



# Meaningful Metrics

## Measuring Success of Software Integration Testing Labs

Christian Hagen ■ Steven Hurt ■ Andrew Williams

The U.S. military is moving from a world dominated by advanced hardware to one of fully integrated, complex systems of both hardware and software—a move that makes it even more relevant for the military to understand how to measure and test systems with data-driven metrics and easily measurable results.

Weapon systems program offices have developed full-system and subsystem integration laboratories with the primary mission of testing and certifying integrated hardware and software during the systems' development, modernization and sustainment. These labs play a critical role in delivering a war-winning software and hardware capability to the warfighter in the battlefield. As a result, these labs have become essential to our country's defense and support of our foreign policy.

---

**Hagen** is a partner in A.T. Kearney's Strategic Information Technology Practice and is based in Chicago. He advises many of the world's largest organizations across multiple industries, including government and defense contractors. **Hurt** is a partner in A.T. Kearney's Public Sector and Defense Services Practice and is based in Dallas. He has worked with several of the U.S. Air Force's highest visibility programs to drive affordability in both software and hardware sustainment. **Williams** is a manager in A.T. Kearney's Dallas office and works with military programs to drive improvement and cost reduction in software engineering operations.

However, each lab throughout the Department of Defense (DoD) has developed its own unique processes, specific to individual programs, for measuring its progress and success. This nonstandard, and often ad hoc, approach has caused confusion among DoD program leaders about what the metrics mean, which ones matter and how to make fully informed command decisions about the software integration labs.

Without meaningful metrics that can illuminate the labs' actual performance, program leaders are unable to make even minor decisions—let alone major ones—about running an individual software lab or groups of labs within the DoD. They simply cannot manage the labs effectively. Leaders can't answer questions about whether a lab is running cost-effectively, about what a lab's efficiency is or if that level of efficiency is good or bad, or about whether to send more or less work to a particular location. Moreover, military and software leaders don't know how much money to invest in updating a lab, whether it would be best to close a lab and move the testing somewhere else or even whether buying a new piece of equipment would reduce the lab's overall costs and improve its performance. Without an appropriate approach to software integration laboratory metrics, leaders are operating the labs in the dark with little visibility on whether their decisions improve or hurt development, sustainment and modernization.

Program leaders are now making decisions with the engineering- and technology-based metrics favored by those who are far more concerned with what's needed to test, say, a third-generation radar unit than with the cost, efficiency and performance of running a lab. They have no valid metrics relevant to those who must make command-level decisions from a holistic, business perspective. Having this information on testing productivity has never been more important. We need to look no further than the F-35 program, whose software has expanded to about 24,000 source lines of code (see Figure 1). The indications are that much of the F-35's well-publicized delays are the result of its inability to test software.

Recently, the leaders of a major DoD program tried to determine what the impact on its operations would be if they moved a specific lab to another geographic location. Because they had no standard set of metrics, a new approach would be needed to make a decision based on concrete information. To determine which lab was better run, they had to significantly improve the way they looked across multiple

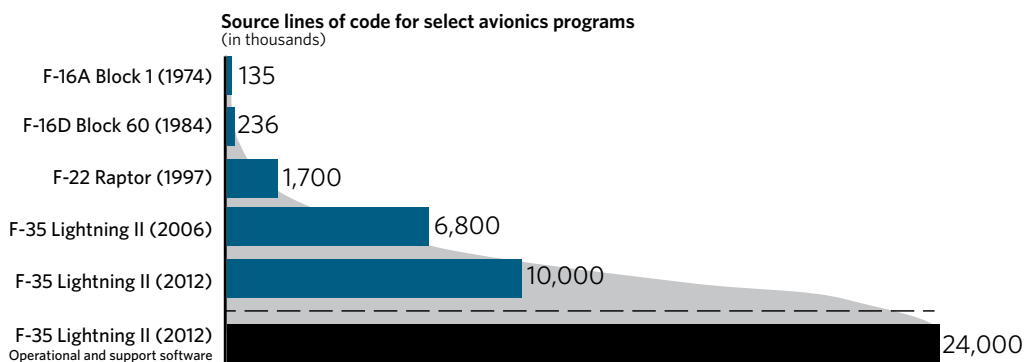
program labs to compare operating costs, performance and other key metrics. Their contractors also were unable to provide the needed metrics to compare operations—impossible, they said, because the labs used different technologies and because they tested different equipment and had completely different workloads.

Such conclusions need to be revisited, especially given the importance of software to our weapon programs and soldiers. You wouldn't tell an automotive manufacturer that it can't compare two factories because one builds compact cars and the other builds SUVs. In fact, determining the most meaningful metrics for decision makers in the software integration labs will come by examining operations with similar processes, such as the aforementioned automotive factories. These factories input parts, assemble them, and output completed vehicles. The labs input software code and hardware, run tests against the code, and put out a report on whether the code is good or bad. The processes are similar, and the metrics can be similar as well (see Figure 2).

These metrics—capacity, efficiency, effectiveness, and capability, derived from the body of work in manufacturing excellence—will enable decision makers not only to measure and improve each software lab's cost and performance but to manage all their labs effectively as they test the software systems that are fast becoming the strategic weapons on which the military's future success depends. Although these metrics are not yet completely adopted by decision makers who manage software integration labs, they are used throughout automotive manufacturing and are recognized as paramount by executives running similar operations across industries.

As manufacturing improved, a discipline known as overall equipment effectiveness (OEE) was developed to measure how effectively a process was executed. The metrics were designed to allow leaders to compare processes across factories and industries and to provide metrics that decision makers needed to understand if they were to manage their

**Figure 1. The Amount of Software in Military Avionics Systems Has Skyrocketed**



Note: Source lines of code for the F-16 and F-22 are at first operational flight. F-35 source-line data are from first test flight and from current estimates and sources. Sources: "Delivering Military Software Affordably," *Defense, Acquisition, Technology, and Logistics*, March–April 2013; A. T. Kearney analysis.

businesses and operations. The meaningful metrics for integration labs closely follow the OEE framework, with tweaks to make them more relevant for software and hardware development.



It is easy to see why these metrics are equally appropriate for measuring lab operations. The comparisons are straightforward. For capacity, auto manufacturers look at the number of cars produced per hour; labs look at the number of test points executed per hour. For efficiency, manufacturers check the number of “lemons” produced per hour; labs check the number of tests executed “on condition.” For effectiveness, manufacturers count the number of quality assurance fixes; labs count the number of software defects. For capability, manufacturers explore the functionality of their equipment and what each lab can make; labs explore the abilities of each lab to meet the overall program requirements.

These metrics can give program leaders the kind of manufacturing-environment benefits that are valuable in software integration lab measurement, including:

- **Transparency.** With a clear, communicable set of metrics, program leaders can quickly and accurately assess performance and capacity. In addition, fact-based, apples-to-apples comparisons will enable them to contrast each lab’s performance against that of other labs.
- **Cost savings.** Cost advantages between labs, which have historically been buried beneath nonrelevant metrics, will be clear when decision makers use equal, meaningful metrics that highlight cost-saving opportunities within the current environment.
- **Risk mitigation.** The metrics will take into account current and future lab capacity, allowing for more accurate estimates of cost and potential schedule delays.
- **Negotiations support.** The metrics will provide the facts on which the best negotiations are based and enable the program office to accurately size and negotiate requirements for contracting labs.

Following is a look at the four main metrics (see Figure 3).

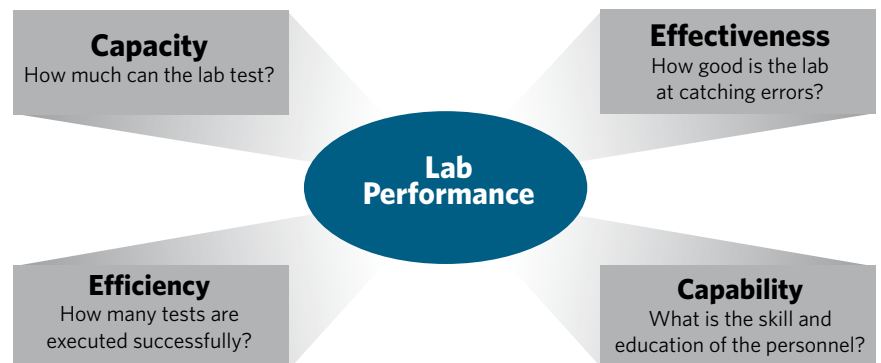
**Figure 2. The Metrics for Manufacturing and for Software Testing Labs Are Similar**

Production Metrics		
	Manufacturing	Software Integration Labs
		
Capacity	Number of cars produced per hour	Number of test points executed per hour
Efficiency	Number of good cars produced per hour	Number of tests executed on condition
Effectiveness	Number of quality fixes	Number of defects found
Capability	What can the factory produce? (for example, Porsche vs. Yugo)	What areas and complexity of tests can the lab execute?

Measured in test points, capacity is the software lab’s throughput per hour in terms of its ability to execute its raw work, which includes integration, verification and registration tests. If the lab runs 24 hours a day, seven days a week, how much work could it get done in total units?

Test points can easily be converted into derivative metrics, such as shift capacity, daily capacity and yearly capacity. As the best proxy for lab size, capacity shows whether the lab corresponds to a big or small factory. Knowing a lab’s capacity will, among other things, enable planners who are considering shifting work between labs to decide whether

**Figure 3. Meaningful Metrics for Software Testing Labs Should Follow an OEE Framework**



Note: OEE is overall equipment effectiveness. Source: A. T. Kearney analysis.

the receiving lab has the maximum capacity to handle the additional work.

Because test points are the basic unit of lab production, comparing dollars per test point is the core indicator of cost in a lab. Using this comparison, decision makers can determine, for example, how much it costs to run a test or how much it costs to find a defect—whether the defect is major (could ground an aircraft) or minor (could prevent a vehicle’s windshield wipers from working).

Efficiency is a quality metric that indicates how well the lab is doing the work. If the lab can do 100 units of work in a day but, on average, only 50 come out correct, then the lab’s efficiency metric is quite low.

This measurement of the lab’s testing procedure shows how many tests must be run before the lab starts finding errors in the testing procedure. The accuracy of this metric depends on several issues, including the quality of the code being input into the lab.

Capability is the skill set of a lab’s workforce and the functionality of its equipment. Capability is used to compare how well each lab can test specific areas of the software and is the result of three factors:

- **Knowledge** is assessed across product, functions, and technology, and is proven through work experience requiring expertise in the product, function and technology areas.
- **Competency** is assessed across current work behaviors

## A business case analysis such as the one done for the DoD can capture a series of deliverables that help leaders better manage their labs and make cost-saving changes that do not hinder the capabilities.

Efficiency is measured with the on-condition metric. “On-condition” is defined as a test executed successfully, according to the checklist and setup procedures handed down by the system engineers, that does not need to be repeated. Efficiency measures the percentage of tests executed correctly—not whether the software being tested passed or failed the test—and is calculated by dividing test points on condition by total test points attempted. “Off-condition” is defined as a test that must be performed again because of an error in testing methods or setup. A false on-condition test is properly executed on condition, but further analysis shows the test package was poorly designed, so the test must be repeated.

Lab capacity and efficiency are tightly linked and are often measured together to provide a clear understanding of their combined effect. Baselines derived from this combination enable leaders to begin making command-level decisions about questions such as how a given action would change the lab’s throughput, how a different action would affect the lab’s cost per hour or cost per defect and how yet another action would impact the lab’s efficiency or capacity.

Effectiveness points out how good the lab is at discovering errors. If an integration lab’s primary purpose is to find defects or certify code, the ratio of work units to defects could be a measure of effectiveness. Effectiveness is measured by the number of test points executed per defect found and is calculated by defect found divided by test points attempted.

and skills required to perform the work and proven by the existence of artifacts, such as current job descriptions and training, which are used to validate managers’ and directors’ scores for their teams and specific knowledge areas.

- **Capacity** is measured by the availability and readiness of the lab’s resources (human and infrastructure) to perform an activity.

Because capability is also directly affected by the lab’s equipment composition, this composition must be analyzed in any lab-to-lab comparison.

Capability plays a major role in the program leaders’ overall management decisions because it has an implicit effect on the other three meaningful metrics. Therefore, its impact on each of these metrics must be understood before making changes to the size, experience or skill set of the workforce.

### Meaningful Metrics for DoD

These meaningful metrics for software integration labs were recently used for a DoD laboratory that tests large, complicated systems. The lab had a complex software- and system-testing environment that lacked performance transparency.

The meaningful metrics were developed during the assessment to enable appropriate comparisons across the current lab footprint, which spanned multiple sites with differing approaches to software integration testing. They provided the necessary method to accurately measure and compare lab

performance across the footprint and were essential to the DoD.

In essence, the metrics drove the study, allowing the direct lab comparisons needed for the analysis. With them, the team created a business case to model future scenarios and compare cost savings, transition risks, and steady-state capacity risks across scenarios.

### Approach

The assessment objective was to evaluate the current strategy for software integration labs and explore alternative models that might deliver better value. Specifically, the assessment was designed to reduce the life-cycle costs of the labs by moving testing from its current lab to potential alternatives and to do so without degrading current performance.

It also was designed to answer four key questions:

- What are the key attributes of the current lab footprint?
- What are the proposed alternatives to the current lab environment?
- What are the costs, benefits and risks of the current plan and the proposed alternatives?
- What is the recommended strategy (current plan versus proposed alternatives)?

The objective was met with a thorough analytical review of the current long-term strategy and potential alternatives and was shaped by qualitative insights gained during the assessment. The best value alternative would result in the lowest life-cycle cost with manageable risk while not degrading lab capabilities or performance.

### Results

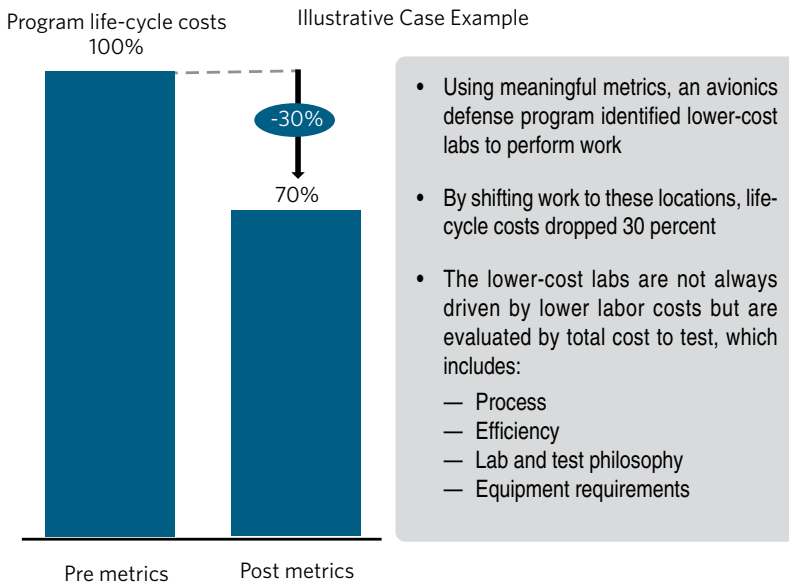
The team recommended that the DoD transition testing from its current lab to alternative labs while maintaining the same performance and the same operator and equipment capability as the current lab, resulting in less risk during transition and normal operation.

The recommendation would also reduce program life-cycle costs by more than 30 percent, for a total net present value savings of hundreds of millions of dollars (see Figure 4).

The team also created clear, communicable metrics that would reflect lab capacity, efficiency, effectiveness and capability—and allow leadership to manage the labs more effectively.

Finally, the team modeled various courses of action from the present day through the perceived end of life. And it recommended a clear course of action for the transition,

**Figure 4. Focusing on Meaningful Metrics Can Reduce Life-Cycle Costs**



Source: A.T. Kearney analysis

including the expected cost savings, transition risks and operational risks.

### Where Can This Help?


A business case analysis such as the one done for the DoD can capture a series of deliverables that help leaders better manage their labs and make cost-saving changes that do not hinder the capabilities. Potential deliverables include:

- **As-is baseline:** an evaluation of the current-state capacity, efficiency, effectiveness and capabilities of the software integration labs and the development of relevant metrics that will allow further insights into the lab footprint
- **Cost-saving opportunities:** analytical comparisons between labs revolving around the proven metrics, the ability to quantify and estimate previously hidden proficiencies and the generation of plausible future-state scenarios
- **Scenario modeling:** analytical modeling for each of the potential variables; sensitivity, tipping point and worst-case analysis around key input variables; and risks to schedule and the estimated costs to mitigate schedule delays

Software labs, which are expensive and vital DoD assets, often suffer from testing overruns, under-deliveries on initiatives, and intricate projects that make software testing and laboratory management complex. Meaningful metrics can reduce or resolve such problems. These metrics are relevant to a number of different applications faced by software and lab program managers, who might want to consider refreshing their lab performance metrics to realize several objectives in the areas of lab performance, transparency, monitoring and continuous improvement.

Additionally, these meaningful metrics will give DoD technology leaders the information needed to develop a baseline of their current operations, with which they can put in context their decisions and the impact of those decisions. With this baseline, they will know the effect of making small changes, such as how adding capacity will affect a lab's costs, how reducing costs will change the lab's capability and efficiency, and how hiring employees with different skill sets will change the on-condition efficiency. They will know the effect of command-level decisions, such as those they must make when answering questions about whether the labs are effective, whether they have talent or skill deficiencies, or whether significant changes need to be made to improve overall software testing. And, what is perhaps most important, they will know

whether their throughput and quality meet the demands of the DoD and individual defense programs.

Finally, these metrics will cut through the confusion that leaders now feel and give them the concrete measures they need for making decisions, not just on technical performance and operations but on fiscal performance. As the DoD's capabilities in developing software and integrated systems mature, these metrics will become even more vital in the department's overall effort to drive efficiencies and savings in its programs to give the warfighter the best, most advanced systems available anywhere. 

The authors can be reached at [christian.hagen@atkearney.com](mailto:christian.hagen@atkearney.com), [steven.hurt@atkearney.com](mailto:steven.hurt@atkearney.com) and [andrew.williams@atkearney.com](mailto:andrew.williams@atkearney.com).

## Buying What Works Case Studies in Innovative Contracting Released

The first version of *Innovative Contracting Case Studies* was released Aug. 21 by the White House Office of Science Technology Policy (OSTP) and the Office of Management and Budget's Office of Federal Procurement Policy (OFPP). "*Innovative Contracting Case Studies* is an iterative, evolving document that describes a number of ways federal agencies get more innovation per taxpayer dollar under existing laws and regulations," according to a joint OSTP-OFPP announcement.

"For example, NASA has used milestone-based payments to promote private sector competition for the next generation of astronaut transportation services and moon exploration robots," the announcement stated. "The Department of Veterans Affairs issued an invitation for short concept papers that lowered barriers for nontraditional government contractors, which led to discovery of powerful new technologies in mobile health and trauma care. The Department of Defense has used head-to-head competitions in realistic environments to identify new robot and vehicle designs that will protect soldiers on the battlefield."

Over the years, there has been much progress on helping federal agencies gain greater access to the innovation and synergies generated by the commercial marketplace. Still, the standard procurement processes on which agencies rely to meet most of their needs may remain highly complex and enigmatic for companies that are not traditional government contractors. Many such companies can offer federal agencies valuable new ways of solving longstanding problems and cost-effective alternatives for meeting everyday needs.

As budgetary constraints continue to reduce available resources, the need increases for new innovative contracting models that can help agencies reach these entrepreneurs and reduce the complexity and cost of doing business with the government. "Such tools allow federal agencies to pay contractors for results, not just best efforts," the announcement stated.

The document stated that the OSTP and OFPP "seek to encourage greater innovation in federal contracting. ... OSTP compiled the collection of agency case studies to highlight different models that have been successfully tested by agencies to meet a range of needs related to research, prototyping, and market testing."

In the joint announcement, officials of OSTP and OFPP said: "We encourage both private sector stakeholders and public servants to engage in a sustained public discussion, identifying new case studies and improving this document's usefulness in future iterations. At the same time, federal government employees can join a community of practice around innovative contracting by signing up for the new 'Buyers Club' e-mail group (open to all .gov and .mil e-mail addresses). This 'Buyers Club' group should provide a useful forum for troubleshooting and sharing best practices across the federal government, serving everyone from contracting officers with deep expertise in the Federal Acquisition Regulation (FAR) to program managers looking for new ways to achieve their agencies' missions."

Note that OSTP compiled these case studies based partly on feedback from external experts, and that the *Innovative Contracting Case Studies* document does not necessarily reflect the views of the federal departments and agencies that are cited as examples. The availability and use of different innovative contracting methods will require consideration of legal authorities and the desired outcome/goals of the specific activity, the study cautioned.

### See:

- <http://www.whitehouse.gov/blog/2014/08/21/buying-what-works-case-studies-innovative-contracting-0>
- Summaries: Find summaries of programs collected at the following URL:
  - [http://www.whitehouse.gov/sites/default/files/microsites/ostp/innovative\\_contracting\\_case\\_studies\\_2014\\_-\\_august.pdf](http://www.whitehouse.gov/sites/default/files/microsites/ostp/innovative_contracting_case_studies_2014_-_august.pdf)