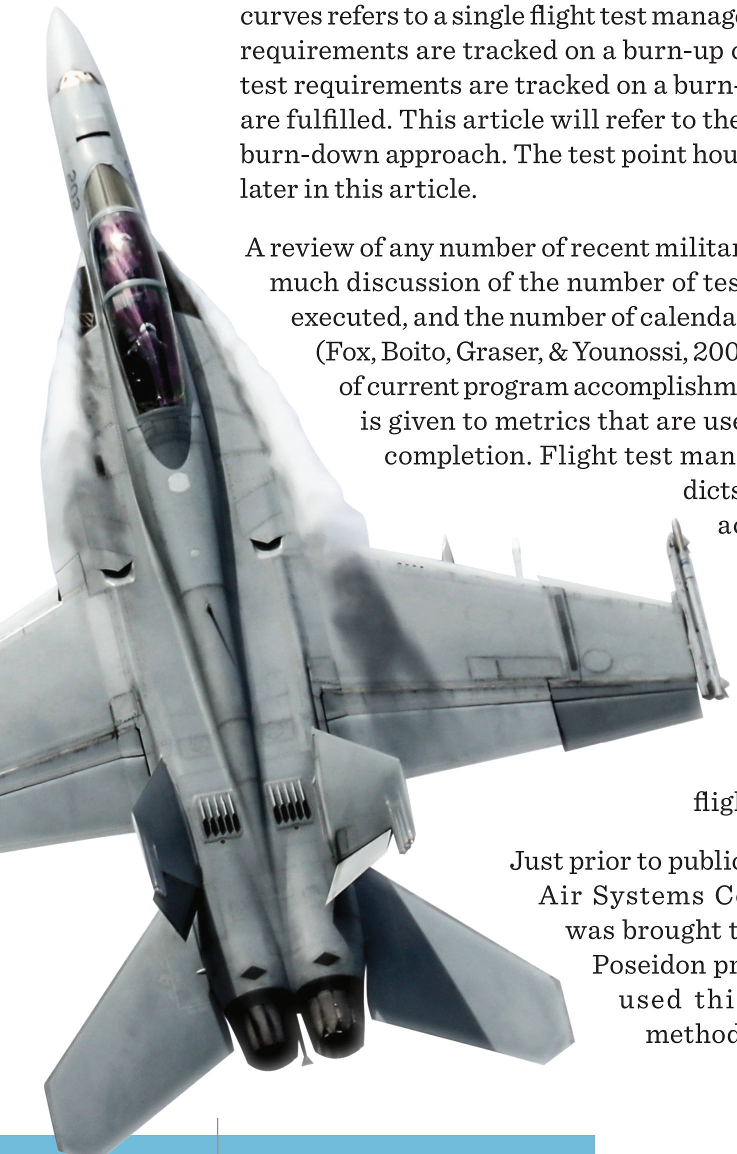
AN EARNED VALUE MANAGEMENT APPROACH to TIME-WEIGHTED Test Point Program Management

Lt Col C. Matt Ryan, USAF

Flight test programs have long struggled with finding metrics that are predictive in nature, which would allow program managers to make decisions about how to sequence test events and what to do about program delays. Actionable data have been difficult to find because many test programs have relegated themselves to metrics that are often meaningless—test point burn-down. Test program delays have traditionally been handled by cutting test points and endless revisions to the test schedule that never reflect an accurate finish date. The test program manager's challenge is to find the right test points to eliminate or defer, and to understand the effect on the overall progression of the test program. This article proposes a better approach to test management—the use of time-weighted test points combined with earned value management and earned schedule methods to provide predictive information to test program managers.

DOI: *https://doi.org/10.22594/dau.17-790.25.02*
Keywords: *Earned Schedule, Flight Test, Test Point Hour, Time On Conditions, Predictive Test Metrics*

The flight test community has long agreed that not all test points are created equal, yet programs continue to be managed with test point burn-up and burn-down curves. A review of two of the largest flight test programs in history—the F-22 Engineering and Manufacturing Development (EMD) and F-35 System Development and Demonstration (SDD) test programs—shows that both used test point burn-up and burn-down curves. Additionally, the F-35 test program introduced a useful approach to provide more actionable data to test managers. The B-2 EMD program introduced a unique metric—the test point hour—but continued to use traditional burn-down and burn-up curves to manage the program. The use of test point burn-down and burn-up curves refers to a single flight test management approach. All the added test requirements are tracked on a burn-up curve. All the baseline or planned test requirements are tracked on a burn-down curve as test requirements are fulfilled. This article will refer to the use of these curves as simply the burn-down approach. The test point hour (TPH) will be further explained later in this article.



A review of any number of recent military flight test programs would offer much discussion of the number of test points, sorties, and flight hours executed, and the number of calendar months to complete the program (Fox, Boito, Graser, & Younossi, 2004). These metrics are all indicative of current program accomplishments. However, very little attention is given to metrics that are useful as predictors of test program completion. Flight test managers need an approach that predicts program completion and provides actionable data for management decisions throughout execution. Combining a new metric—the time-weighted test point—with concepts used in earned value management (EVM) and earned schedule (ES) approaches may provide the data needed for better flight test management decisions.

Just prior to publication, an approach used by Naval Air Systems Command (NAVAIR) circa 2009 was brought to the author's attention. The P-8 Poseidon program and a few other programs used this approach. The management method used a concept called test point mass (TPM) and a process model to measure flight test program progress and predict when a program would finish. We will briefly discuss the NAVAIR methodology later in this article.

Background

Before we examine our three example flight test programs, we will briefly review the metrics of several modern military fixed-wing aircraft development test programs.

From the data in Tables 1 and 2, we can make some broad observations. First, the most recent flight test programs, some of which we will examine shortly, suffered significant delays, overly optimistic schedule estimates, and underestimated work scope. The only exception was the F-18E/F, which was a derivative program. Second, the scope of flight test programs is generally increasing over time. Again, the exception is the F-18E/F flight test program. While not shown in the tabular data (Tables 1 and 2), one other observation can be made from a review of the most recent military flight test programs. Recent military flight test programs have labored under significant instability in the schedule and planned scope (flight hours, test points, etc.) (General Accounting Office [GAO], 1995, 1996, 1998a, 1998b, 1999, 2000b, 2001; Government Accountability Office [GAO], 2008, 2011a, 2011b). These observations point out the need for a better flight test management approach.

Why Not Test Points?

Let's examine some of these test programs to point out specific weaknesses of past approaches and to highlight useful concepts that will be applied to the proposed methodology. The F-22 flight test program used a test point burn-down management approach. The program tracked the efficiency and effectiveness of fulfilling test point requirements and then applied these factors to determine the number of sorties needed. The F-22 program used test operations per hour as a measure of efficiency, where test operations included test information sheet (TIS) points flown, TIS points re-flown, and test point burn-up. The TIS points were accepted formal test requirements (test points) that were part of the program baseline. Test points re-flown were differentiated from test points repeated. A repeat occurred real-time when test conditions were not met, while a test point was re-flown on a later sortie because post-mission analysis showed that the test requirement had not been met for any number of reasons. The test point burn-up included added test requirements that were not part of the program's baseline, test

TABLE 1. HISTORICAL FIXED-WING FLIGHT TEST PROGRAM DATES

Program	Program Start[a]	First Flight (FF) Date		End of Initial Operational Test & Evaluation (IOT&E)	
		Planned[b]	Actual	Planned[b]	Actual
F-15A/B	Dec 1969[6]		Jul 1972[6]		
A-10	Jan 1973[1]	Dec 1974[1]	Feb 1975[1]		Mar 1976[6]
F-16A/B	Apr 1975[2]	Jan 1977[12]	Dec 1976[6]	Dec 1978[12]	
F-18A/B	Dec 1975[1,3]	Jul 1978[1,3]	Nov 1978[1,6]	Dec 1981[1]	Nov 1982[1,20]
F-117	Nov 1978[4]	Jul 1980[4]	Jun 1981[13]	Nov 1982[13,c]	
B-1A	Jun 1970[5]	Apr 1974[5]	Dec 1974[6]		
B-1B	Jan 1982[6]	Apr 1983[14]	Oct 1984[15]	Jun 1986[14]	
C-17	Nov 1987[7]	Jun 1991[16]	Sep 1991[7,8]		Jun 1995[7, 8]
B-2	Nov 1981[8]	Dec 1987[17]	Jul 1989[8]	Dec 1993[21]	Mar 1998[8]
F-22	Jun 1991[9]	Sep 1995[18]	Sep 1997[9]	Sep 1999[18]	Dec 2004[9]
F-18E/F	May 1992[10]	Oct 1995[10]	Nov 1995[8,23]	Feb 2000[10]	Nov 1999[8]
F-35	Oct 2001[11]			Mar 2012[19]	Feb 2019[11,d]
F-35A		Nov 2005[19]	Dec 2006[11]		
F-35B		Apr 2006[19]	Jun 2008[11]		
F-35C		Jan 2007[19]	May 2010[11]		

Note. Adapted from (Drezner & Smith, 1990)[1]; (Tyson, Harmon, & Utech, 1994)[2]; (GAO, 1979)[3]; (Smith, Shulman, & Leonard, 1996)[4]; (GAO, 1973)[5]; (Rothman, 1987)[6]; (DoD, 2010a)[7]; (Fox, Boito, Graser, & Younossi, 2004)[8]; (DoD, 2010b)[9]; (DoD, 1992)[10]; (DoD, 2016)[11]; (Acker, 1983)[12]; (R. Moseley, personal communication, 1983)[13]; (GAO, 1983)[14]; (Boeing, n.d.)[15]; (DoD, 1990)[16]; (GAO, 1990b)[17]; (DoD, 1991)[18]; (DoD, 2001)[19]; (GAO, 1985)[20]; (420th, 1997)[21]; (GAO, 1989)[22]; (DoD, 2012)[23].

TABLE 2. HISTORICAL FIXED-WING FLIGHT TEST PROGRAM METRICS

Program	Flight Test Program Duration (months)		Sorties	
	Planned[a]	Actual[b]	Planned	Actual
F-15A/B		60		
A-10		14		
F-16A/B		43	240[1]	
F-18A/B	38	49		
F-117	22	29	850[2]	496[2]
B-1A				
B-1B		53		
C-17	22	46		1,134[4]
B-2		105		1,013[4]
F-22	66	88		3,496[5]
F-18E/F		49		3,141[4]
F-35[e]			6,979[3]	9,201[11]
F-35A	64	147		
F-35B	46	129		
F-35C	23	106		

Note. (GAO, 1977)[1]; (R. Moseley, personal communication, 1983)[2]; (GAO, 2008)[3]; (Fox, Boito, Graser, & Younossi, 2004)[4]; (Hehs & Rhodes, 2012)[5]; (GAO, 1990a)[6]; (GAO, 1999)[7]; (GAO, 2000a)[8]; (GAO, 2001)[9]; (GAO, 2006)[10]; (Lockheed Martin, personal communication, December 2017)[11].

[a]Planned duration as of the program's first flight date: F-18A/B (Drezner & Smith 1990); F-117 (R. Moseley, personal communication, 1983); C-17 (DoD, 1990); F-22 (DoD, 1997); F-35 (DoD, 2007). [b]Derived from the data in Table 1. [c]Accomplished 2,291 test point hours (420th Flight Test Squadron, 1997). [d]Flight test plan included 2,720 test point hours as of 1995; testers planned

Initial Operational Capability (IOC)		Schedule Overrun (Months)		
Planned[b]	Actual	FF	IOT&E	IOC
Jul 1975[2]	Sep 1975[6]			2
Jun 1977[1]	Oct 1977[1]	2		4
	Oct 1980[12]	-1		
Sep 1982[1,3]	Mar 1983[1,2]	4	11	6
Jun 1982[13,c]	Oct 1983[13]	11		16
		8		
	Sep 1986[22]	18		
	Jan 1995[7]	3		
		19	51	
Sep 2003[18]	Dec 2005[9]	24	63	27
Sep 2000[10]	Sep 2001[23]	1	-3	12
			83	
Jun 2011[19]	Aug 2016[11]	13		62
Apr 2010[19]	Jul 2015[11]	26		63
Apr 2012[19,e]	Aug 2018[11]	40		76

[a]Date of Milestone decision (Milestone II or B) that started full program development. [b]Planned dates from Milestone II/B decision. [c]F-117 flight test program plan as of February 1980 (more than 1 year after full-scale development began). [d]Current planned date of F-35 IOT&E completion. [e]Current planned date of F-35C IOC.

Flight Hours		Test Points	
Planned	Actual	Planned	Actual
	4,096		
1,350[2]	721[2]		
1,105	~1,900		
1,000			
2,277[4]			5,623[4]
3,600[6,d]	5,197[4]		19,897[4,c]
~4,000[7,8,9]	7,616[5]		>29,500[5]
	4,620[4]		>15,000[4]
>12,000[10]	17,054	5,792[11,f]	19,325[11,f,h]
		13,774[11,g]	13,830[11,h]
		18,953[11,g]	18,536[11,h]
		13,762[11,g]	13,800[11,h]

to accomplish testing in 4,400 flight hours (GAO, 1995, 1996). [e]Data in this row apply to all F-35 variants. [f]These are mission systems test points, which are applicable to all variants. Test points listed by variant are flight sciences test points. [g]Planned test points are from the test program baseline as of January 1, 2010 (approximately 3 years after first flight of AA-1). [h]Actual test points are as of December 31, 2017.

point repeats, regression test points added due to hardware and software updates, and real-time build-up test points that were added during a test sortie when flight envelope boundaries were being approached. Flight test efficiency was a measure of how much flight time was spent in pursuing TIS points. The more test operations accomplished per flight hour, the more efficient the sortie (J. Pieper, personal communication, 2005).

Test effectiveness was measured by TIS points closed whether through execution of the TIS point or due to analysis of other executed TIS points. A sortie was more effective if it led to a large number of TIS points closed. Using TIS points closed per flight and assuming consistent sortie durations, an estimate of how many sorties were needed to complete the program could be calculated. Finally, a ground abort rate was applied to determine the number of missions that needed to be scheduled (J. Pieper, personal communication, 2005).

"This article proposes the combination of the TWTP concept with the methods of EVM and ES to provide the actionable information that flight test program managers need.

The difficulty introduced by the test point burn-down method is that the scope of work remaining to complete the flight test program is based on averages. The F-22 flight test program did not differentiate between the scope of work required to complete different types of test points. For example, flutter test points, performance trim shots, throttle bodies, weapons drops and sensor performance verification test points all require a different amount of time to execute. The program broke down separate execution metrics for flight sciences and mission systems testing, but this was the depth of granularity used to determine work remaining (J. Pieper, personal communication, 2005).

The data in Tables 1 and 2 reveal the challenges of the F-22 flight test manager. First flight was delayed by 2 years and the end of initial operational test and evaluation (IOT&E) was delayed by 5 years. Using the number of flight hours as a measure of work scope, the flight test program more than doubled. Not all of the problems encountered by the test program could have been remedied with any management techniques. Most prominent were the repeated delays in producing test aircraft for the test program (GAO, 2002). In addition to delayed aircraft availability, the program also suffered from less productivity on each test flight than was planned (GAO, 2000b, 2001, 2002). The test program was re-baselined on multiple occasions after first flight, but program test managers were unable to accurately forecast the scope of test or the test schedule remaining (GAO, 2002; J. Pieper, personal communication, 2005). For example, the June 2001 re-plan effort assumed a December 2003 completion date for IOT&E. With more than 3 years of flight test execution performance to leverage, the program test managers continued to plan for a program that would see an additional 12-month delay to completion. This re-plan also reduced the number of test points remaining by 31%, which proved inaccurate in the years that followed as more than 4,000 test points were added during the years after the 2001 re-plan (GAO, 2002; J. Pieper, personal communication, 2005).

Useful Concepts That Can Be Built Upon

The F-35 program also used a test point burn-down management approach, which will not be discussed in detail since it is very similar to the approach already described. What is of interest for our discussion was the categorization of test points by capability. (The F-22 EMD program tracked test points by subsystem in the pursuit of verifying specification compliance.) In addition to using a test point burn-down management approach for the aggregate test program, test requirements were also tracked as smaller groups against particular capabilities. For example, test points associated with weapons were tracked specifically for that capability area in addition to the aggregate test point burn-down. The program applied the test point burn-down method to 13 different capability areas. Within each capability area, the same methodology was applied to another set of subgroups. For example, radar test points were tracked as a subgroup to one of the capability areas (Joint Strike Fighter Program Office, personal communication, June 2013; Lockheed Martin, personal communication, December 2017).

The added depth of this accounting approach provided useful information to the test program manager to more easily recognize which capabilities were in jeopardy of delayed delivery and some insight as to why. However, this additional feature does not address the fundamental flaw with the test point burn-down method that all test points are not equal in airborne work scope. So the flight test manager still does not have a proper metric to forecast the time required for capability or program completion.

While the B-2 EMD flight test program also used a burn-down management method, the B-2 test program introduced a new metric in addition to the test point. Instead of focusing solely on tracking test points, the B-2 test program tracked test point hours (TPH) using the burn-down method (420th Flight

Test Squadron, 1997). The TPH metric used time-weighted test points (TWTP) to track the time required to be on conditions (altitude, airspeed, vehicle configuration, etc.) for test point completion, which provided a much better measure of work scope for test program managers (420th Flight Test Squadron, 1997; GAO, 1995, 1996). The TWTP method assigns a specified number of minutes to each test point that are required for its execution (start run to stop run). These times are based on test experience and engineering knowledge of how much data need to be collected to verify a function or capability. (The TPH does not include the flight hours spent getting to an "on conditions" state, but this overhead can be accounted for separately and is generally easier to quantify by using averages [GAO, 1996].)

Still unclear is why this new (at the time) metric was not adopted by other programs that followed (like the F-22 and F-35), but there are at least two plausible reasons. First, many aspects of the B-2 EMD test program were highly classified. Next, the prime contractor for the B-2 weapons system is different than the prime contractor for the F-22 and F-35. Because the prime contractor for F-22 and F-35 had total system performance responsibility for those development programs, it is likely that the flight test manager applied familiar approaches from past programs.

Regardless, the B-2 EMD flight test program did not combine the TPH concept with an appropriate management methodology that provided predictive and actionable information to the program manager.

Determining the Value of a Test Point

This article proposes the combination of the TWTP concept with the methods of EVM and ES to provide the actionable information that flight test program managers need.

The TWTP provides a relevant metric for test program managers because a significant item of interest for all test programs is maintaining schedule. Extending the test schedule contributes directly to cost increases and delays delivery of capability to combat forces. Fundamentally, two parameters drive test program schedules: test capacity and system maturity. Test capacity determines how much testing can be accomplished over any period of calendar time. Flight test capacity is most often measured in air vehicle months, or months when a properly configured test vehicle is available for test. System maturity determines when testing will be ready for execution, and determines the order of test points available for execution. An excess of flight test capacity will not make up for a lack of system maturity. The proposed method does not address system maturity, but it does allow a manager to understand how much flight test capacity is needed to meet schedule, and it can provide insight into critical capabilities, which can be affected by system immaturity. In short, it provides the test manager a predictive tool.

The first implication of the TWTP is that we have now solved the problem of managing disparate test points that are not equal in "earned value." Test points that require 10 minutes to execute show up as much more significant to program execution metrics than test points that require 1 minute to execute. As with TPH, the TWTP earned value metric still does not account for the time required to get to an "on conditions" state. The overhead spent "off conditions" during each sortie can be accounted for by using an efficiency factor that is generally consistent for a particular flight test program and can be applied as an aggregate with reasonable accuracy. The many enablers of an "on conditions" state (e.g., support aircraft, range time, and other range resources) are important considerations and cannot be overlooked in the test program plan. However, these items do not indicate program progress or performance against a schedule, and are therefore not accounted for in the TWTP metric.

The TWTP defines the currency of the flight test manager in test time. Tracking TWTP burn-down can still provide the same insight as test points tracking. For example, TWTP executed ($TWTP_E$) per flight hour provides a measure of efficiency, and TWTP closed ($TWTP_C$) per flight hour provides a measure of effectiveness. It is important to distinguish between TWTP executed and TWTP closed. When any particular test point is executed during a mission, then the budgeted time assigned through the time-weighting process for each test point has been spent. If the same test point is repeated once, then the actual cost is twice the budgeted cost of the TWTP. A test point is only closed after the resulting data have been analyzed and determined to meet the data requirement or a test point is determined to be no longer required.

Simply tracking a burn-down of TWTP will not allow the program manager to make decisions to trade off capability delivery and test capacity development (test capacity is not free) to maintain overall test program schedule. These predictive data are the field of EVM and ES. To apply these management methods, we need one more TWTP concept—TWTP planned ($TWTP_P$).

Most programs will have thousands of minutes of TWTP that need to be accomplished. The TWTP can only be executed if test points are available for execution, i.e., the air vehicle configuration (both hardware and software) is correct, test planning is complete, and other go/no-go criteria can be met. All of these factors are part of the test plan. As each detailed test plan is developed, the engineers assign (or budget) a specific time-weighting to each test point that will not change unless the test program is re-baselined. All of the budgeted TWTP from each detailed test plan is scheduled over the span of the test program in the master program test plan. Once execution begins, the planned schedule for TWTP accomplishment and the time-weighting budget for each test point cannot be changed unless the flight test program is re-baselined. This management method demands consistency in the cost (time-weighting) to work (all the effort and resources required to execute test points) relationship.

“At only the 25% planned completion point of the planned test program, the program manager would have an indication that the program duration was likely to significantly increase without some intervention.

One challenge for flight test programs, particularly large new weapon system developments like F-35, is the task of identifying all the required test points before the program starts. Using time-weighted test points offers the flight test manager several ways to tackle the ambiguity of flight test planning. First, the program could use an error margin on the time-weighting of individual test points, particularly those test points that pose a greater risk of unknowns. Second, when test managers build assumptions for inefficiencies (repeats, regression, and re-fly), they can multiply the time-weighting of the test points that are expected to be more challenging to accomplish. Last, flight test managers can build a management reserve into the program up front to acknowledge the challenge of identifying every test point that will be required to verify the system's functions and capabilities. This would account for added test points due to discovery.

Once execution begins, as test points are added to the program, their time-weighting and closure status must be tracked closely. However, time from these added test points cannot be included in the TWTP closed until the program is re-baselined to include these test points. The time spent executing the unplanned test points does need to be included in the TWTP executed metric.

One may ask what happens when the engineer budgets 5 minutes to accomplish a test point, but test execution proves that it actually takes 7 minutes to accomplish. In this case, the accounting system uses the budgeted time-weighting with no changes as previously stated. However, one of the opportunities of the TWTP metric is that the actual time required to accomplish each test point can be logged in the verification matrix so that the test planning phase for a future increment of the program can leverage this knowledge. When budgeting the time-weighting of test points, the follow-on program increment can build a more accurate test program plan.

Let's revisit the NAVAIR program mentioned earlier in this article. The TPM method used test point weighting to measure the value of a test point as proposed here. However, the weighting scheme attempted to capture test risk and complexity in addition to time required. The weighting method was more challenging to apply than the proposed approach; however, it proved useful as a way to measure test point value. Once combined with the process model, the P-8 test program was able to provide insight about future outcomes based on past events and management decisions currently being considered (F. DiBonaventuro, personal communication, March 7, 2018; P. Leard, personal communication, 2009).

Can EVM and ES Management Methods Help the Flight Test Manager?

We will now introduce the EVM concept to the discussion. Those familiar with EVM will recognize the following terms: Actual Cost for Work Performed (ACWP), Budgeted Cost for Work Performed (BCWP) or earned value, and Budgeted Cost for Work Scheduled (BCWS) or planned value (Lipke, 2003). These terms correspond to TWTP executed, TWTP closed, and TWTP planned, respectively. So TWTP executed represent the actual testing performed, but the TWTP closed represent the smaller amount of test requirements that are complete. Test requirements are typically tracked as open or closed (complete) in a master database or verification matrix. The TWTP planned represent the planned cost of the work scheduled as of a reporting date. For the test program, the analog of budget at completion (BAC) is the total amount of TWTP planned for the test program (total budgeted TWTP), and this value does not change unless the program is re-baselined. Using the TWTP provides a better unit of measure to relate test requirements to cost and schedule for the flight test program manager. These parameters can be used to calculate the flight test program performance indicators analogous to the EVM performance indicators: cost variance (CV), cost performance index (CPI), and to complete performance index (TCPI) (Defense Acquisition University, 2017). We will use the following terms for our performance indicators: execution variance (EXV), execution performance index (EPI), and completion execution index (CEI).

$$EXV = TWTP_C - TWTP_E \tag{1}$$

$$EPI = \frac{TWTP_C}{TWTP_E} \tag{2}$$

$$CEI = \frac{TWTP_{P,Total} - TWTP_C}{TWTP_{P,Total} - TWTP_E} \tag{3}$$

Just like traditional EVM that uses dollars as the currency, a positive execution variance and EPI ≥ 1.0 are favorable performance indicators for the TWTP currency. These performance indicators will help the flight test manager understand how effective the program is concerning test point completion.

While the EVM methodology can provide actionable cost performance information to a program manager, it does have some known flaws with respect to monitoring schedule performance. First, EVM often fails to provide a good schedule performance indicator as a project nears completion. Regardless of how far behind schedule the program runs, the schedule variance and schedule performance index tend toward zero and one, respectively, which are indications of "on schedule" performance (Crumrine &

Ritschel, 2013; Lipke, 2003; Lipke & Henderson, 2006). However, the ES methodology developed by Walt Lipke using the currency of time to measure performance is directly applicable to our problem of finding an appropriate metric for a test program. Lipke defines ES as the number of completed planned value (PV) time increments exceeded by earned value (EV) plus the fraction of the next incomplete planned value increment (Lipke, 2003; Lipke & Henderson, 2006).

$$ES = C + \frac{EV_C - PV_C}{PV_{C+1} - PV_C}\text{, where} \tag{4}$$

C = Number of completed planned value time increments (number of time increments where EV ≥ PV)

EV_C = Earned value (BCWP) of time increments where EV ≥ PV

PV_C = Planned value (BCWS) at the completed time increment

PV_{C+1} = Planned value (BCWS) at the next time increment to be completed

So the ES is the program duration planned to accomplish the current earned value (recall that earned value for this approach is all the closed TWTP). Now, using our test program parlance, we introduce the following equation for ES:

$$ES = CPD + \frac{TWTP_C - TWTP_{CPD}}{TWTP_{CPD+1} - TWTP_{CPD}}\text{, where} \tag{5}$$

CPD = Completed planned duration increments corresponding to the current earned value ($TWTP_C$). For example, if the original baseline program plan assumed 10 hours of TWTP would be closed in the first 2 months of flight test, then when 10 hours of $TWTP_C$ is accomplished, CPD is 2 months regardless of the actual time required to complete this amount of testing.

$TWTP_C$ = Current time-weighted test points closed

$TWTP_{CPD}$ = Planned value (planned TWTP) associated with CPD.

$TWTP_{CPD+1}$ = Next increment of planned value to be completed. This next increment of planned value is associated with the next planned duration increment. For our example, the program's planned value (expected $TWTP_C$ of 10 hours) was associated with a duration of

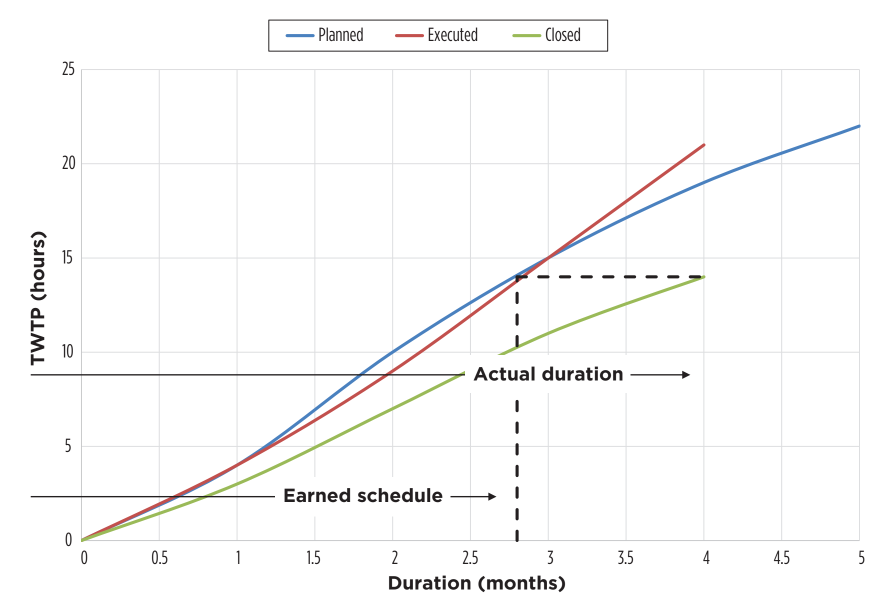
2 months. So continuing with our previous example, the next increment of duration would be the third month, when the program's integrated master schedule called for 15 hours of TWTP to be closed ($TWTP_{CPD+1}$ = 15 hours).

Now, to complete the example, assume that only 14 hours of $TWTP_C$ had been achieved after 4 months. Then equation 5 would yield the following:

$$ES = 2 + \frac{14 - 10}{15 - 10} = 2.8 \text{ months}$$

So the program is already 1.2 months behind schedule (4 – 2.8 months). Figure 1 depicts the ES concept using our example integrated master schedule.

FIGURE 1. EARNED SCHEDULE CONCEPT



Now that we have defined ES, we can again reference Lipke's work to define our analogs for the time-based schedule performance indicators: schedule variance (SV(t)), schedule performance index (SPI(t)) (Lipke, 2003), and to complete schedule performance indicator (TCSPI(t)) (Henderson, 2004). We will use the following naming convention: test schedule variance (TSV), test schedule performance index (TSPI), and completion schedule performance index (CSPI).

$$TSV = ES - AD \tag{6}$$

$$TSPI = \frac{ES}{AD} \tag{7}$$

$$CSPI = \frac{TPD - ES}{TPD - AD}\text{, where} \tag{8}$$

AD = Actual current duration of test program (e.g., 5 months after first flight, *AD* = 5 months).

TPD = Total planned test program duration (i.e., the number of months planned to complete the test program). This value does not change unless the program baseline changes, and is analogous to the BAC.

A final note on convention—first flight marks the beginning of actual schedule duration, not the planned first flight date (i.e., $TWTP_P$, $TWTP_E$, and $TWTP_C$ all start at the same time—first flight). While the test program is accruing cost to the program long before first flight, the proposed management method is tracking a different currency than dollars.

Again, just like our cost performance indicators, a positive variance (ES exceeds actual duration) and a performance index that is greater than 1.0 are both positive performance indicators. The ES method was chosen to track schedule performance because the ES method has been shown to provide more accurate schedule indicators than EVM (Crumrine & Ritschel, 2013).

The test manager can use these performance indicators to set execution goals for the program. As the test program is developed, the test manager can analyze the impact of inefficiencies by projecting the cost and schedule impacts of different indices and variances. As previously discussed, when the test manager builds management reserve into the flight test plan, it can be based upon an assumed performance index and variance. Failing to meet the performance goals may be the leading indicator to the test manager that the plan will need to be modified.

> **"The use of test point burn-down and burn-up curves refers to a single flight test management approach.**

Next, we will discuss the powerful part of this methodology—program predictors. Once again, we adapt the formulas from EVM and ES to our test terminology (Defense Acquisition University, 2017; Lipke & Henderson, 2006). For our proposed flight test management methodology, the analog to estimate at completion (EAC) will be called estimated test required (ETR). The analog to independent estimate at completion for schedule (IEAC(t)) will be called estimated test duration (ETD). The ETR parameter uses the to-date actual work and schedule executed (not earned), the remaining work, and performance indices to forecast what the flight test program will cost in our flight test currency (total executed TWTP). Likewise, we can forecast when the flight test program will complete (total flight test program duration in months).

$$ETR = TWTP_E + \frac{TWTP_{P,Total} - TWTP_C}{EPI} \qquad (9)$$

$$ETD = AD + \frac{TPD - ES}{TPSI} \qquad (10)$$

Equation 11 is similar to equation 10, and offers a second approach to calculating the ETD using a composite performance factor that multiplies our analogs for CPI and SPI(t) (Henderson, 2004; Lipke & Henderson, 2006). The ETR parameter takes into account the test program's effectiveness and efficiency at closing TWTP. For example, test point repeats, re-fly, regression points, and added test points will all lead to an increase in $TWTP_E$ required to complete the test program. Likewise, all the added test execution will increase the duration of the test program unless test capacity is increased to compensate for the increased scope of work.

$$ETD_{comp} = AD + \frac{TPD - ES}{ESP \cdot TSPI} \quad (11)$$

The following Tables 3, 4, and 5 present all our equations alongside the EVM and ES equations—similar to the DAU Gold Card.

TABLE 3. FLIGHT TEST EXECUTION GOLD CARD

Flight Test Execution	Earned Value Management	Earned Schedule
	Cost Performance Parameters	
$EXV = TWTP_C - TWTP_E$	$CV = BCWP - ACWP$	
$EPI = \frac{TWTP_C}{TWTP_E}$	$CPI = \frac{BCWP}{ACWP}$	
$CEI = \frac{TWTP_{P,Total} - TWTP_C}{TWTP_{E,Total} - TWTP_E}$	$TCPI = \frac{BAC - BCWP_{Cum}}{BAC - ACWP_{Cum}}$	

ACWP	**Actual Cost of Work Performed**	Cost actually incurred in accomplishing work performed (actual cost)
BAC	**Budget At Completion**	The sum of all budgets for the contract through any given Work Breakdown Structure/Organization Breakdown Structure level
BCWP	**Budgeted Cost for Work Performed**	Value of completed work in terms of the work's assigned budget (earned value)
CEI	**Completion Execution Index**	Efficiency needed from "time now" to achieve total planned TWTP
TCPI	**To Complete Performance Index**	Efficiency needed from "time now" to achieve a Cost Target (BAC or EAC)
$TWTP_C$	**Closed Time-Weighted Test Point**	Test points verified (earned value)
$TWTP_E$	**Executed Time-Weighted Test Point**	Test points executed (actual cost)
$TWTP_P$	**Planned Time-Weighted Test Point**	Time-phased budget plan for time-weighted test points execution (planned value)
$TWTP_{P,Total}$	**Total Planned Time-Weighted Test Points**	The sum of all budgeted time-weighted test points for the flight test program

Performance Indicators

CEI	Completion Execution Index
CPI	Cost Performance Index
CV	Cost Variance
EPI	Execution Performance Index
EXV	Execution Variance
TCPI	To Complete Performance Index

Note. Adapted from (Defense Acquisition University, 2017); (Henderson, 2004); (Lipke, 2003, 2006); (Lipke & Henderson, 2006).

TABLE 4. FLIGHT TEST SCHEDULE GOLD CARD

Flight Test Execution	Earned Value Management	Earned Schedule
	Schedule Performance Parameters	
$ES = CPD + \frac{TWTP_C - TWTP_{CPD}}{TWTP_{CPD+1} - TWTP_{CPD}}$		$ES = C + \frac{EV - PV_C}{PV_{C+1} - PV_C}$
$TSV = ES - AD$	$SV = BCWP - BCWS$	$SV(t) = ES - AT$
$TSPI = \frac{ES}{AD}$	$SPI = \frac{BCWP}{BCWS}$	$SPI(t) = \frac{ES}{AT}$
$CSPI = \frac{TPD - ES}{TPD - AD}$		$TCSPI(t) = \frac{PD - ES}{TEAC - AT}$
AD	**Actual Duration**	Duration incurred in accomplishing testing performed
AT	**Actual Time**	Duration incurred in accomplishing work performed (actual duration)
BCWP	**Budgeted Cost for Work Performed**	Value of completed work in terms of the assigned budget (EV)
BCWS	**Budgeted Cost for Work Scheduled**	Time-phased Budget Plan for work scheduled (planned value)
CPD	**Completed Planned Duration**	Completed planned duration increments corresponding to the current earned value (closed TWTP)
CSPI	**Completion Schedule Performance Index**	Efficiency needed from "time now" to achieve a flight test program Schedule Target (TPD)
ES	**Earned Schedule**	Schedule of completed work in terms of the work's assigned duration
EV	**Earned Value**	Budgeted cost of work performed
PD	**Planned Duration**	Budgeted duration of the project
PV	**Planned Value**	Budgeted cost of work scheduled
TCSPI(t)	**To Complete Schedule Performance Index**	Efficiency needed from "time now" to achieve a Schedule Target (TEAC)
TEAC	**Time Estimate At Completion**	Estimate of total duration
TPD	**Total Planned Duration**	Total planned flight test program duration
$TWTP_C$	**Closed Time-Weighted Test Point**	Test points verified (earned value)
$TWTP_{CPD}$	**Planned Time-Weighted Test Point**	Planned value (planned TWTP) associated with CPD
Performance Indicators		
SPI	Schedule Performance Index	
SPI(t)	Time-Based Schedule Performance Index	
SV	Schedule Variance	
SV(t)	Time-Based Schedule Variance	
TSPI	Test Schedule Performance Index	
TSV	Test Schedule Variance	

Note. (Defense Acquisition University, 2017); (Lipke & Henderson, 2006); (Lipke, 2006); (Henderson, 2004); (Lipke, 2003).

TABLE 5. FLIGHT TEST FORECAST GOLD CARD

Flight Test Execution	Earned Value Management	Earned Schedule
	Estimate at Completion	
$ETR = TWTP_E + \frac{TWTP_{P,Total} - TWTP_C}{EPI}$	$EAC = ACWP_{Cum} + \frac{BAC - BCWP_{Cum}}{CPI_{Cum}}$	
$ETD = AD + \frac{TPD - ES}{TSPI}$		$IEAC(t) = AT + \frac{PD - ES}{SPI(t)}$
$ETD_{comp} = AD + \frac{TPD - ES}{EPI \cdot TSPI}$		$IEAC(t)_{comp} = AT + \frac{PD - ES}{CPI \cdot SPI(t)}$

ACWP	**Actual Cost of Work Performed**	Cost actually incurred in accomplishing work performed (actual cost)
AD	**Actual Duration**	Duration incurred in accomplishing testing performed
AT	**Actual Time**	Duration incurred in accomplishing work performed (actual duration)
BAC	**Budget At Completion**	The sum of all budgets for the contract thru any given Work Breakdown Structure/Organizational Breakdown Structure level
BCWP	**Budgeted Cost for Work Performed**	Value of completed work in terms of the work's assigned budget (earned value)
EAC	**Estimate At Completion**	Estimate of total Cost for the contract through any given level
ES	**Earned Schedule**	Schedule of completed work in terms of the work's assigned duration
ETD	**Estimated Test Duration**	Estimate of total Duration for the flight test program
ETR	**Estimated Test Required**	Estimate of total TWTP that needs to be executed to achieve total TWTP planned
IEAC(t)	**Independent Estimate At Completion**	Estimate of total Duration for the project
PD	**Planned Duration**	Budgeted duration of the project
TPD	**Total Planned Duration**	Total planned flight test program duration
$TWTP_C$	**Closed Time-Weighted Test Point**	Test points verified (earned value)
$TWTP_E$	**Executed Time-Weighted Test Point**	Test points executed (actual cost)
$TWTP_{P,Total}$	**Total Planned Time-Weighted Test Points**	The sum of all budgeted test points for the flight test program

Performance Indicators

CPI	Cost Performance Index
EPI	Execution Performance Index
SPI(t)	Time-Based Schedule Performance Index
TSPI	Test Schedule Performance Index

Note. (Defense Acquisition University, 2017); (Lipke & Henderson, 2006); (Lipke, 2006); (Henderson, 2004); (Lipke, 2003).

Combining Concepts for Greater Insights

Lipke (2006) discusses the application of the ES method to critical path analysis, which applies the ES method to a group of segregated tasks comprising a program in addition to the program aggregate. Applying this approach to a flight test program could provide valuable insights that would otherwise be missing. First, we will need to use the concept used by the F-35 program to group and track TWTP by different capability areas of interest. The verification of each capability area would be planned at the beginning of the program no differently than the aggregate flight test program. The TWTP would be scheduled over the duration of the test program using the build-up principles that are always applied to program planning while observing the planned progression of function verification. An example flight test program plan is shown below in Figure 2.

FIGURE 2. CAPABILITY-BASED TWTP TRACKING CONCEPT

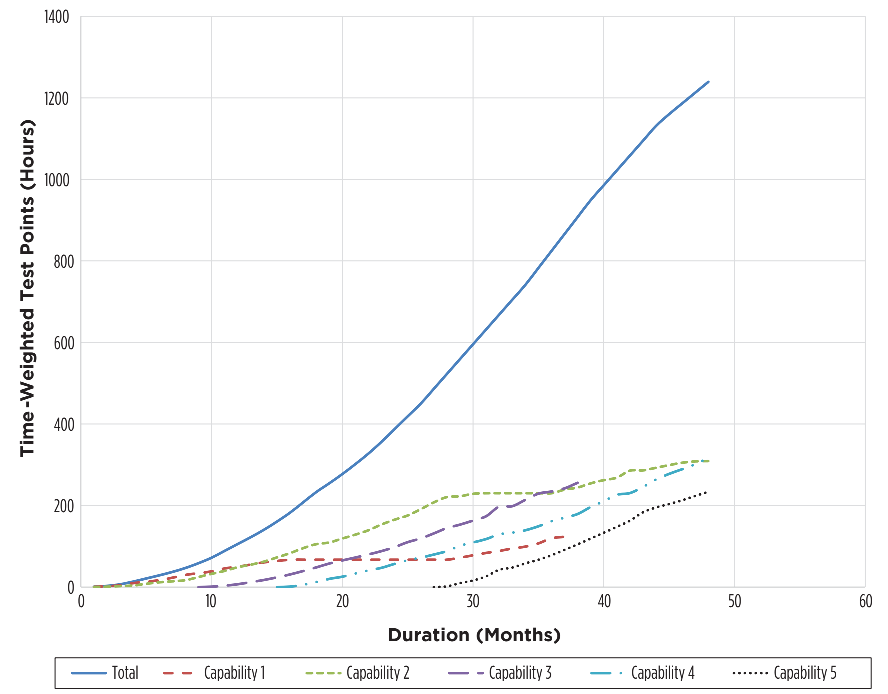


Figure 2 illustrates that each capability area's planned TWTP offers the same opportunity to apply the approach presented above using the same equations for performance indicators and completion predictors. Now we can track progress toward the aggregate program and the various capability areas simultaneously. Each capability could be further segregated into subgroups for various subsystems or functions required to achieve each capability.

Combining the approach used by the F-35 program (grouping test points under capability development threads) with this EVM- and ES-based approach provides insight into the development and verification status of each major capability of interest. Thus, the program test manager is able to work with program stakeholders to make informed trade-off decisions between verifying different capabilities when resources are constrained or certain functions are falling behind expected maturity needs. This could lead to greater efficiency as it avoids overflying the system's maturity in a capability area, which usually leads to re-flying test points after software updates have been made. Last, this added insight from the TWTP concept allows the flight test manager to see the functions and resulting capabilities that will pace the program's completion (i.e., critical path).

If the program manager determines that a function or capability needs to be removed from the current weapon system development increment, those TWTPs will need to be removed from the re-baselined test program. When the test program is re-baselined, the flight test manager will also need to add the time for all the new test points that have been discovered. As TWTP are removed from a test program, the test manager will have an indication of how much test scope is traveling to the next increment of the system if the user still needs those functions and capabilities.

Methodology

The author attempted to gather historical data from the B-2 EMD (or later upgrade) flight test program since the B-2 is the only known program to have used a time-weighted test point metric. However, the attempt to obtain program data of sufficient detail to use the proposed method was unsuccessful. Instead, the author used a Microsoft Excel spreadsheet to simulate notional data. For this example, the original flight test program plan (as of first flight) assumed a 48-month duration program to complete (close) 1,239 hours of TWTP. The "actual" data were simulated by starting with a nominal test capacity build-up of aircraft months, assuming a number of sorties and flight hours per aircraft month. Flight hours were converted to $TWTP_E$ by multiplying by an efficiency factor that was determined using the bounded random function (RANDBETWEEN). The bounds used on the efficiency factor (and other assumptions) were based on 15 years of personal experience in the flight test career field and are shown in Table 6. The efficiency factor provides for the difference between total flight hours and the time spent executing test points. Next, bounded random functions were used to emulate historically representative test program execution (regression, re-fly, repeat, and added/subtracted test points).

TABLE 6. ASSUMPTIONS USED TO CREATE TWTP EXECUTED AND TWTP CLOSED PROFILES

TWTP Per Month	Assumption
Executed in pursuit of planned test points	20–35% of flight hours/month (months 1 to 6) 20–50% of flight hours/month (month 7 to end)
Added execution for regression test points	2 hrs every 9 months
Executed burn-up for repeats, re-fly, added test points	A linear profile was used to represent the shifting focus from flight sciences (FS) to missions systems (MS) verification that occurs in flight test programs. The profile follows: • Months 1 to 3: 90% FS and 10% MS work • Months 4 to 60: linear change from 10% to 80% MS • Months 61 to end: 20% FS and 80% MS work To account for the difference in efficiency between FS and MS testing that occurs in many modern flight test programs, the profiles described above were multiplied by the following efficiency factors: • Flight sciences: 5–35% of $TWTP_E$ per month • Missions systems: 20–50% of $TWTP_E$ per month

The TWTP burn-up (additions to TWTP due to re-fly, repeat, regression, and test point additions) was booked against TWTP closed for that month. If there was no burn-up for a month, then TWTPC would equal $TWTP_E$ (i.e., every test point was flown satisfactorily, resulting in a closed test point after analysis and no test point burn-up occurred during the month). The test point burn-up causes $TWTP_E$ and $TWTP_C$ to diverge in the simulated data. If test points were removed from the program, then those test points would be booked as TWTP closed, and it would be possible for $TWTP_C$ to be greater than $TWTP_E$ during a single month. During flight test, there are numerous reasons a test point might be removed. Examples include: deferred capability delivery, program re-baselines that expect to leverage data from a source other than flight test or reduced intermediate (build-up) test points, closing a test point based on analysis of other test point results, etc.

Analysis

The approach described in the previous section showed promise as a methodology to provide actionable projections of cost (TWTP) and schedule for a test program manager. Figure 3 below shows sample data from a single simulated program using the assumptions in Table 6. The figure depicts the original burn-down plan for test points ($TWTP_P$), and the "actual" $TWTP_E$ and $TWTP_C$ over time, where time is tracked in months shown by the abscissa. The predictors (estimate at completion) results are also shown for TWTP and program duration. The predictors converge on the final TWTP and duration as expected. The most noteworthy aspect of this chart is that the predictors are providing valuable information to the program manager within the first 12 months. As early as the 12th month, the ETR parameter predicts the program plan is approximately 500 hours short of TWTP execution needed to complete the program. This parameter tracks consistently throughout the duration of the program toward the actual $TWTP_E$ required to complete the test program. The schedule predictor (ETD) also shows that the planned duration is more than a year short of what will be needed to complete the test program. The composite schedule predictor (ETD_{comp}) shows an even greater deficit in the program plan. Again, both schedule predictors trend toward the simulated "actual" duration as the program progresses.

Table 7 summarizes the predictors (ETR, ETD, and ETD_{comp}) every 12 months for the first 5 years throughout this simulated test program, which can be compared to the "actual" executed TWTP (1,757 hours) and flight test program duration (75 months). Note how all three predictors forecast a need to add duration or capacity to make up the disparity between the planned program (TWTP of 1,239 hours and 48-month duration) and the simulated "actual" test program.

FIGURE 3. SAMPLE OUTPUT OF A SINGLE TEST PROGRAM TRIAL

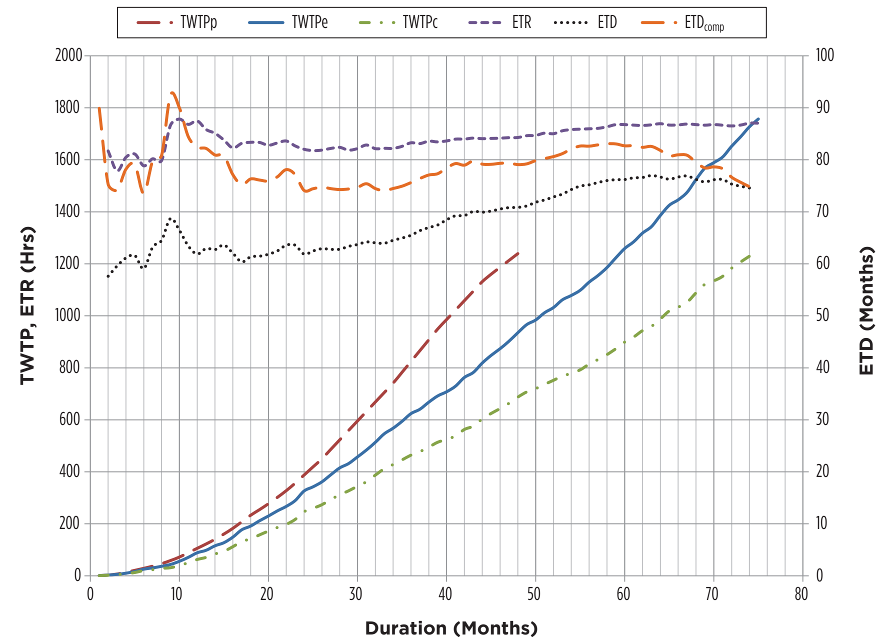


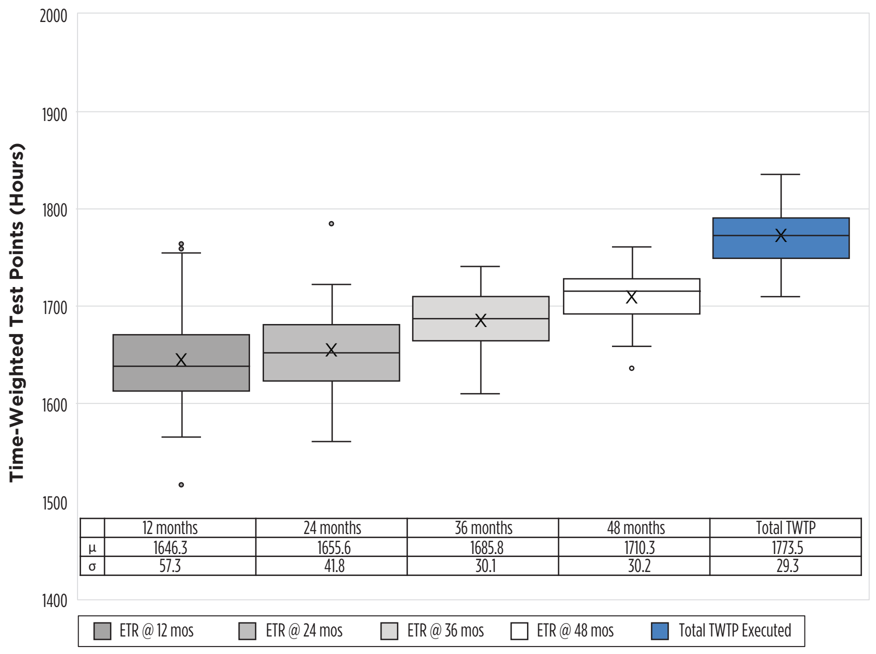
TABLE 7. DATA FROM A SINGLE SIMULATED TEST PROGRAM

Month of Program	ETR (hours)	ETD (months)	ETD_{comp} (months)
12	1750	62	82
24	1640	62	74
36	1665	65	76
48	1685	71	79
60	1735	76	82

At only the 25% planned completion point of the planned test program, the program manager would have an indication that the program duration was likely to significantly increase without some intervention.

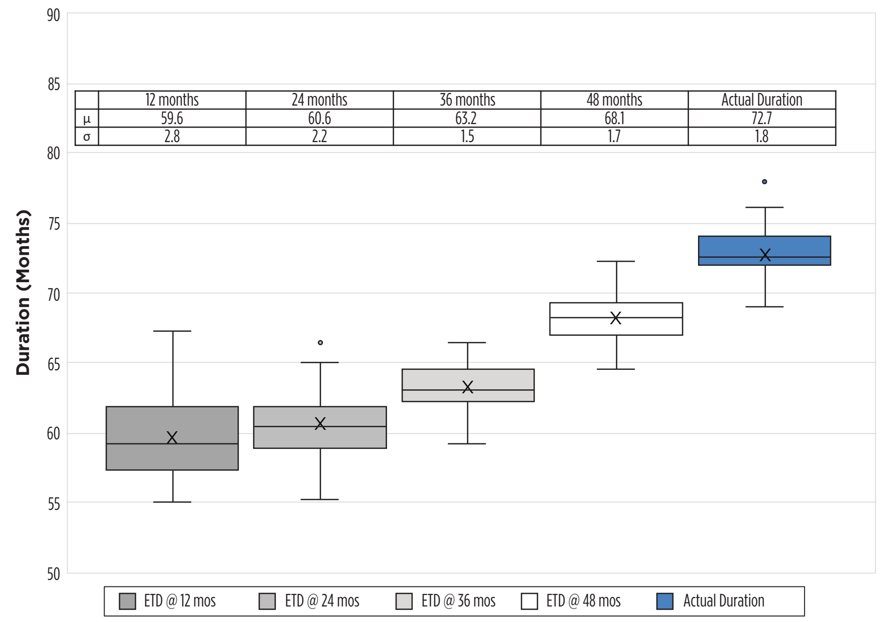
To further explore the viability of this approach, the same program plan and assumptions were used to simulate the outcome of 50 flight test programs. Statistical data from these simulations are shown in Figures 4 to 6. The box and whisker plots in each figure show the distribution of outcomes for the 50 "actual" simulated test programs on the right side of each figure. Moving from left to right in each figure, we see the box and whisker plots of our predictors (ETR, ETD, and ETD_{comp}) over time. The predictors are shown at four different moments in time (12, 24, 36, and 48 months after first flight). Recall that our objective is to forecast the actual cost of work performed ($TWTP_E$) and actual duration. Therefore, the performance of the predictor parameters early in the program are of particular interest.

FIGURE 4. ETR PREDICTOR AT 12-MONTH INTERVALS AND SIMULATED ACTUAL $TWTP_E$



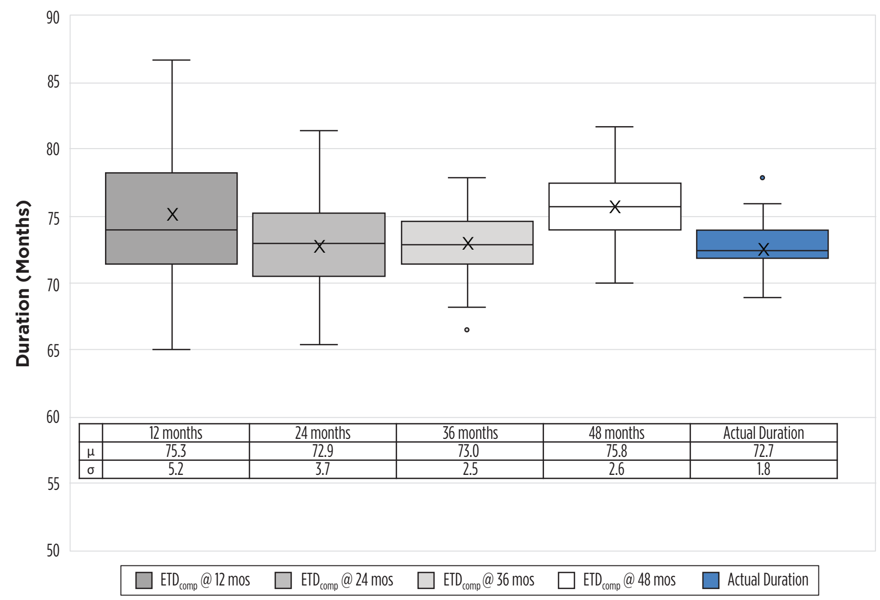
	12 months	24 months	36 months	48 months	Total TWTP
μ	1646.3	1655.6	1685.8	1710.3	1773.5
σ	57.3	41.8	30.1	30.2	29.3

FIGURE 5. ETD PREDICTOR AT 12-MONTH INTERVALS AND SIMULATED ACTUAL DURATION



	12 months	24 months	36 months	48 months	Actual Duration
μ	59.6	60.6	63.2	68.1	72.7
σ	2.8	2.2	1.5	1.7	1.8

FIGURE 6. ETD_{COMP} PREDICTOR AT 12-MONTH INTERVALS AND SIMULATED ACTUAL DURATION



	12 months	24 months	36 months	48 months	Actual Duration
μ	75.3	72.9	73.0	75.8	72.7
σ	5.2	3.7	2.5	2.6	1.8

The box and whisker plot was chosen to depict these data because it is one of the most succinct ways to depict descriptive statistics for a data set. Further, the purpose here is to show whether or not the predictors are converging toward the "actual" execution data for the simulated programs. Finally, the box and whisker plot quickly depicts the variability (standard deviation and outliers) and skew in each data set for a quick comparison. When viewing a box plot, it is important to know what defines an outlier as the ends of the whiskers can represent different values depending on the chosen convention. These charts were created in Microsoft Excel, which uses the convention attributed to John Tukey where data are considered outliers if they lie beyond 1.5 times the interquartile range from either end of the box (also called the inner fences). Recall that the ends of the box are always defined by the first and third quartiles. The whiskers extend to the most extreme sample in the data set that is within each of the two inner fences (Dawson, 2011).

From these 50 simulated programs, we see that our observations from the single trial described above hold—the ETR predictor provides actionable information to the program manager. This information is useful as early as the first year of the program and it trends in the correct direction throughout the program. After 12 months, all 50 trials predict a need to amend the program plan for more test execution (ETR ranges between 1,518.4 to 1,764.1 hours of TWTP execution compared to the 1,239 hours of total TWTP planned). Additionally, the distribution of ETR improves (converges toward the distribution of the total "actual" TWTP executed) with each successive 12-month interval of simulated execution. The data distributions of the ETR parameter over time indicate that it would consistently alert the flight test manager to an impending cost overrun (recall that we are measuring cost by TWTP executed instead of dollars). The standard deviations and mean values for each parameter and time period are also noted in Figure 4.

The ES approach using TWTP as the currency also yields actionable information for a flight test manager as evident by the box plots in Figures 5 and 6. The 48-month planned program ballooned on average to nearly 73 months (similar performance to some of the modern military aircraft test programs discussed earlier). Using equation 10 for ETD, the proposed methodology provided mean estimated program durations of 59.6, 60.6, 63.2, and 68.1 months at the 12-, 24-, 36-, and 48-month points in the program, respectively. The range of data for ETD at the 12-month point was 55.0 to 67.2 months. So at only the 25% planned completion point of the planned test program, the program manager would have an indication that the program duration was likely to significantly increase without some intervention.

The duration predictor (ETD) tends to be optimistic as indicated by comparing the data distributions for ETD and the "actual" duration of the simulated test programs. However, the distributions of duration predictions are converging toward the distribution of "actual" duration at each successive interval of time, indicating that the predictions are more accurate as the program progresses.

Using the composite ETD predictor for duration (equation 11) also yields an early indication of schedule problems. Just 12 months into execution of the program, the mean of the ETD_{comp} distribution indicated that the program would require 75.3 months to complete. The range of ETD_{comp} after 12 months was 65.2 to 86.8 months. The mean of ETD_{comp} at the 12-, 24-, 36-, and 48-month points in the program were 75.3, 72.9, 73.0, and 75.8 months, respectively. The ETD_{comp} parameter often provided a slightly pessimistic prediction of performance, but was always reliable in alerting the program manager to a pending problem. Also, using ETD_{comp}, the predicted test program duration always converged to the final "actual" duration as in the single sample shown in Figure 3. The distributions of data for both duration predictors indicate that both parameters would alert the flight test manager to impending schedule overruns as desired.

Conclusions

When test program managers are given the tools discussed in this article, they can make timely and actionable decisions to either increase test capacity, reduce the scope of the test program, or negotiate an appropriate extension to the program with the acquisition executive. If the TWTPs are grouped into capability groupings (similar to the F-35 test program), then the test program manager can begin to understand how much effort is needed to verify a capability or to complete a test thread. (A test thread is a particular line of testing, typically under a single detailed test plan, e.g., structural testing.) This added detail also affords the manager the opportunity to trade effort toward a capability that is not maturing as expected for another capability that is on-track and ready for further verification. Understanding these relationships allows the program manager to work with other stakeholders to determine what functions and capabilities can be delayed to a follow-on increment of the program. It also gives the manager a sense of critical path because the level of effort (number of flight hours and sorties) to verify a series of functions and capabilities that build on one another is better understood when using the TWTP. If the program manager decides to extend the program rather than cut capability or increase capacity (which is usually difficult or unrealistic to accomplish in the short term), the program at least has an opportunity to realistically re-baseline the program to something executable with the current resources. With an appropriate program re-baseline, stability is now embedded (in the budget and the schedule) that has been lacking in recent flight test programs.

This approach has at least one other benefit. As the program executes, the time-weighting expected for each type of test point can be noted in accordance with the time actually required during execution. While these data would not be applied to the test program already in progress, they would allow better up-front scope and schedule estimates for the follow-on test program of the next system block increment.

Keys to successfully using this approach are:

- Define an accounting method for all the variables, and then be consistent throughout the program's execution.
- Be vigilant in tracking all the sources of additional (or subtracted) TWTP.
- Keep up with test point closure metrics (do not fall behind in data analysis).
- Combine the EVM and ES methods with capability-based metrics (the methodology used by F-35) to guide program decisions.

Industry leaders depend on setting goals for execution and holding themselves accountable by tracking the right metrics. With a proper methodology and proper metric, flight test programs can achieve better execution outcomes.

References

420th Flight Test Squadron. (1997). B-2 program status review: Flight week of June 23–29, 1997 [PowerPoint slides]. Phoenix, AZ: 420th Flight Test Squadron, Test Program Planning Office.

Acker, D. D. (1983, April). *Evaluation of the effectiveness of the Defense Systems Acquisition Review Council (DSARC)* (Vol. 2, Pt. 1). Retrieved from http://www.dtic.mil/docs/citations/ADA129796

Boeing. (n.d.). B-1 Lancer Bomber historical snapshot. Retrieved from http://www.boeing.com/history/products/b-1b-lancer.page

Crumrine, K. T., & Ritschel, J. D. (2013, Spring). A comparison of earned value management and earned schedule as schedule predictors on DoD ACAT I programs. *The Measurable News.* Retrieved from http://www.mycpm.org/news-events/measurable-news/

Dawson, R. (2011, July). How significant is a boxplot outlier? *Journal of Statistics Education, 19*(2). Retrieved from http://ww2.amstat.org/publications/jse/v19n2/dawson.pdf

Defense Acquisition University. (2017). Earned value management gold card [Instructional card]. Retrieved from https://www.dau.mil/tools/Lists/DAUTools/Attachments/19/20170720-gold%20card%201%20sided.pdf

Department of Defense. (1990). *C-17 Selected Acquisition Report.* Available from Defense Acquisition Management Information Retrieval (DAMIR) information repository. Retrieved from https://ebiz.acq.osd.mil/damir

Department of Defense. (1991). *F-22 Selected Acquisition Report.* Available from DAMIR information repository. Retrieved from https://ebiz.acq.osd.mil/damir

Department of Defense. (1992). *F/A-18E/F Selected Acquisition Report.* Available from DAMIR information repository. Retrieved from https://ebiz.acq.osd.mil/damir

Department of Defense. (1997). *F-22 Acquisition Program Baseline.* Available from DAMIR information repository. Retrieved from https://ebiz.acq.osd.mil/damir

Department of Defense. (2001). *F-35 Acquisition Program Baseline.* Available from DAMIR information repository. Retrieved from https://ebiz.acq.osd.mil/damir

Department of Defense. (2007). *F-35 Acquisition Program Baseline.* Available from DAMIR information repository. Retrieved from https://ebiz.acq.osd.mil/damir

Department of Defense. (2010a). *C-17A Selected Acquisition Report.* Available from DAMIR information repository. Retrieved from https://ebiz.acq.osd.mil/damir

Department of Defense. (2010b). *F-22 Selected Acquisition Report.* Available from DAMIR information repository. Retrieved from https://ebiz.acq.osd.mil/damir

Department of Defense. (2012). *F/A-18E/F Selected Acquisition Report.* Available from DAMIR information repository. Retrieved from https://ebiz.acq.osd.mil/damir

Department of Defense. (2016). *F-35 Selected Acquisition Report.* Available from DAMIR information repository. Retrieved from https://ebiz.acq.osd.mil/damir

Drezner, J. A., & Smith, G. K. (1990). *An analysis of weapon system acquisition schedules* (Report No. R-3937-ACQ). Retrieved from https://www.rand.org/pubs/reports/R3937.html

Fox, B., Boito, M., Graser, J., & Younossi, O. (2004). *Test and evaluation trends and costs for aircraft and guided weapons* (MG-109-AF). Retrieved from https://www.rand.org/pubs/monographs/MG109.html

General Accounting Office. (1973). *B-l weapon system* [Staff Study]. Retrieved from https://www.gao.gov/assets/80/78824.pdf
General Accounting Office. (1977). *Status of the F-16 aircraft program* (Report No. PSAD-77-41). Washington, DC: U.S. Government Printing Office.
General Accounting Office. (1979). *Need to demonstrate F-18 Naval Strike Fighter weapon system effectiveness before large-scale production* (Report No. PSAD-79-25). Washington, DC: U.S. Government Printing Office.
General Accounting Office. (1983). *The B-l Bomber program—A new start* (Report No. GAO/MASAD-83-21). Washington, DC: U.S. Government Printing Office.
General Accounting Office. (1985). *Production of some major weapon systems began with only limited operational test and evaluation results* (Report No. GAO/NSIAD-65-66). Washington, DC: U.S. Government Printing Office.
General Accounting Office. (1989). *Strategic bomber: B-1B cost and performance remain uncertain* (Report No. GAO/NSIAD-89-55). Washington, DC: U.S. Government Printing Office.
General Accounting Office. (1990a). *Status of the B-2 Bomber program* (Report No. GAO/T-NSIAD-90-16). Washington, DC: U.S. Government Printing Office.
General Accounting Office. (1990b). *Strategic bomber: B-2 program status and current issues* (Report No. GAO/NSIAD-90-120). Washington, DC: U.S. Government Printing Office.
General Accounting Office. (1995). *B-2 bomber: Status of cost, development, and production* (Report No. GAO/NSIAD-95-164). Washington, DC: U.S. Government Printing Office.
General Accounting Office. (1996). *B-2 bomber: Status of efforts to acquire 21 operational aircraft* (Report No. GAO/NSIAD-97-11). Washington, DC: U.S. Government Printing Office.
General Accounting Office. (1998a). *F-22 aircraft: Progress in Achieving Engineering and Manufacturing Development Goals* (Report No. GAO/NSIAD-98-67). Washington, DC: U.S. Government Printing Office.
General Accounting Office. (1998b). *F-22 aircraft: Progress of the engineering and manufacturing development program* (Report No. GAO/NSIAD-98-137). Washington, DC: U.S. Government Printing Office.
General Accounting Office. (1999). *F-22 aircraft: Issues in achieving engineering and manufacturing development goals* (Report No. GAO/NSIAD-99-55). Washington, DC: U.S. Government Printing Office.
General Accounting Office. (2000a). *Defense acquisitions: Progress in meeting F-22 cost and schedule goals* (Report No. GAO/T-NSIAD-00-58). Washington, DC: U.S. Government Printing Office.
General Accounting Office. (2000b). *F-22 Aircraft: Development Cost Goal Achievable If Major Problems Are Avoided* (Report No. GAO/NSIAD-00-68). Washington, DC: U.S. Government Printing Office.
General Accounting Office. (2001). *Tactical aircraft: F-22 development and testing delays indicate need for limit on low-rate production* (Report No. GAO/NSIAD-01-310). Washington, DC: U.S. Government Printing Office.
General Accounting Office. (2002). *F-22 delays indicate initial production rates should be lower to reduce risks* (Report No. GAO/NSIAD-02-298). Washington, DC: U.S. Government Printing Office.

Government Accountability Office. (2006). *Joint Strike Fighter: DoD plans to enter production before testing demonstrates acceptable performance* (Report No. GAO-06-356). Washington, DC: U.S. Government Printing Office.

Government Accountability Office. (2008). *Joint Strike Fighter: Recent decisions by DoD add to program risks* (Report No. GAO-08-388). Washington, DC: U.S. Government Printing Office.

Government Accountability Office. (2011a). *Joint Strike Fighter: Program restructuring should improve outcomes, but progress is still lagging overall* (Report No. GAO-11-450T). Washington, DC: U.S. Government Printing Office.

Government Accountability Office. (2011b). *Joint Strike Fighter: Restructuring places program on firmer footing, but progress still lags* (Report No. GAO-11-325). Washington, DC: U.S. Government Printing Office.

Hehs, E., & Rhodes, J. (2012, June). F-22 milestones, Part 2. *Code One Magazine.* Retrieved from http://www.codeonemagazine.com/article.html?item_id=101

Henderson, K. (2004, Spring). Further developments in earned schedule. *The Measurable News.* Retrieved from http://www.earnedschedule.com/Papers.shtml

Lipke, W. (2003, Summer). Schedule is different. *The Measurable News.* Retrieved from http://www.earnedschedule.com/Papers.shtml

Lipke, W. (2006, Fall). Applying earned schedule to the critical path and more. *The Measurable News.* Retrieved from http://www.earnedschedule.com/Papers.shtml

Lipke, W., & Henderson, K. (2006, November). Earned schedule: An emerging enhancement to earned value management. *CrossTalk.* Retrieved from http://www.crosstalkonline.org/back-issues/

Rothman, M. B. (1987). *Aerospace weapon system acquisition milestones: A data base* (Report No. N-2599). Retrieved from https://www.rand.org/pubs/notes/N2599.html

Smith, G. K., Shulman, H. L., & Leonard, R. S. (1996). *Application of F-117 acquisition strategy to other programs in the new acquisition environment* (Report No. MR-749-AF). Retrieved from https://www.rand.org/pubs/monograph_reports/MR749.html

Tyson, K. W., Harmon, B. R., & Utech, D. M. (1994). *Understanding cost and schedule growth in acquisition programs.* Retrieved from http://www.dtic.mil/docs/citations/ADA284321

Author Biography

Lt Col C. Matt Ryan, USAF,

is a flight test engineer and is currently the commander of a test squadron. He received an MS in Aerospace Engineering and Systems Engineering from the University of Dayton and Air Force Institute of Technology, respectively. Lt Col Ryan is a graduate of USAF Test Pilot School and has served in flight test and test management roles for a variety of aircraft programs.

(E-mail address: charles.ryan.1@us.af.mil)