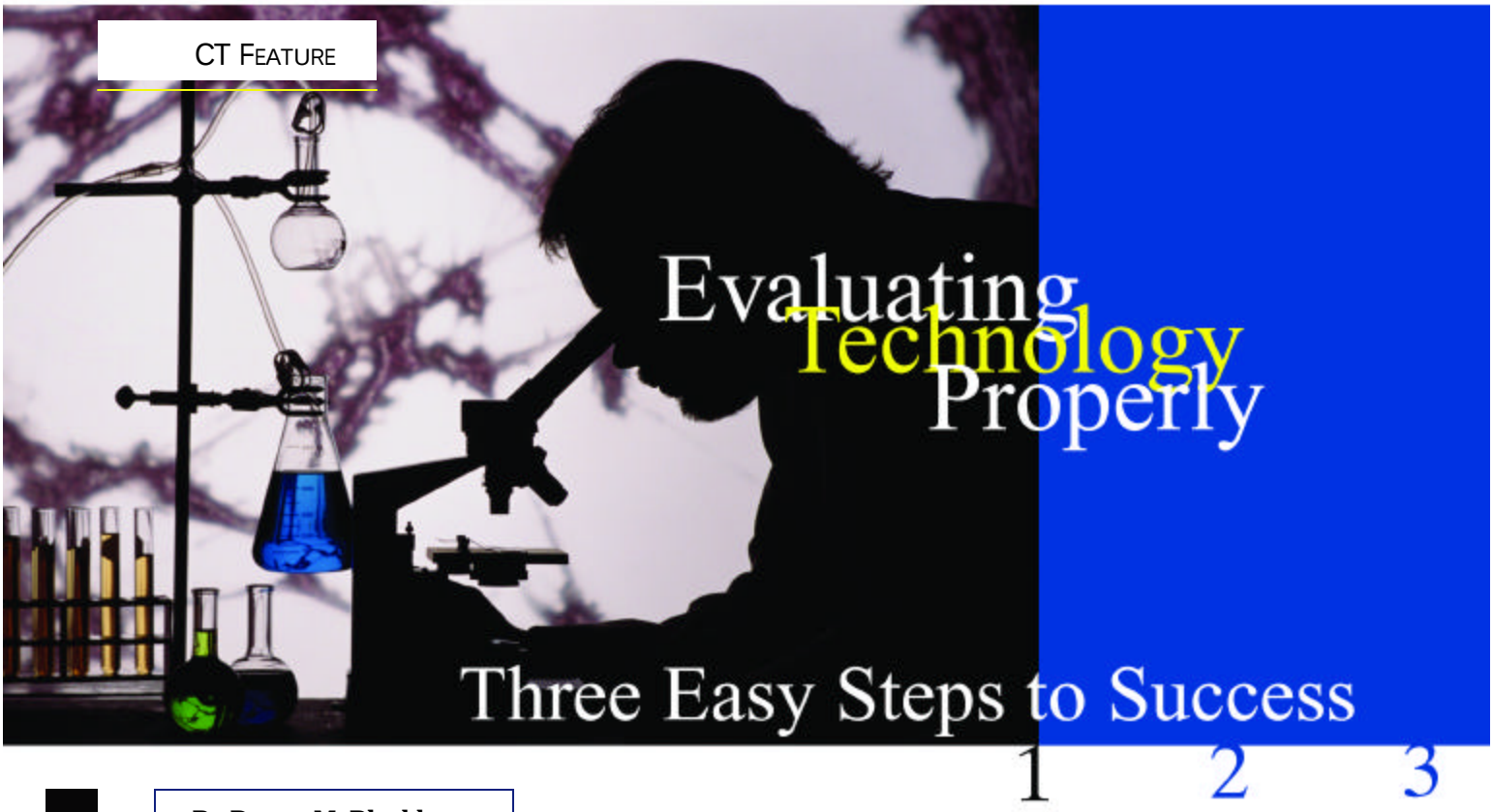

# Evaluating Technology Properly

## Three Easy Steps to Success

1 2 3

**By Duane M. Blackburn**

As a program manager for federal agencies devoted to developing technology, I often encounter the evaluation dilemma. Whether a project is in the concept, development, demonstration or commercial phase, the question of how to properly evaluate the technology always arises. Frequently, evaluations are proposed simply so someone can say the technology was evaluated and was successful at whatever it was intended to do. Unfortunately, not much thought goes into these evaluations, which makes their usefulness extremely limited. This article analyzes the thought process and steps necessary to turn the typical evaluation into a thorough and widely usable evaluation.

## How a "Smoked Pig" Will Make Or Break Your Evaluation

Growing up, I was very involved in the Boy Scouts and I was elected senior patrol leader when I was 14 or 15. The senior patrol leader was responsible for planning and executing all activities, with guidance given by the scoutmaster and his assistants.

Each fall, our Boy Scout troop had a pig roast and awards banquet. The adults took care of roasting the pig, but the planning was up to the senior patrol leader. Until that point, my management approach was very basic — delegate simple tasks to those who I thought could do it and then do everything else myself. That is what happens when you cross a workaholic with an "if you want something done right, do it yourself" mentality. No need for advanced planning, I thought, since I worked better under pressure anyway and always had gotten the job done in the past. However, an assistant scoutmaster had other plans.

His lesson started with the simple phrase: "Proper planning prevents poor performance." This was a big change from my regular thought process. He showed me that to prepare for a flawless pig roast, we first had to determine what we wanted the result to be. Then we could take gradual planning steps back to determine our course of action. We had a menu in place and knew which of these foods we wanted to serve hot, warm and cold.

Next, we had to plan the preparation steps so we could serve the food at the correct temperatures. In this example, the planning steps were simple. For example, corn took X minutes to get from the stalk to being ready to serve and bread took Y minutes to get from bags of flour to piping-hot dinner roles. Of course, the planning was bound to get more complicated. Unless I wanted to be responsible for a bunch of boys cooking on more than three fires at a time, I had to decide which foods to cook first and which ones to put aside. It took the assistant multiple meetings to convince me that this was the proper planning approach, and I did not fully believe him until the successful, stress-free pig roast was over.

Often, I have wondered if this assistant developed this approach himself or if he had derived it from studying the Toyota Production System's Just in Time (JIT) approach to automobile assembly that was gaining popularity at the time. A friend recently told me about a television sitcom in which a main character decided she wanted to have her first child before reaching 35. She then unconsciously used a JIT-like approach to learn

that to meet her goal, with all the goals she wanted to attain before she had the child, she needed to get married within the next couple months — not the best method to decide one's soul mate, but an interesting example.

As countless American companies and that sitcom character, learned, you cannot use someone else's approach (Toyota's JIT model) and expect it to be successful for every application. Nevertheless, it has been my experience that applying a JIT-like approach to planning an evaluation of technology is successful.

So, what do we want our technology evaluations to show? What are the evaluation objectives? In the pig roast example, the objective was properly heated food served at the correct time. For the sitcom character, the objective was to have a child by age 35. If you are a vendor wanting to sell products, the objective is to show that your product works better, faster and for less money than any other competitor's products. Hopefully we, as developers and practitioners, do not fall into the trap of designing evaluations with this desired objective. What we should want to know is whether a technology works and why. We should not simply want a test that ranks specific products. We should want an evaluation that also is of sufficient depth to understand why the products were ranked in that manner. We also want to be able to show areas of strength and weakness so we understand where we can deploy the current technology and where future development efforts should be focused. Now that we understand what our desired objective is, we can look at several evaluation ideals and an evaluation structure that will help us meet that objective.

## Ideals to Follow for a Successful Evaluation

When I was establishing the evaluation methodology for the Facial Recognition Vendor Test (FRVT) 2000 (http://www.dodcounterdrug.com/facialrecognition), I was fortunate to have been given a draft version of an article that eventually was published in the February 2000 edition of *IEEE Computer*, titled "An Introduction to Evaluating Biometric Systems." The ideals presented in this paper impressed me, and I developed the FRVT 2000 Evaluation Methodology around them. Although this paper was tailored to evaluating biometric technologies, we can use the ideals presented in it to evaluate any type of technology.

The first ideal presented in the paper is that successful evaluations must be administered by independent groups that will not reap any benefits should one system outperform the other. If this is not so, conflicts of interest, even if only perceived, will cast significant doubt on the validity of the evaluation results. Sometimes you must study these "independent groups" carefully as they may be funded by vendors or entities with alliances to vendors.

A second ideal is to use test data that none of the systems being tested have previously seen. System developers are very smart. After all, they are the ones who developed the new technology you want to evaluate. They also are excellent marketers. Otherwise, you would not be interested in their systems. You can bet that these system developers will learn the properties of previously seen test data and tune their systems for maximum performance.

A third ideal is that the evaluation itself must not be too easy, nor too difficult. If the evaluation is too easy, all the systems will perform well and will group together at one end of the capabilities spectrum. If the evaluation is too difficult, none of the systems will perform well and will group together at the other end of the spectrum. In either case, the evaluation will fail to produce results that will enable you to accurately distinguish one system from another.

A fourth ideal is that the evaluation itself must be repeatable and made available to the technical and practitioner communities. Repeatable does not necessarily mean the same test data and the same test results, but a comparable test that will statistically return the same results. An excellent example of this is the SAT taken by high school students. There are multiple versions of this test given every year, but any one student should expect to receive approximately the same score on any of the tests. To be repeatable in our technology evaluations, we must document all phases of the evaluation including the gathering of test data, evaluation protocol, testing procedures, performance results and examples of test data.

There are two reasons to document all these phases. The first is so the technical and practitioner communities accept the validity of the evaluation. More vendors will be willing to participate in an evaluation if the process is described beforehand. They will know that those performing the evaluations understand what they are doing and, thus, will not be afraid of having to answer questions about a poorly designed evaluation. The second reason to document these phases is so evaluators and other readers can accurately determine how each presented result was obtained. Irregularities in test results often can be explained via a thorough analysis of the test protocol. Documenting the results along with the test protocol also allows others to improve the evaluation protocol for future evaluations. If the cycle continues, you will be able to continually reap the benefits of evaluations of which you were not even a part.

A final ideal, which I did not take from this article, is that you must understand the true requirements for your application to learn whether results from any evaluation show that a technology investment is warranted. There is a difference between true requirements and desired requirements and, typically, these are incorrectly mixed. For example, assume that you currently are achieving 15 percent to 20 percent on some measurable and that you have a desired requirement of 90 percent once you install the new technology. Now assume that your evalua-

tions show that you should expect to see a score of 70 percent with the new technology. The technology is a failure, right? Not necessarily — deeper analysis could show that an improvement of 30 percent to 40 percent on this measurable makes up for the costs incurred due to the addition of technology. This would be your true requirement. The benefits of the new technology far exceed the true requirement and would be a huge success, even though it did not meet the desired goal.

## Three Steps to a Complete Evaluation

The article, "An Introduction to Evaluating Biometric Systems," provides a structured approach to a complete evaluation that moves from the general to the specific through three major steps: a technology evaluation, a scenario evaluation and an operational evaluation.

The most general type of evaluation is a technology evaluation, the goal of which is to learn the underlying technical capabilities of a particular technology. The testing is performed in laboratories using a standard set of data that a universal sensor collected. In the vast majority of technologies, the same data can and should be used as input for each system. Technology evaluations usually are reproducible and typically take a short time to complete, depending on the type of technology being evaluated.

The next step in the structured evaluation approach is a scenario evaluation, which aims to evaluate the overall capabilities of the entire system in a specific scenario, rather than a subset of the system in technology evaluations. For example, in evaluating facial recognition systems, the technology evaluation would study the face recognition algorithms only, but the scenario evaluation studies the entire system, including camera and camera-algorithm interface, in a given scenario. When evaluating drug detection devices, the technology evaluation would determine the minimum level of detection capabilities for each of the device types, and the scenario evaluation studies how well the entire system performs for a specific scenario. In a scenario evaluation, each tested system would have its own acquisition sensor and would, thus, receive different data. Consequently, scenario evaluations are not always completely reproducible, but the approach used can always be completely repeatable. Scenario evaluations typically take a few weeks to complete because multiple trials — and for some scenario evaluations, multiple trials of multiple subjects/areas — must be completed.

The most specific step in the structured evaluation approach is an operational evaluation, which is very similar to a scenario evaluation except it is performed at the actual site using the actual subjects/areas. Operational evaluations usually are not reproducible unless the operational environment naturally creates reproducible data. Operational evaluations typically last from several weeks to several months.

The three steps described in this structured evaluation approach not only flow from the general to the specific, but also flow from one to

another. Technology evaluations are performed on all applicable technologies that could conceivably meet your requirements. Results from the technology evaluations will be of immediate interest to the vendors as well as the evaluators. The technology evaluation results will provide the vendors a direction toward what developments they will need to undertake to improve their product. It also will help the evaluator determine which, if any, of the technologies could match your stated requirements now. The evaluator then can select a subset of these technologies for a scenario evaluation. Once the scenario evaluation has been completed, the evaluator can select one, or possibly two, systems for an extended operational evaluation at the actual site. If the operational evaluation is successful, the evaluator then can decide to implement the technology permanently onsite.

There are multiple reasons not to skip the technology evaluation and go straight for a scenario evaluation, or even worse, an operational evaluation. The first is that you would be selecting technology based on a whim rather than scientific analysis. Are any of us truly wanting to explain to our superiors why we spent so much money doing field tests of a specific vendor's product without first studying how it compares to competing systems? Another reason is that if a scenario evaluation is successful, or fails, we will not truly understand why unless we have the technical information from the technology evaluations to analyze with the scenario evaluation results. An example is the scenario evaluations in the FRVT 2000 evaluations. There were two different, yet similar, scenarios tested and the results from each evaluation varied widely. There were two main variances in the setup of the scenario evaluations. By looking at data from the technical evaluation, I could determine that one of the setup variances did not contribute to the result variances. This indicated that the other variance was the likely culprit. Without having the data from the technology evaluation, I would not have determined where the limitation in the systems occurred.

## Who Should Perform The Evaluations?

Who should perform these evaluations — technologists or practitioners? The answer is that both should be involved in all three phases of the structured evaluation process, but the level of participation of each varies for each phase. Technology evaluations should be performed by technologists who are experts in the subject field. These technologists, however, need to understand the community's objective for the evaluations so they can tailor the technology evaluations to deliver the data needed to establish a scenario evaluation. Only practitioners can provide this insight. Scenario evaluations require equal

input from both technologists and practitioners. The practitioners need to develop the scenario so it resembles the activity envisioned, while the technologists need to develop the test protocol so that useful data can be found for analysis. Practitioners should perform operational evaluations but, again, technologists should assist and advise the testing and evaluation aspects of the evaluation.

## How to Share Results

After you have set up, performed and documented your evaluations, you will have a thorough knowledge of the capabilities of the technologies that could be beneficial for your situation. This knowledge is of limited value, however, until you share it with those outside your evaluation group. We should make every effort to make all our evaluations available to the widest possible audience by making them available for release to the public.

The federal government has established several free ways to do this. The first is to provide your evaluation documentation to the National Law Enforcement and Corrections Technology Centers (NLECTC) (www.nlectc.org). NLECTC is a program of the National Institute of Justice that provides criminal justice professionals with information on technology, guidelines and standards for these technologies, objective testing data, and science and engineering advice and support to implement these technologies. If you contact NLECTC with evaluation documentation, the organization will work with you to distribute it throughout the community.

You also may place a copy of your evaluation documentation in the Counterdrug Technology Information Network (CTIN) (http://www.ctin.com). CTIN is sponsored by the Department of Defense Counterdrug Technology Development Program Office and serves as a location to freely share information about technologies that are applicable to the counterdrug efforts of federal, state and local governments.

## Conclusion

Whenever we are faced with a decision about new technology, we always want to have a clear understanding of how well it will assist us in our efforts. By performing an evaluation that follows the ideals and structure presented in this article, you, and others who read your evaluations, will obtain that desired understanding and will be able to successfully field the technology.

### REFERENCES

Phillips, P.J., A. Martin, C. Wilson and M. Przybocki. 2000. Introduction to evaluating biometric systems. *IEEE Computer*, 56-63. Piscataway, N.J.: IEEE Press (February).

*Duane M. Blackburn is a program manager for the National Institute of Justice and the Department of Defense Counterdrug Technology Development Program Office.*