state, to a high degree of accuracy, that a system's LCC is simply the sum of its total acquisition costs and its total O&S costs.

Of these two cost components, the DoD has historically placed significantly greater emphasis on the acquisition side of the equation. Over the years, a plethora of control and oversight accountability mechanisms have been implemented with the expressed purpose of improving the execution and/or management of the acquisition phase of defense programs. Meanwhile, the O&S sustainability considerations have been perennially neglected or subordinated to acquisition requirements or program survival (DoD 2009; Choi et al. 2009).

By virtue of its traditional focus on the acquisition component of a system's life cycle, the DoD has managed to gain a variety of valuable insights into the nature of the acquisition costs of defense, including how accurate acquisition cost estimates are and how they tend to behave over time. These insights have provided the framework for many revisions to the acquisition process and provided the opportunity for numerous improvements to the acquisition cost component of a system's LCC.

The same cannot be said for the O&S cost projections. Consider that between 1945 and 2008, there were at least 130 separate studies and commissions focused on the acquisition of DoD systems, dozens of which involved the accuracy of the cost estimates (DoD 2009). During this same time period, there appears not to be a single published study pertaining to the accuracy of O&S cost estimates.

Therefore, *the DoD simply still does not know how O&S cost estimates compare to reality.* This means that DoD emphasis on a program's life cycle cost is effectively a hollow requirement. Consider the following statement by the Director of the Cost Assessment and Program Evaluation—

*“The strategic intent is to emphasize how O&S cost estimates will actively support key decisions throughout the system life cycle, rather than calling for O&S cost estimates simply for the sake of having an estimate”* (OSD/CAPE 2011).

Without knowledge of the validity of a program's O&S cost estimates, the Pentagon cannot have confidence in its LCC estimates. And without confidence in LCC estimates, efforts to reduce LCCs are largely nullified, and attempts to meaningfully discern the value of competing systems or designs based on their respective LCCs are rendered futile.

The premise of this paper is thus anchored in the notion that it is critical for the DoD to be able to characterize the accuracy of LCC estimates for its programs. The authors have completed a comprehensive study of major defense programs in an attempt to achieve this goal. The results presented in this paper have important implications to decisions based on O&S cost estimates, and potentially to how these estimates are constructed and interpreted. It is worth emphasizing, however, that the sole purpose of this paper is to report objective results; the authors do not attempt to interpret the results or offer recommendations on how to improve cost estimating performance.

## **BACKGROUND**

In part due to the recognition of these deficiencies related to the O&S component of life cycle cost, the DoD has sought to better formalize its emphasis on the importance of life cycle costs, most notably through the Weapons Systems Acquisition Reform Act of 2009 (DoD 2009). The result has been a relative flurry of studies (i.e., four) over the last couple of years that examine O&S costs for defense programs (DoD 2009; Choi et al. 2009; Balaban, Devers and Roark 2010; GAO 2010). While this is a welcome development, unfortunately these studies have limited applicability to the task at hand. As detailed by the authors in a previous paper, only one of these published studies includes results that employ a methodology that allows for genuine qualitative characterization of the accuracy of O&S cost estimates. Moreover, the scope of this study was extremely narrow, only involving two weapon systems (Ryan et al. 2012). This further speaks to the potential importance of the results presented in this paper.

Although the enactment of WSARA has served as a catalyst for increased focus on O&S cost issues in the DoD, it is also the case that an empirical analysis of O&S costs for defense systems would simply not have been possible until relatively recently. To conduct the analysis requires three elements: a valid source of predicted costs from the acquisition phase, a valid source of actual costs from the O&S phase, and enough elapsed time for a large number of programs to have accrued representative data from both phases.

Regarding predicted O&S costs for major defense programs, the obvious source is the Selected Acquisition Report (SAR). SARs are required to be submitted at least annually for all major defense programs until they have been 90 percent acquired (DAU 2012). Starting in 1989, programs were required to include as part of each SAR a “full life-cycle cost analysis,” and soon thereafter (about 1990 in most cases), programs began providing estimates of Annual Unitized

O&S cost (i.e., average O&S cost per unit, per year). Starting in 2001, most programs also began providing an estimate of Total O&S cost.

Not long after O&S cost estimates were required in the SARs, the DoD mandated that each military service maintain an historical database of *actual* O&S costs for its systems. The effort became known as Visibility and Management of Operating and Support Costs (VAMOSC), with each service component managing its own instantiation. Though the central focus of VAMOSC is on future planning and the development of O&S cost estimates, the nature of the database often allows for actual O&S costs to be broken out by weapon system and year (DoD 1992; DAU 2012). Accordingly, VAMOSC data can serve as “ground truth” for actual system O&S costs, thus enabling an accuracy assessment of the O&S cost estimates found in the SARs.

So with a consistent, time-phased, and relatively reliable source for predicted O&S costs in one hand, and a viable source for obtaining actual O&S costs segregated by system in the other, all that remains is allowing enough time to pass for a sufficient amount of data to be collected to enable a valid comparison between the two cost figures. Since programs can take many years to develop and field, it is reasonable to suppose that it could take multiple decades to amass sufficient data for a substantive analysis. In fact, the authors have screened all the data, and found that two decades (~1990-2010) may be enough time to obtain reasonably valid O&S cost estimates and actuals for over three dozen major defense programs. In other words, it has taken some time for the salient data elements involving a study like this to be established and amassed, and now the time is ripe to exploit the data.

## METHODS

At this point, the basic research methodology should be evident. Our first step is to annotate the predicted O&S cost estimate from every SAR for a given system. We next use the VAMOSC data to establish the actual O&S costs for that system. Finally, we compare the actual O&S cost of the system to what it was predicted to be each year during its acquisition phase and characterize the accuracy of that estimate over time. The principle is simple enough, but there are a number of obstacles that complicate the proposed analytical methodology, which are outlined in detail in the authors’ predecessor paper, which will henceforward be referred to as simply the *associated Methodology Paper* (Ryan et al. 2012). At minimum, however, the reader should be familiar with the two different O&S cost measures used in this study.

### O&S Cost Measures

There are two distinct measures of O&S cost that will be assessed via this proposed methodology: *Total O&S Cost* and *Annual Unitized O&S Cost (AUC)*. Each measure—within the context of this study—has its own strengths and weaknesses. The Total O&S cost is a readily intuitive metric that offers a direct means of inferring program LCCs via summation with the final (or most recently reported) total acquisition cost for that program. LCCs are arguably the most comprehensive and facile cost indicator for system value assessments. Consequently, the Total O&S cost measure (after being summed with acquisition costs to produce an LCC) serves as the basis for our first dependent variable: *Error in LCC Estimate*. Values are reported as percentages, with negative values indicating that the LCC estimates were lower than the actual life cycle costs, and positive values indicating that the LCC estimates were too high.

Although the AUC metric does not provide direct insight into the system LCC, the data tends to be more broadly available—both in terms of estimates and actuals—thereby enabling analysis of a greater number of programs over a longer span of time. Moreover, unitized O&S costs are a

commonly employed metric for assessing sustainment costs, in part because the estimates tend to exhibit reduced variability over time (e.g., less perturbed by quantity fluctuations) than total O&S costs. In addition, unitized O&S costs (as opposed to total O&S costs) often provide a more valid comparative measure across similar contemporary or antecedent systems, in part because the AUC is not dependent on the operational service life assumption<sup>1</sup>. Therefore, this measure of O&S costs is also useful and will form the basis of the second dependent variable in this study: *Error in AUC Estimate*. Like the first dependent variable (Error in LCC Estimate), error values are reported as percentages, with negative values corresponding to AUC estimates lower than the actual reported costs.

## Data Weighting

As noted in the associated Methodology Paper, potential data reliability concerns often arose while compiling and reducing the data. Sources of concern encountered on the estimate side (i.e., in the SARs) included unstated assumptions, incorrect units of measurement, inconsistencies between tabular values and accompanying narrative text, and uncertainty in the final acquisition cost associated with less mature programs. For the cost actuals (i.e., VAMOS data), reduced confidence in the data could result from having too few years of O&S cost actuals and/or the occasional, inexplicably high variability in year-to-year reported costs of the program.

In an attempt to account for these data quality concerns, a detailed rating system was developed to score the overall reliability of each row of the data relative to both dependent variables. This rating was used as a weighting factor in the analysis with the expectation that it would improve the accuracy of the findings. However, after a thorough comparison of results obtained both with and without the weighting applied, we found that the weighting factors exhibited only a minor effect on the results. The values for measures of central tendency were a bit different in some cases, but never more than ten percent, and there were no instances in which conclusions regarding statistical significance were affected by the use of the weighting factor.

The fact that an attempt to weight the analysis by perceived reliability of the data had no significant impact on the results of the study could indicate one of at least three things. First, it could indicate the data was more reliable than it seemed to be. Second, it may be that the weighting system was not effectively constructed or implemented. Third, it may be the case that the underlying statistical trends in the data are strong enough not to be significantly influenced by a few invalid entries. We are inclined to suspect the latter explanation, which would suggest to others pursuing this type of analysis that a comprehensive weighting strategy may not be worth the additional effort and complexity.

## Mixed Models

The associated Methodology Paper is largely focused on the task of structuring the study and reducing the data. Specific *analytical* methodologies, however, were not addressed in that paper, and thus will be briefly discussed here.

This effort is fundamentally a longitudinal data analysis, and the nature of how the data is organized precludes the possibility of assuming an identical and independent distribution (i.i.d.) of the random variables. Because the data is clustered into programs, with repeated measurements of each program over time, there necessarily exists a correlation between the repeated

---

<sup>1</sup> Every major DoD system being acquired must specify the expected operational service life for the purpose of sustainment budgeting and LCC estimating.

measurements for a given program—and therefore the statistical errors of the observations—that must be accounted for in the statistical analysis. Further, we expect these correlations to be greater for data points close in time, such as for successive SARs from the same program. This also means that the statistical errors will be correlated as well.

Importantly, the fact that we expect correlated errors for the programs in this study invalidates the underlying assumptions of simple analysis of variance and regression models, namely i.i.d. observations. To compensate for this, we instead employ mixed model techniques for the data in this study. For readers not familiar with mixed models, the term refers to the use of both fixed (i.e., entire population) effects and random (i.e., subject-specific) effects within the same analysis. In this study, the subject-specific effects can be regarded as those that are unique to each program. The key distinction between mixed models and simple regression models is that the former can produce valid models even if the subject observations are not independent. In essence, mixed models allow the data to exhibit inherent correlations and non-constant variability that arise from the program-specific effects. This allows one to effectively model not only the measures of central tendency for the data, but also the covariance structure attributable to the repeated measurements (Diggle, Liang and Zeger 1994; Verbeke and Molenberghs 2000).

## RESULTS

### Summary Statistics

Before presenting the results of this study, it is worth reviewing the nature of the data being analyzed. This study encompasses 37 major defense programs: 25 Navy, and 12 Air Force. (For reasons detailed in the associated Methodology Paper, Army programs were not included in this study (Ryan et al. 2012)). Because of issues related to inflation and joint procurement (again see the associated Methodology Paper), it was frequently necessary to create artificially separate analytical units for some programs (e.g., “F-22 base year 1990” costs and “F-22 base year 2005” costs, or the “Air Force AMRAAM” and “Navy AMRAAM”). In this study, each of these lower-level analytical units is referred to as a *SubProgram*, which may or may not be the same as the Program. Using this terminology, there are 46 SubPrograms in this study: 28 Navy, and 18 Air Force. Table 1 lists all these SubPrograms (and Programs), along with the lead DoD service component, the type of system, and the count and time range of available SARs since 1987.

The SubProgram is the fundamental unit of analysis in this study. Based on this measure, the data is clearly dominated (61 percent) by Navy programs. Another way to view the distribution of data is to consider the relative proportion of SARs, with each SAR effectively representing one “row” of data. Figure 1 shows the distribution of SARs by DoD Component, and by this measure, the bias toward Navy programs is even more pronounced (66 percent).

With respect to the type of system, the categories of *Aviation* and *Maritime* are self-explanatory. The *Munition* system type category includes any expendable weapon such as bombs or missiles. Systems which do not fit into these first three categories are designated as *Other*. As shown in Figure 2, the large majority of the systems (83 percent) are categorized as either *Aviation* (50 percent) or *Maritime* (33 percent). Because of the relatively small sampling of systems that do not fit into these two categories, as well as concerns regarding the basic validity of using a unitized O&S cost metric for non-traditional system types (as addressed in the associated Methodology Paper), most analyses are conducted both on all programs as well as the subset of *Aviation* and *Maritime* system types. Wherever this segregation of system types affects the results, it will be noted.

**Table 1. Summary of Programs Included in Study**

#	Program Name	SubPrgm Name	Lead Comp.	Sys. Type	SAR Yrs.	# of SARs
1	AIM-9X	AIM-9X (Navy)	Navy	Munition	1996-2001	6
2	AMRAAM	AMRAAM-AF	Air Force	Munition	1988-1992	5
3	AMRAAM	AMRAAM-Navy	Air Force	Munition	1988-1992	5
4	AMRAAM	AMRAAM-Jt	Air Force	Munition	1992-2010	18
5	AOE 6	AOE 6	Navy	Maritime	1988-1997	11
6	AV-8B REMAN.	AV-8B REMAN.	Navy	Aviation	1994-2002	10
7	C-130J	C-130J	Air Force	Aviation	1996-2010	13
8	C-17A	C-17A (BY81)	Air Force	Aviation	1987-1994	10
9	C-17A	C-17A (BY96)	Air Force	Aviation	1995-2010	14
10	C/MH-53E	C/MH-53E	Navy	Aviation	1987-1994	9
11	CVN 68	CVN 68 (74/75)	Navy	Maritime	1987-1998	13
12	CVN 68	CVN 68 (76)	Navy	Maritime	1994-2002	9
13	DDG 51	DDG 51	Navy	Maritime	1987-2010	25
14	E-2C	E-2C	Navy	Aviation	1994-2006	14
15	EELV	EELV	Air Force	Other	1998-2007	13
16	F-14D	F-14D	Navy	Aviation	1987-1993	9
17	F-16C/D	F-16C/D	Air Force	Aviation	1987-1994	8
18	F-22	F-22 (BY90)	Air Force	Aviation	1991-2004	16
19	F-22	F-22 (BY05)	Air Force	Aviation	2005-2010	6
20	F/A-18C	F/A-18C	Navy	Aviation	1987-1994	10
21	F/A-18E/F	F/A-18E/F (BY90)	Navy	Aviation	1992-1999	9
22	F/A-18E/F	F/A-18E/F (BY00)	Navy	Aviation	2000-2010	10
23	GLOBAL HAWK	GLOBAL HAWK	Air Force	Aviation	2001-2010	11
24	JASSM	JASSM	Air Force	Munition	1999-2009	12
25	JPATS	JPATS (BY1995)	Air Force	Aviation	1995-1999	5
26	JPATS	JPATS (BY2002)	Air Force	Aviation	2001-2010	9
27	JSOW	JSOW	Navy	Munition	1997-2010	14
28	JSTARS	JSTARS (BY83)	Air Force	Aviation	1989-1996	10
29	JSTARS	JSTARS (BY98)	Air Force	Aviation	1997-2003	6
30	KC-135R	KC-135R	Air Force	Aviation	1987-1994	8
31	LHD 1	LHD 1	Navy	Maritime	1987-2005	18
32	LPD 17	LPD 17	Navy	Maritime	1996-2010	16
33	MH-60R	MH-60R (BY93)	Navy	Aviation	1994-2005	14
34	MH-60R	MH-60R (BY06)	Navy	Aviation	2006-2010	5
35	MH-60S	MH-60S	Navy	Aviation	1998-2010	17
36	MHC 51	MHC 51	Navy	Maritime	1991-1998	8
37	NESP	NESP	Navy	Other	1992-2004	14
38	PREDATOR	PREDATOR	Air Force	Aviation	2009-2010	2
39	SSGN	SSGN	Navy	Maritime	2002-2007	6
40	SSN 21	SSN 21	Navy	Maritime	1987-1999	15
41	SSN 774	SSN 774	Navy	Maritime	1995-2010	16
42	STRAT. SEALIFT	STRAT. SEALIFT	Navy	Maritime	1993-2001	11
43	T-45TS	T-45TS (BY84)	Navy	Aviation	1987-1993	10
44	T-45TS	T-45TS (BY95)	Navy	Aviation	1994-2007	14
45	T-AKE	T-AKE	Navy	Maritime	2001-2010	10
46	T-AO 187	T-AO 187	Navy	Maritime	1987-1994	8

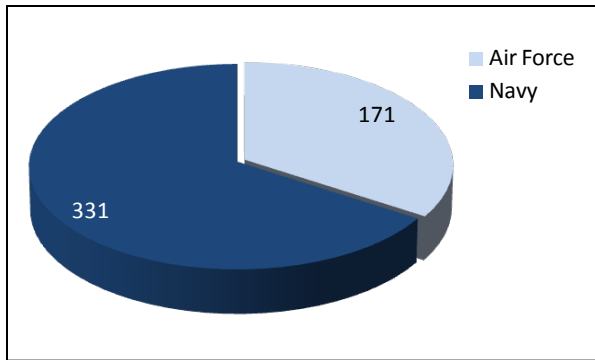


Figure 1. Distribution of SARs by DoD Service Component

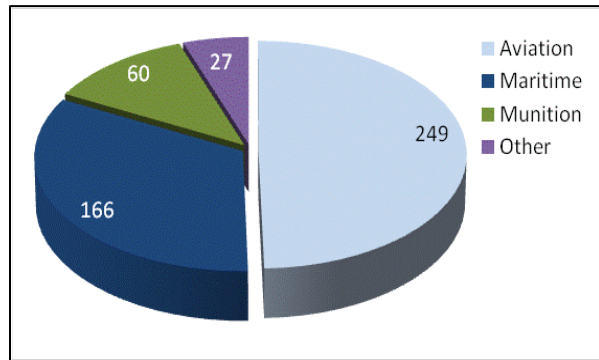


Figure 2. Distribution of SARs by System Type

The most logical measure for ascertaining the accuracy of O&S estimates is to calculate the mean value across all the SubPrograms. However, this is not as straightforward as it might seem. Because of the fixed effects associated with a given SubProgram, we cannot simply calculate the mean value across all 502 SARs, as that would bias the data toward those SubPrograms with more SARs. Nor can we calculate the mean of all the individual SubProgram means as that would provide equal weighting to all of the SubPrograms, which is not appropriate given the disparate standard errors and random variances associated with each SubProgram. The solution is to analyze the data using the mixed model approach, which inherently accounts for these fixed and random effects. The following discussion provides a detailed characterization of the observed estimating errors for the two dependent variables, beginning with the *Error in AUC Estimate* and then proceeding to the *Error in LCC Estimate*.

### Characterizing the Accuracy of AUC Estimates

Figure 3 and Figure 4 show the mean value for the Error in AUC Estimate as deviation contrasts between DoD Component and Type of System, respectively. In these plots, the horizontal red line indicates the overall estimated mean and the vertical bars represent the 95 percent confidence deviation contrasts. In any case where the contrasts do not touch the red line, we conclude there is a statistically significant difference between the parameter values with respect to Error in AUC Estimate. Thus, although the data in Figure 3 suggests that the AUC for Air Force programs tends to be underestimated, while the AUC for Navy programs tends to be overestimated, the difference between the two services is not statistically significant ( $p = 0.15$ ). In the case of System Type, however, the difference between the two major system types is significant ( $p < 0.01$ ). The mean AUC estimating error for Aviation systems is  $-0.088$  ( $SE = 0.065$ ), and the mean error for Maritime systems is  $0.289$  ( $SE = 0.081$ ).

Based on these contrasts, it is not clear whether it is better to attribute the observed bias to system type or to service component. By segregating the data further to control for these variables, we can gain additional insight. This is easily accomplished because the dataset consists of a large number of aviation systems split roughly equally between the two service components. Thus, in Figure 5, we control for Type of System by comparing the 12 Naval Aviation SubPrograms to the 13 Air Force Aviation SubPrograms. The results are still not significant ( $p = 0.10$ ).

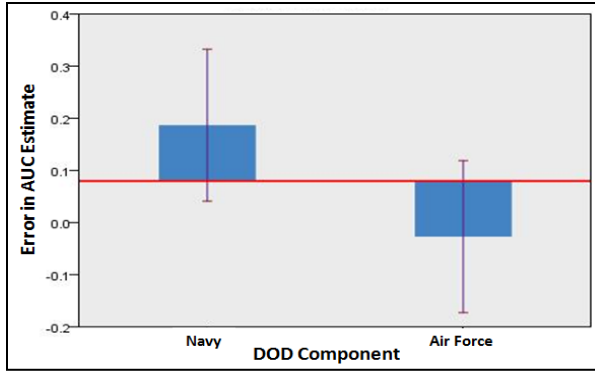


Figure 3. Mean Error in AUC Estimate by DoD Component

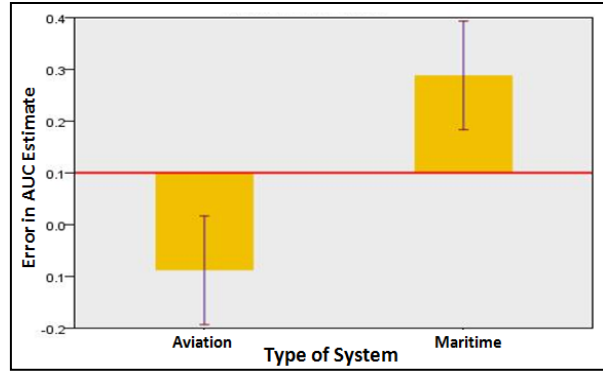


Figure 4. Mean Error in AUC Estimate by Type of System

We then control for DoD Component by comparing all (i.e., 12) of the Navy Aviation SubPrograms to all (i.e., 13) of the Maritime SubPrograms. Figure 6 shows the resulting deviation contrast, and the difference is now significant ( $p = 0.02$ ). The mean AUC error for Naval Aviation programs is 0.009 ( $SE = 0.089$ ) and the mean AUC error for Naval Maritime programs is 0.293 ( $SE = 0.078$ ). Taken together, these results indicate that while there may be a statistical difference in the accuracy of AUC estimating based on lead service component, the system type being fielded is clearly a more significant contributing factor.

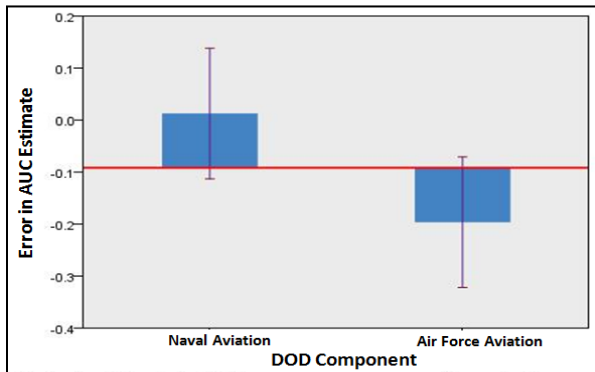


Figure 5. Mean Error in AUC Estimate by DoD Component (Aviation Only)

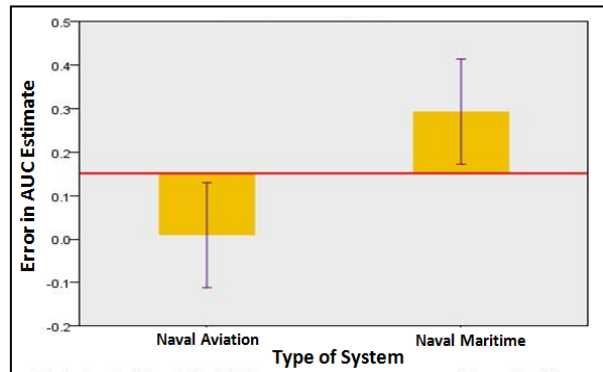


Figure 6. Mean Error in AUC Estimate by Type of System (Navy Only)

An important question is how the accuracy of these estimates change over time. Conventional thinking is that, as programs mature, the errors in AUC estimates are reduced. However, this notion is not supported by the results of this study. Figure 7 shows the mean error in AUC estimates as a function of six gradations of *Program Maturity* ranging from early development to late production. The dashed red line indicates a perfect estimate (i.e., zero percent error), and the magnitude of the vertical lines designate the range of the 95 percent confidence interval for the mean value.

The trend in Figure 7 suggests that estimates are best in the early and middle development phases, but worsen as the program matures. This is an erroneous inference, however. What is



happening in these earlier cases is that the extreme biases—both high and low—coincidentally average to a value close to zero. This fact can be seen more clearly in Figure 8, which plots the *absolute value* of the error in AUC estimates. Here, we do see a very slight downward trend over time (i.e., the magnitude of the error estimate is decreasing) as conjectured, but this change is not statistically significant.

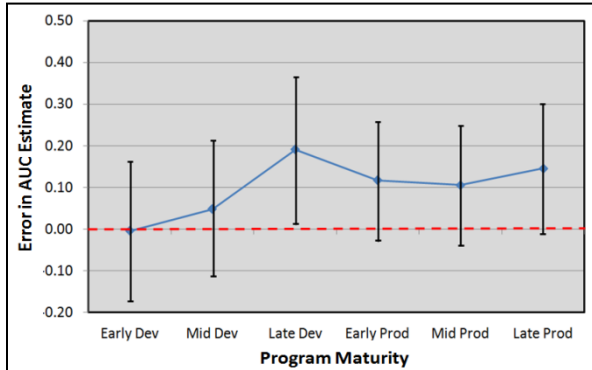


Figure 7. Mean Error in AUC Estimate by Program Maturity

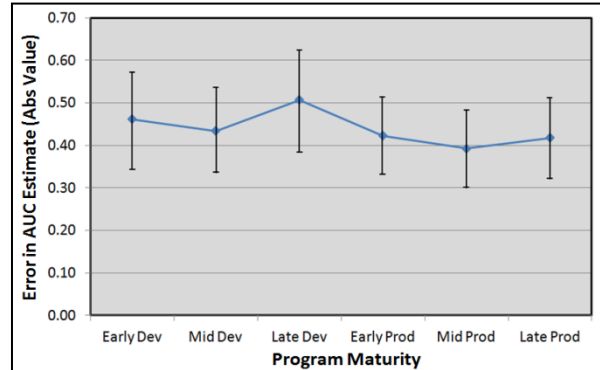


Figure 8. Error in *Absolute Value* of AUC Estimate by Program Maturity

Figure 8 is important for another reason, as it provides a stark insight into the state of current AUC estimating. In all previous plots, the magnitude of extreme biases for many SubPrograms was masked by the averaging function, thereby leaving the impression that estimating errors were actually not too bad in an aggregate sense. Figure 8 exposes this notion as false, indicating that the magnitude of the mean errors in AUC estimates are, on average, about 40 to 50 percent too high or too low throughout a program's life.

Figure 9 provides aggregated insight into the apparent deficiencies of DoD O&S cost estimates. On the y-axis, the AUC estimate error is plotted for every aviation and maritime system SAR for which an AUC estimate was provided and could be compared to valid actual costs (339 cases). The SAR year (i.e., the number of years since program initiation) is plotted on the x-axis. Clearly, there is a tremendous amount of dispersion in the AUC estimates. Of the 339 estimates in this figure, only 58 (17 percent) were within 15 percent of the actual AUC of the program, and only 123 (36 percent) were within 25 percent. Even allowing for an estimate error of 50 percent above or below the actual cost of the system, over one-quarter of the estimates (90 of 339) were still outside this range. The plot also shows the significant difference in estimating accuracy between aviation and maritime systems as well as the *slight* improvement as the programs mature.

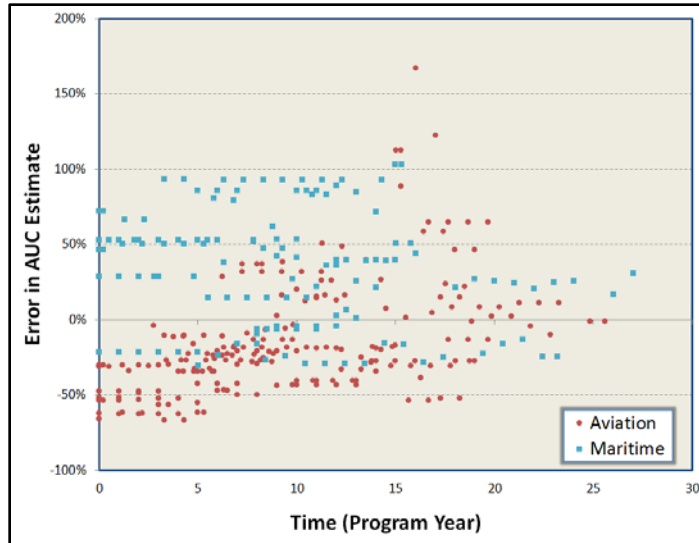


Figure 9. Mean AUC Estimate Errors as a Function of Time

### Characterizing the Accuracy of LCC Estimates

Next, we examine the Error in Life Cycle Cost estimates. Figure 10 shows the mean value for the Error in LCC Estimate as a deviation contrast between DoD Components. Clearly, the difference is statistically significant ( $p < 0.01$ ). The mean value for Air Force programs is  $-0.191$  ( $SE = 0.049$ ), and the mean value for Navy programs is  $0.055$  ( $SE = 0.037$ ). We see this same significant relationship in Figure 11 when we control for system type by only considering aviation systems ( $p = 0.05$ ). The mean LCC estimate error for Naval Aviation systems is  $0.056$  ( $SE = 0.066$ ); for Air Force Aviation systems, the mean LCC estimate error is  $-0.138$  ( $SE = 0.065$ ). Unlike AUC estimates, however, in the case of LCC, we find no significant relationship between system type and the accuracy of the estimates.

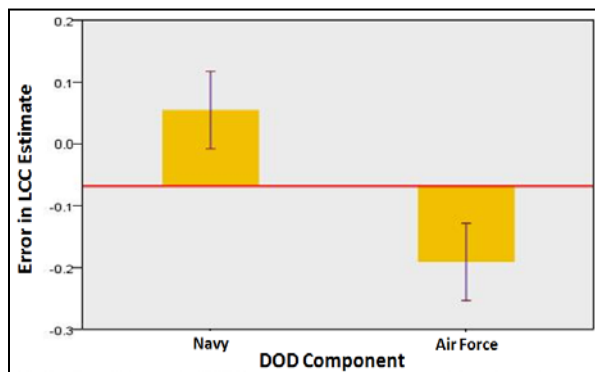


Figure 10. Mean Error in LCC Estimate by DoD Component

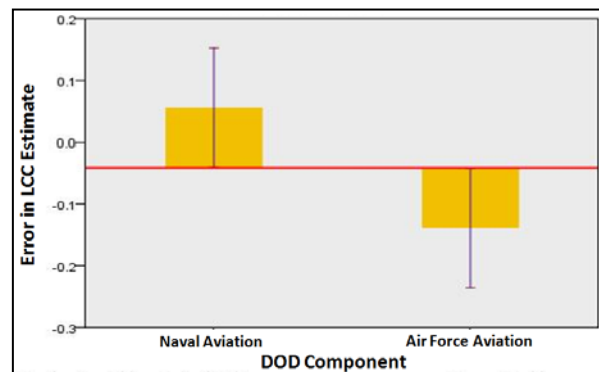


Figure 11. Mean Error in LCC Estimate by DoD Component (Aviation Only)

Now we examine LCC estimates over time. Figure 12 shows the mean Error in LCC Estimates as a function of Program Maturity. This plot suggests that LCC estimates are biased low

during early and middle development, but that estimates created during the production phase are, on average, very good. This is a false impression, however, and once again examining the absolute value of these errors provides a clearer picture. From Figure 13, we see that LCC estimate errors remain sizeable even into the production phase: the magnitude of the errors tends to be around 30 percent during the development phase, whereas they are “only” about 20 percent during the production phase. Although this represents an improvement in estimating accuracy relative to the AUC estimates, this is still a substantial error, especially considering that the median final LCC estimate in this study is nearly \$19.0 billion.

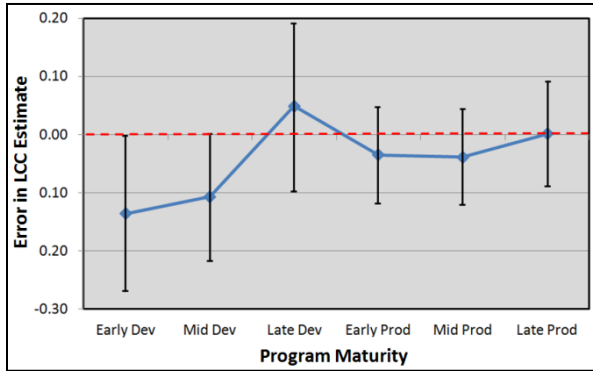


Figure 12. Error in LCC Estimate by Program Maturity

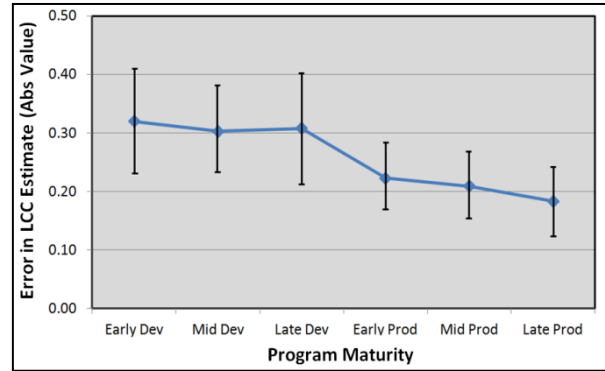


Figure 13. Error in *Absolute Value* of LCC Estimate by Program Maturity

Finally, we consider the LCC estimate errors in aggregate. Figure 14 is much like Figure 9 in that mean estimate errors are plotted on the y-axis (this time for all system types), and the x-axis is the SAR year. Here, however, the grouping variable is DoD Component vice System Type. Of the 333 LCC estimates in this figure, 148 (44.4 percent) were within 15 percent of the actual LCC of the program, and 204 (61.3 percent) were within 25 percent. With respect to the worst estimates, only 6.6 percent (22 of 333) of all estimate errors were not within 50 percent of the actual cost of the system. Based on these results, the error in LCC estimates is clearly not as poor as what we observed for AUC estimates, although the magnitude of the errors is still considerable.

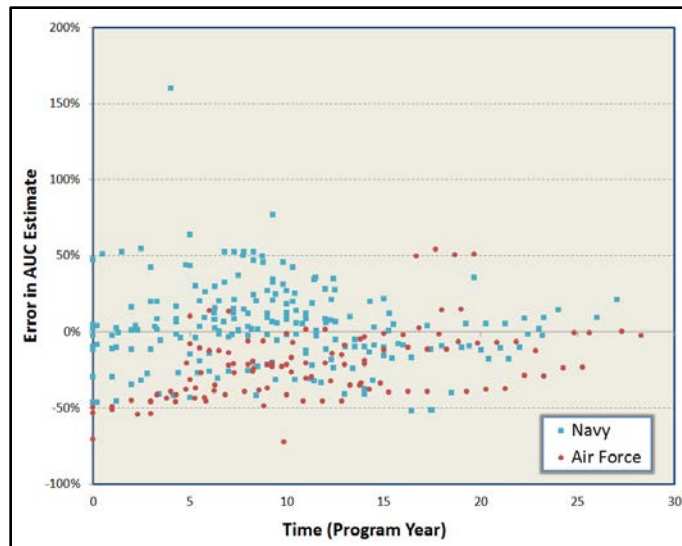


Figure 14. Mean LCC Estimate Errors as a Function of Time

## DISCUSSION

### Key Findings

Regarding the Annual Unit O&S Cost (AUC) estimate, the results show that the type of system being procured significantly influences its accuracy. On average, AUC estimates for aviation programs tend to be too low, while AUC estimates for maritime programs tend to be too high. This disparity persisted even after controlling for DoD service component. Although service component may be a contributing factor in the accuracy of the AUC estimates, it was not shown to be statistically significant.

With respect to the accuracy of Life Cycle Cost (LCC) estimates, however, the DoD service component was a significant factor. LCC estimates developed by the Air Force were, on average, significantly lower (and negatively biased) than the LCC estimates produced by the Navy. Further, by comparing Naval Aviation systems to Air Force Aviation systems, we showed that this relationship is independent of the type of system being procured.

The preceding biases are worthy of note and, assuming they are valid, certainly warrant further investigation into potential causal factors. However, the more remarkable finding in this study is *how poor both the AUC and LCC estimates are on average, and how little they improve over time*. The accuracy of the AUC estimates is particularly poor. Not only are the magnitudes of the mean errors nearly 50 percent throughout the development phase, but there is virtually no improvement in the accuracy of these estimates as the program matures, even late into production after a majority of the units have been fielded. This observation may be surprising to many, but is consistent with the findings of a recent GAO report which notes that many defense programs are failing to take advantage of actual O&S cost records to update their “going-forward” O&S cost estimates (GAO 2010).

Relative to AUC estimates, LCC estimate behavior over time is certainly better, but by no means impressive. The mean magnitude of LCC estimate errors during the development phase is closer to 30 percent vice the 50 percent we observe for AUC estimate errors. Moreover, there appears to be some marginal improvement in the accuracy of LCC estimates during the produc-

tion phase. Unfortunately, this relatively good performance is likely specious, an artifact of the accounting methodology employed in this study relative to acquisition costs. By the time that a program has entered the production phase, the total acquisition cost tends to be relatively stable and fairly precisely known. Although acquisition costs typically do not constitute the majority of life cycle costs, they do generally comprise about 25 to 40 percent of a program's life cycle cost (DoD 2009), thereby serving to substantially reduce the overall range of potential variability in the LCC figure. This effectively dampens the variability in the LCC estimate errors, which was borne out by a separate analysis (not provided here due to space considerations) which excludes acquisition costs. Examining the error in the estimates of Total O&S cost only, we find error magnitudes at least as great as those for AUC estimates.

If it is true that Pentagon estimates related to O&S costs are highly inaccurate and do not markedly improve over time, there are some potentially important implications to DoD life cycle management. The most obvious impact is budgetary. Overestimating cost liability creates opportunity loss by precluding investment in additional capabilities; underestimating cost liability adversely affects the subject program and/or other programs from which budget shortfalls must be reconciled. And while inaccurate and volatile cost projections can greatly complicate the budgeting process, this is typically less problematic if these deficiencies are known, or at least not unexpected. The greater potential concern here is that many decision-makers may be under the (apparently) false impression that the AUC and LCC estimates they are seeing late in the program are substantially more accurate than earlier estimates. This would certainly be a reasonable assumption given the traditional behavior of acquisition costs. However, based on the results of this study, it appears that O&S cost estimates behave fundamentally differently from acquisition cost estimates in that they simply *do not converge on an accurate value within the time spans of consideration*. In other words, DoD leadership should be wary of relying on the accuracy of O&S-related cost estimates—particularly AUC estimates—even if the estimates are being provided when the program is well into production and ample O&S cost actuals are available.

As a short digression, the notion that the most recent O&S estimate is bound to be an accurate (or at least, the most accurate) estimate has been an explicit or tacit assumption in all the previous studies the authors found which attempted to characterize O&S costs for DoD systems (Ryan et al. 2012). In these studies, the final O&S estimate was used—in either all, or the majority of cases—as a proxy for actual costs. The findings of this study indicate that such an approach is flawed. Any study that seeks to truly characterize the accuracy, absolute or relative, of O&S cost estimates must consider O&S cost *actuals*.

While projected O&S cost figures play an important role in budget formulation, often times the more valuable—and perhaps more valid—use of O&S cost estimates is as a decision management tool. The AUC metric is frequently used to differentiate design alternatives within a given program, and the LCC metric is a standard discriminator between competing programs or as a means of comparing the cost-effectiveness of modifying an existing system vs. initiating a new acquisition. For each of these types of comparative analyses, the *absolute* accuracy of an estimate is clearly less important than its *relative* accuracy.

However, there is nothing in this study that would allow one to reasonably suppose that the relative accuracy of these cost estimates is any more reliable than their absolute accuracy. Even for the same service component, or the same system type, we see tremendous within-subject and temporal variability of estimate accuracy. Furthermore, our analysis found that several other program parameters (the results of which were not detailed in this paper due to space considera-

tions) also played a significant role in the accuracy of AUC and LCC estimates. These other parameters included the nature and type of the cost variance to date, official program cost estimate history, and procurement change trends. In essence, AUC and LCC estimates are not only characterized by high levels of inaccuracy and variance, but they appear to also exhibit *random* inaccuracies and variances as well. Clearly, this calls into question the validity of value decisions substantively based on AUC and LCC estimates.

## Methodological Concerns

There are a number of concerns regarding the underlying premise of this study and how valid it is to make comparisons between O&S cost estimates and actuals. For the most part, these concerns are addressed in the associated Methodology Paper (Ryan et al. 2012), but a couple of points bear repeating and/or emphasis.

The overall validity of the results of this study is highly dependent on whether the O&S actual costs obtained from the VAMOSOC systems are valid. The authors have worked diligently to account for validity concerns related to the VAMOSOC systems, but there is an obvious potential source of *systemic* error that should be noted. Only one program in this study has reached its end of life; thus, there is only one instance in which we may *directly* calculate actual system life cycle costs or the true annualized unit O&S cost over the life of the system. In all other cases, we must infer these values based on the average annual unitized O&S costs seen to date (directly in the case of the AUC, and prorated for the remainder of the assumed operational service life in the case of the LCC). This is a valid point of concern given that the average operational maturity (i.e., the proportion of the planned operational service life observed thus far) of the systems in this study is only 34 percent.

In addition, there is evidence that for at least certain types of weapons systems, the O&S costs do tend to increase as the system ages (CBO 2001) (and the data in this study provides some support to this notion). Consequently, a simple proration across the remaining years of the program could be expected to yield a negative bias in the AUC and LCC actual cost figures. It is important to recognize, however, that even if the actual costs used in this study are biased in this manner, the fundamental data trends we observe would likely be unchanged. Although mean errors for specific programs would change, we would expect the shape of the error distribution to remain the same, such that improvements in the estimate accuracies for one program would likely be offset by diminished accuracy in another program. In other words, even if the O&S cost actuals were shown to be biased, as long as they were *consistently* biased, then it is likely that Figure 8 and Figure 13 would be unaffected, and that the data points in Figure 9 and Figure 14 would simply be shifted down, en masse.

Another potential bias is the temporal distribution of cost actuals. Given the phasing of program development and fielding, along with the steadily improved robustness of the VAMOSOC systems, it is reasonable to expect that the accrual of actual costs will be skewed toward more recent dates. This is, in fact, the case, as 80 percent of actuals in this study are from 2001-2010. Since the United States has been at war during that entire time, one might suspect that the O&S cost estimates for these systems (which have all assumed peacetime operations) may be systemically low. This may be a valid concern, but could not be substantiated from the data in this study. For the 15 programs in which O&S cost actuals spanned 2001, we found no significant difference in the mean costs accrued before and after this date.

The other principal objection to a study of this nature is that it is invalid to compare O&S cost estimates to cost actuals. The basic rationale for this position is that the assumptions used to

construct the estimates are often fundamentally different from what occurred in reality, and thus such an endeavor lacks utility or relevance. These *mismatched assumptions* may either be deliberate (e.g., assuming peacetime operations during a time of war) or the natural consequence of the uncertain impacts and outcomes bound to be encountered in the long life of a weapon system (e.g., fuel prices, healthcare costs, etc.). In either case, the argument is rooted in the notion that the same standard of accountability applied to acquisition cost estimates should not apply to O&S-based cost estimates.

We offer two rhetorical responses to this objection. One is to note that the acquisition phase of a system is also extremely lengthy and characterized by myriad sources of uncertainty: why would we insist on accuracy and accountability for cost estimates that pertain to the acquisition phase, but not to cost estimates that pertain to the sustainment phase? The second response to this objection is to simply call into question the purpose and value of developing O&S cost estimates at all. Is the purpose of an O&S cost estimate only to be accurate relative to a static baseline amid fixed constraints, or do we want the estimate to be accurate in the real world, which is indelibly infused by uncertainty? Regarding the value of these estimates, if we do not believe there is a way to reasonably ascertain the accuracy of O&S cost estimates, why should anyone be inclined to trust them? More bluntly, if we do not expect our O&S estimates to be a valid indicator of the actual costs eventually incurred, *why do we expend the effort of constructing the estimates and basing budgetary and programmatic decisions upon them?*

## **Future Work**

There are a number of ways in which this study can be improved or expanded in order to improve its reliability or increase its applicability. Relative to the discussion above regarding potential bias in the actual O&S costs, we may seek to apply factors that account for increases in actual AUC as a system ages. We also need to consider potential nonlinearities and nonparametric distributions in the data to determine whether more complicated models are required to accurately represent the data.

What we believe to be the most important future effort related to this study is already underway. Now that we have evidence to indicate that there are major deficiencies in the accuracy of O&S cost estimates, the obvious challenge becomes how to improve these estimates. We have analyzed the data along approximately 200 dimensions and found a number of significant independent variables that suggest there is an underlying distribution to the errors being observed in the AUC and LCC estimates. This presents the enticing possibility of creating a stochastically-derived mixed model to predict these estimate errors at any point in a program's life. And if we can predict the estimate errors, we can apply correction factors to improve their accuracy.

## **CONCLUSION**

Historically, the cost to acquire a system has been the principal focus of the defense acquisition community. In recent years, however, there has been a growing emphasis on the importance of fully accounting for a system's other—and generally greater—cost liability: Operating and Support (O&S) costs. As a result, estimates of O&S costs, and the over-arching life cycle costs, are produced and scrutinized more frequently and with greater rigor than ever before. In fact, a system's Annualized Unit O&S Cost (AUC) and its Life Cycle Cost (LCC) have both become key metrics for not just budgetary decisions, but also as value indicators for discriminating among system designs or between competing programs.

Yet despite this increasing reliance on AUC and LCC estimates, there is an inexplicable lack of knowledge regarding their reliability and accuracy. There have only been a handful of published studies on this topic, all within the last couple of years, and each limited in scope and/or employing inappropriate methodologies, not capable of truly assessing the accuracy of O&S-based cost estimates. Given the magnitude of the decisions being made based on these estimates, and the massive dollars at stake, the authors believe a characterization study like this one is both crucial and timely.

As discussed in this paper as well as the associated Methodology Paper, there are a number of methodological concerns associated with attempting to characterize O&S costs. Nevertheless, the results of this study are intriguing, and in some respects, transcend the methodological concerns. Of note, several program factors were found to have a statistically significant effect on the magnitude and direction of LCC and AUC error estimates. It was also found, rather surprisingly, that the accuracy of O&S-based cost estimates improve little, if at all, over time, suggesting that these estimates are destined not to converge with the actual cost. These results should give pause to DoD leaders that rely on LCC or AUC estimates, as they cast significant doubt on the validity of establishing budgets or making decisions based on these figures.

On a more positive note, even though there appear to be significant errors in the accuracy of the LCC and AUC estimates to date, the nature of the results suggest there may be an opportunity to improve these estimates going forward via a macro-level stochastic cost modeling approach. This effort will be the topic of our next paper.

### **ACKNOWLEDGEMENTS**

The authors wish to acknowledge the invaluable assistance provided by the DAMIR (Defense Acquisition Management Information Retrieval) group in the acquisition and data mining of the SARs used in this study. We also wish to express our appreciation to the tremendous support provided by the AFTOC (Air Force Total Ownership Cost) and Naval VAMOSOC teams in providing and interpreting the VAMOSOC data.

This research is sponsored by the Office of the Secretary of Defense.



## REFERENCES

- Balaban, H., Devers, W., & Roark, L. (2010). "Feasibility and Advisability of Baselines for O&S Costs: C-17 Case Study." Institute for Defense Analyses. Alexandria, VA.
- CBO. (2001). "The Effects of Aging on the Costs of Operating and Maintaining Military Equipment." Congressional Budget Office. Washington, D.C.
- Choi, J., Alper, O., Gessner, M., Jondrow, J., Koopman, M., Groo, E. (2009). "O&S Cost Forecasting: Issues and Evidence." Center for Naval Analyses. Alexandria, VA.
- DAU 2012. *Defense Acquisition Guidebook*. Defense Acquisition University. [Online]. Available at <https://dag.dau.mil/Pages/Default.aspx>. [Retrieved February 2012].
- Diggle, P., Liang, K., & Zeger, S. (1994). *Analysis of Longitudinal Data*. Oxford: Clarendon Press.
- DoD. (1992). *5000.04-M: Cost Analysis Guidance and Procedures*. Department of Defense. Washington, D.C.
- DoD. (2009). "Weapon System Acquisition Reform Product Support Assessment." Under Secretary of Defense for Acquisition, Technology and Logistics.
- DoD CAIG. (2007). "Operating & Support Cost Estimating Guide." Office of the Secretary of Defense, Cost Analysis Improvement Group.
- GAO. (2010). "DoD Needs Better Information and Guidance to More Effectively Manage and Reduce Operating and Support Costs of Major Weapon Systems." Government Accountability Office. Washington, DC.
- OSD/CAPE. (2011). "Annual Report on Cost Assessment Activities."
- Ryan, E., Jacques, D., Colombi, J., & Schubert, C. (2012). "A Proposed Methodology to Characterize the Accuracy of Life Cycle Costs for DoD Programs." Conference on Systems Engineering Research (CSER). St. Louis, MO.
- Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.