IMPROVING PREDICTIVE VALUE OF INDICATORS OF POOR PERFORMANCE

Patricia Fazio Bronson and David Sparrow

The Problem

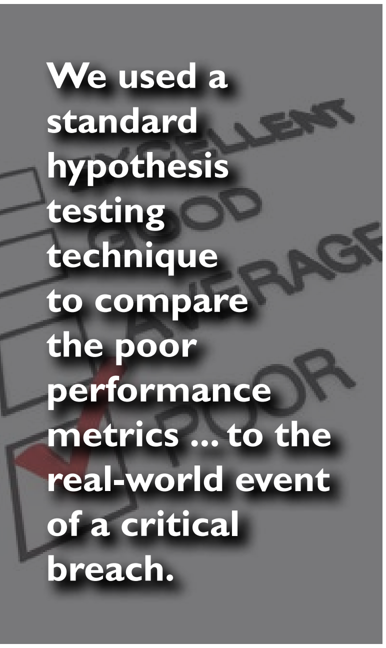
Screening techniques need to be developed to identify Major Defense Acquisition Programs that are likely to experience a critical Nunn-McCurdy breach.

The Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics, Performance Assessments and Root Cause Analyses (OUSD/AT&L/PARCA) asked IDA to develop screening techniques to identify Major Defense Acquisition Programs (MDAPs) that were likely to experience a critical Nunn-McCurdy breach. PARCA also asked IDA to develop performance assessment methods for MDAPs to support their participation in the Defense Acquisition Executive Summary (DAES) and Integrated Product Team (IPT) processes.

This article describes and evaluates a collection of metrics of poor performance that use program data available from Selected Acquisition Reports (SARs) and DAES Web Services, and earned value management (EVM) data from the EVM Central Repository.

The metrics are based on observed events that tend to indicate poor performance:

- Instability in funding and production rate profiles
- Differences between spending forecasts and execution of those forecasts
- Differences between staffing plans and the execution of those plans
- Differences between contract forecasts and funding plans
- Cost growth on mission equipment
- Changes to the estimated costs of developing prime mission equipment
- Persistent use of the Undistributed Budget category
- Rate at which Management Reserve is spent
- Initial investment in system level tasks
- Growth in the cost of system level tasks.



We used a standard hypothesis testing technique to compare the poor performance metrics ... to the real-world event of a critical breach.

Each metric is evaluated for its ability to identify programs that are likely to experience cost growth on the order of a critical Nunn-McCurdy breach. We used a standard hypothesis testing technique to compare the poor-performance metrics generated for eight programs to the real-world event of a critical breach in Program Acquisition Unit Cost (PAUC).

Figure 1 shows a sample plot of the Unit Cost Growth (UCG) for each program on the vertical axis as a function of a poor-performance index value on the horizontal axis. The poor-performance index value is a linear transformation of the poor-performance metrics for all programs, so the minimum index value is 0 and the maximum index value is 1. The horizontal blue line at 25 percent represents the real-world poor-performance event threshold value (the critical Nunn-McCurdy breach limit). The vertical blue line is the threshold value for detecting poor performance.

As with any test of this type, the sensitivity of the test is established by the placement of the vertical blue line. Move the line all the way to the right, and the results are misses and true negatives. Move the line all the way to the left, and the results are all hits and false alarms.

The results of all the tests are summarized in Table 1.

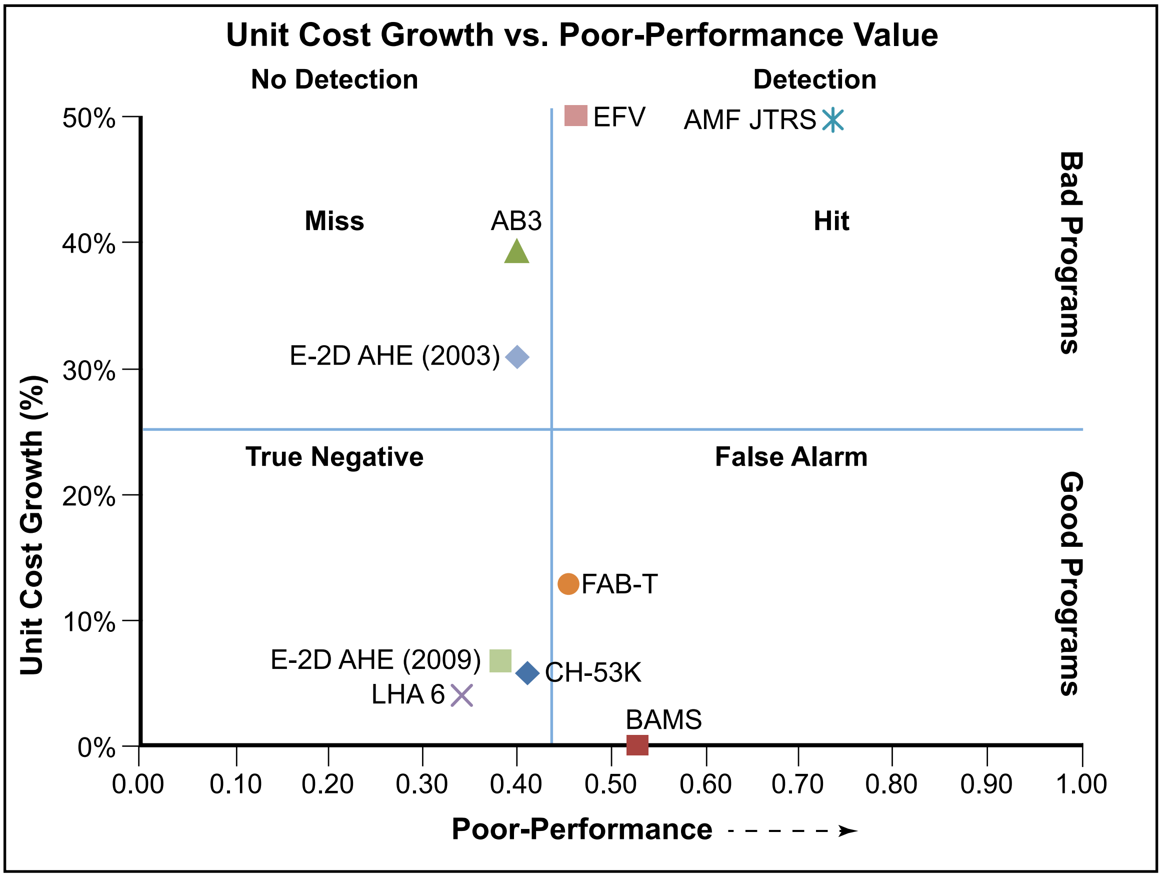


Figure 1. A Scatter Plot of Observed Unit Cost Growth and a Poor-Performance Index

Table 1. Summary of Test Results Showing Successes and Failures

	Instability in Funding and Quantity Profiles			Spending Forecasts			Execution to Staffing Plans				Volatility (#)				
	Funding	Quantity	Magnitude	Scatter	Magnitude	Scatter	Comparison of Contract Forecasts to Funding Plans	Cost Growth on Mission Equipment	Magnitude	Scatter	Undistributed Budget	Management Reserve	Initial Investment in System Level Tasks	System Level Tasks	SE SI ST Growth at 50% Complete
CH-53K								False Alarm	False Alarm						
BAMS								False Alarm					False Alarm		
AB3 Remanufacture	Hit	Miss	Hit	Miss	Miss	Miss	Miss	Miss	Hit	Hit	Miss	Miss	Hit	Miss	Miss
LHA 6													False Alarm		
AMF JTRS	Hit	Hit	Hit	Miss	Hit	Hit	Hit	Hit	Hit	Miss	Hit	Hit	Miss	Miss	Hit
FAB-T	False Alarm	False Alarm	False Alarm				False Alarm								
E-2D AHE (Jun 2003)	Miss	Miss	Miss	Miss	Miss	Miss	Miss	Hit	Miss	Miss	Hit	Hit	Hit	Miss	Miss
E-2D AHE (Jul 2009)								N/A			False Alarm	------- Not Applicable --------			
EFV	Hit	Miss	Miss	Miss	Miss	Hit	Miss	Miss	Hit	Miss	Miss	Miss	Miss	Hit	Hit
Success Rate	78%	56%	67%	56%	67%	78%	56%	50%	78%	67%	67%	78%	50%	63%	75%

The accuracy of the individual performance metrics ranges from 50 percent to 78 percent. Five of the fifteen metrics (33 percent) have a success rate between 75 and 78 percent—better than random but not excellent.

Combining the results of the fifteen metrics with a simple voting scheme yields a poor-performance metric with an accuracy of 89 percent (Table 2).

Placing the detection threshold between 3 and 4 yields an 89 percent success rate with no misses and one false alarm (Figure 2). As noted above, the highest success rate achieved by any of the individual

Table 2. Summary of Test Results Showing Indicators of Poor Performance

	Instability in Funding and Quantity Profiles			Spending Forecasts			Execution to Staffing Plans				Volatility (#)				
	Funding	Quantity	Magnitude	Scatter	Magnitude	Scatter	Comparison of Contract Forecasts to Funding Plans	Cost Growth on Mission Equipment	Magnitude	Scatter	Undistributed Budget	Management Reserve	Initial Investment in System Level Tasks	System Level Tasks	SE SI ST Growth at 50% Complete
CH-53K							1	1							2
BAMS							1	1				1			2
AB3 Remanufacture	1		1						1			1			5
LHA 6												1			1
AMF JTRS	1	1	1	1	1	1	1	1		1	1			1	11
FAB-T	1	1	1			1									4
E-2D AHE (Jun 2003)							1			1	1	1			4
E-2D AHE (Jul 2009)							N/A			1	------- Not Applicable --------				1
EFV	1				1			1						1	5

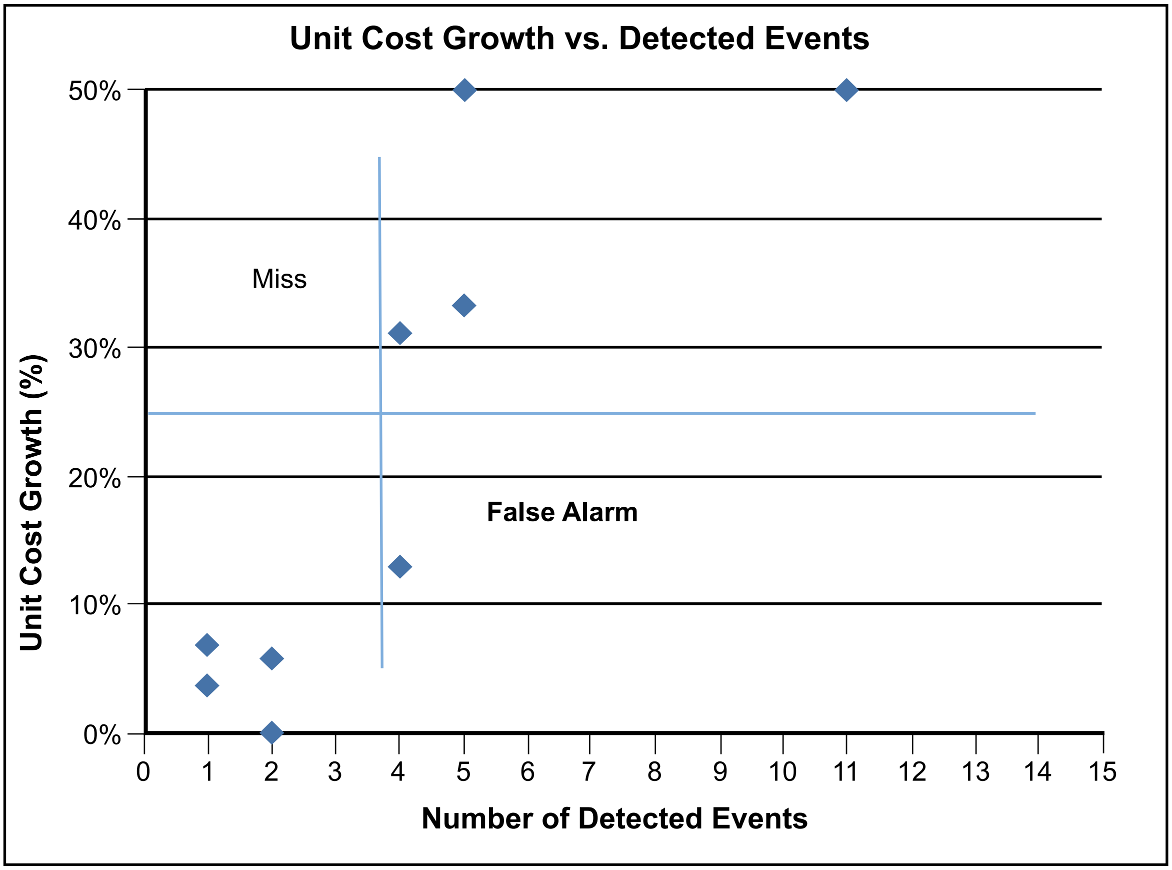


Figure 2. Unit Cost Growth vs. the Number of Events Detected

tests was 78 percent. *This result suggests that the combined test results can discriminate between those programs that are poor performers and those that are not.*

Significance tests also demonstrate that the combination algorithm was the only metric to provide sufficient evidence ($P<0.05$) to distinguish between programs that experienced UCG in excess of 25 percent and those that did not.

The concept of combining poor performance sensors to obtain improved sensor performance has a parallel in the field of radar and sensor fusion (Nicoll et al. 1991). In the early years of radar development, a graphical technique called Receiver Operating Characteristics (ROC) was used to describe how true detections and false alarms would both increase as the threshold for target declarations in a receiver was reduced. This technique recorded the sequence with which true detections or false alarms occur as the sensitivity threshold is varied from no detections to all detections.

Development of the ROC curve for the Funding Profile Instability test results is shown in Figure 3. The progression of steps can be followed by placing the poor performance threshold line (the vertical blue line in the scatter plot) to the far right at 1.0 (no detections) and moving it to

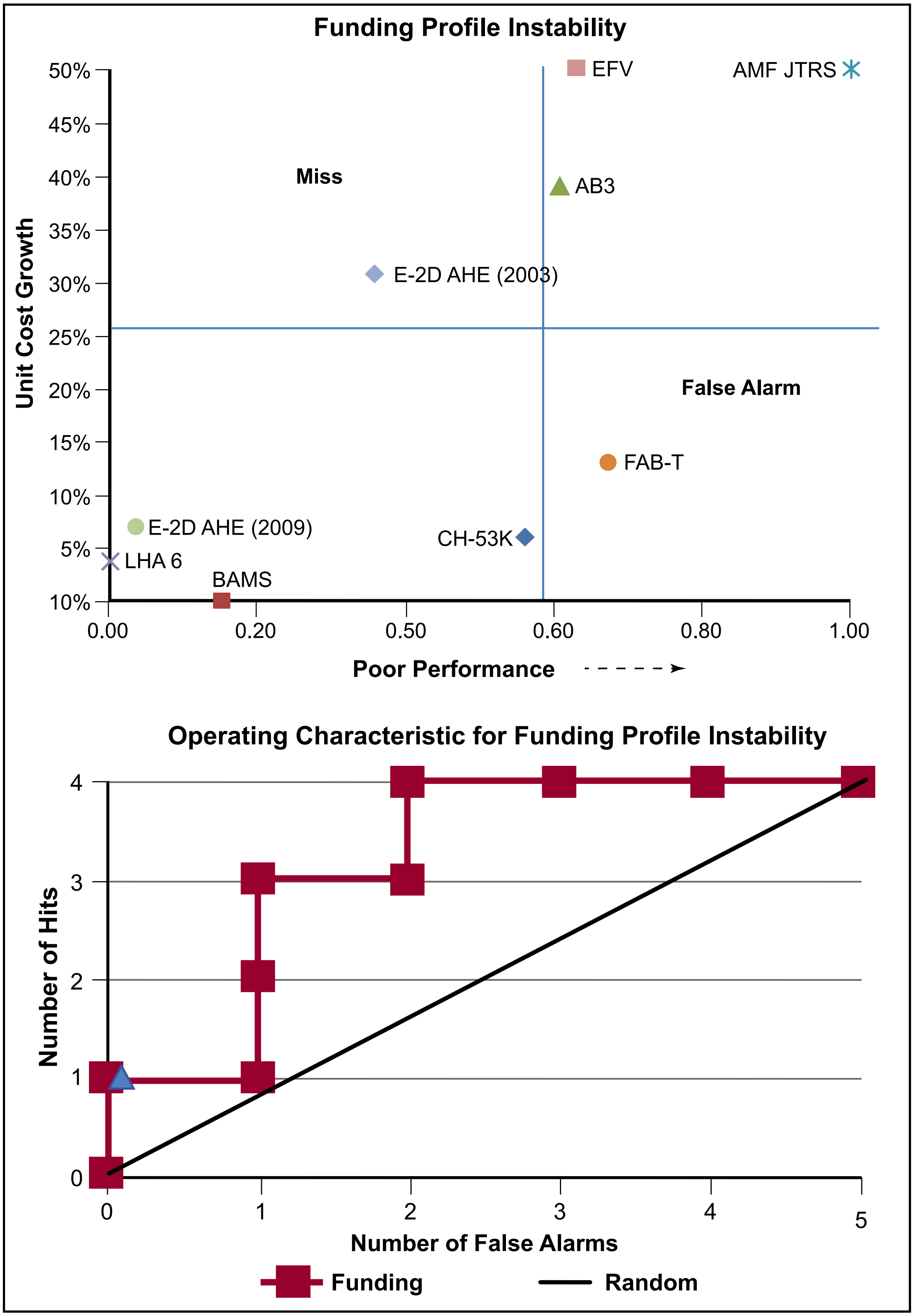


Figure 3. Operating Characteristic for the Funding Instability Metric

the left, recording +1 to the number of hits if the program marker is above the 25 percent UCG limit and +1 to the number of false alarms if the program marker is below the limit. Moving the line all the way to zero (all detections) results in four hits and five false alarms.

The "random outcome" line drawn from the origin is the expected mean for a large, normal population with a probability of detection of 50 percent (p=.5). The ROC curve for a detector with little to no value would lie close to the random outcome line. The further the ROC curve is above the random outcome line, the better the detector is at correctly identifying an event.

Development of the ROC curve for the test of Execution to Forecasts (Magnitude) metric is shown in Figure 4. This ROC curve lies closer to the random outcome line than the Funding Instability curve in Figure 3, and therefore appears to be a poorer detector than the Funding Profile Instability metric.

Figure 5 shows the ROC diagram for the combined (or fused) data in Figure 4.

This ROC diagram has two paths because the E-2D AHE (2003) and FAB-T data points have the same poor-performance index value. Either result is better than any individual test.

SUMMARY AND CONCLUSIONS

This article documents the test results for fifteen metrics of poor performance. All of the metrics use data from SARs or data from the EVM Central Repository.

These poor-performance metrics, which contribute to establishing "situational awareness" for monitoring acquisition programs, are effective tools for identifying and describing some of the problems programs encounter during the acquisition process.

The accuracy of the individual poor performance metrics in predicting UCG on the order of a critical Nunn-McCurdy breach ranges from 50 percent to 78 percent. Five of the fifteen metrics (33 percent) have a success rate between 75 and 78 percent. Combining the results of the fifteen metrics with a simple voting scheme yielded a poor-performance metric with an accuracy of 89 percent. Significance tests also demonstrate that the combination algorithm is the only metric to provide sufficient evidence ($P<0.05$) to distinguish between programs that experienced UCG in excess of 25 percent and those that did not.

The conclusion that combined results from poor detectors can exceed the detection ability of the individual detectors has a parallel in radars and sensor fusion (Nicoll et al. 1991). ROC diagrams are used to demonstrate improved performance of combined (or fused) data. This opens up the possibility that these sensor fusion techniques can be applied more broadly to MDAP-wide acquisition data in the quest for leading indicators for cost growth.

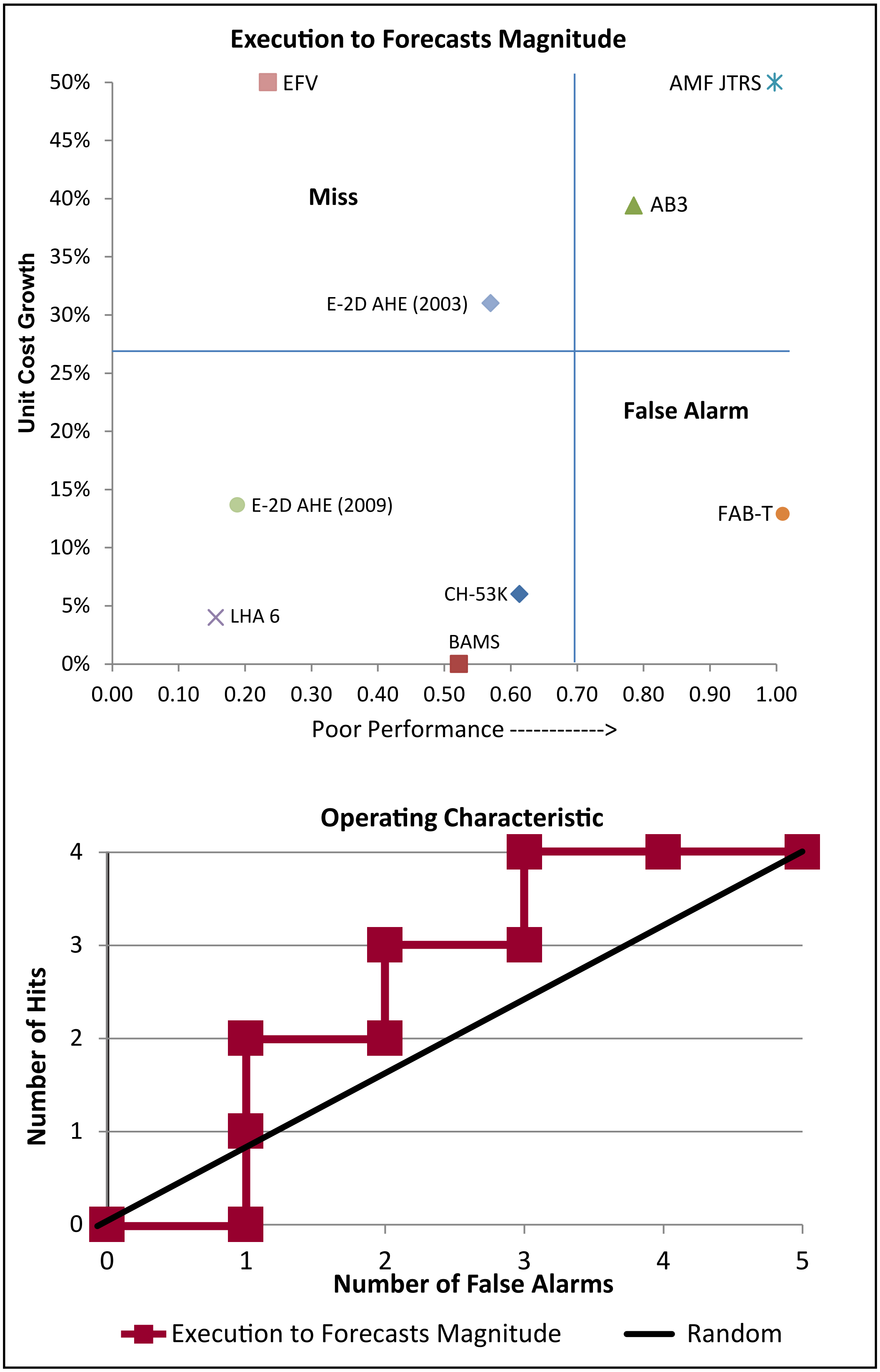


Figure 4. The Execution to Forecasts Metric

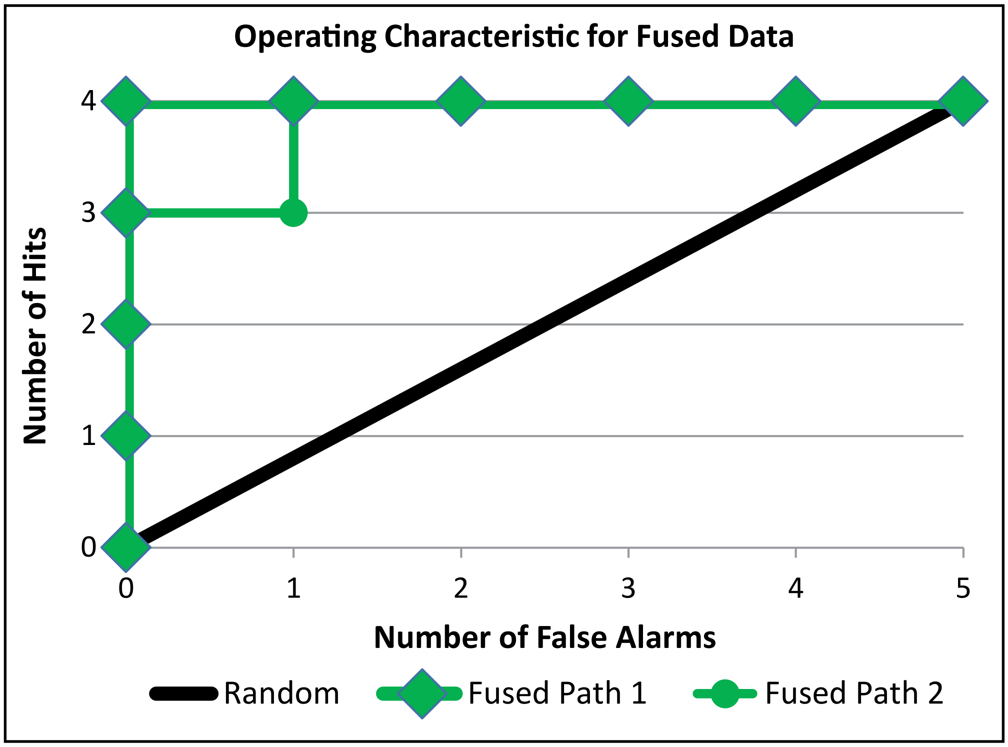


Figure 5. ROC Curve for the Fused Data

Dr. Bronson is a Research Staff Member in IDA's Cost Analysis and Research Division. She holds a doctorate in applied physics from Old Dominion University.

Dr. Sparrow is a Research Staff Member in IDA's Science and Technology Division. He holds a doctorate in physics from the Massachusetts Institute of Technology.

Reference

Nicoll, Jeffrey F., James D. Silk, and David A. Sparrow. Annual Summary Report for FY1990, Task TD-2-748, Assessment of Advanced Sensor Systems, Volume 1: Issues in Automatic Target Recognition. IDA Document D-0923. Alexandria, VA: Institute for Defense Analyses, April 1991.