Taking the Next Step: Improving the Science of Test in DoD T&E

Dr. V. Bram Lillard, Research Staff Member at the Institute for Defense Analyses (IDA) and Advisor to the Director, Operational Test and Evaluation (DOT&E)

Laura J. Freeman, Research Staff Member at the Institute for Defense Analyses (IDA)

Abstract

The current fiscal climate demands now, more than ever, that test and evaluation (T&E) provide relevant and credible characterization of system capabilities and shortfalls across all relevant operational conditions as efficiently as possible. In determining the answer to the question, "How much testing is enough?" it is imperative that we use a scientifically defensible methodology. Design of Experiments (DOE) has a proven track record in Operational Test and Evaluation (OT&E) of not only quantifying how much testing is enough, but also where in the operational space the test points should be placed. Over the last few years, the T&E community has made great strides in the application of DOE to OT&E, but there is still work to be done in ensuring that the scientific community's full toolset is utilized. In particular, many test programs have yet to capitalize on the power of the test design when conducting the data analysis. Employing empirical statistical models (e.g., regression techniques, analysis of variance (ANOVA)) allows us to maximize the information from every data point, resulting in defensible analyses that provide crucial information about system performance that decision-makers and warfighters need to know. DOT&E will continue to work to ensure the highest technical caliber in every DOT&E evaluation, and that Test and Evaluation Master Plans (TEMPs) are adequate to support these robust evaluations. As we improve in our use of these test designs and analysis methods, we need to ensure these practices are institutionalized across the entire T&E community and applied across all phases of DoD testing.

Introduction

In all T&E activities, the evaluation goal and supporting data analysis should drive the determination of test adequacy. More simply put, the evaluation goals should drive the testing. Operational testing should be adequate to characterize system capabilities and shortfalls across all relevant operating conditions. Such full characterization ensures that fielding decisions are made with a clear understanding of system performance, since it is not cost effective to field weapon systems that do not work or provide no clear improvement over existing systems. Full characterization also enables testers to inform the warfighters, whose lives depend upon these systems, about what these systems can and cannot do.

The common test design approaches of the past, such as specialized/singular combat scenarios, changing one test condition at a time, and avoiding any control over test conditions, lack the scientific rigor needed to ensure that testing is both efficient and effective to characterize modern system performance. Especially for complex modern systems for which performance often depends on complex interactions of operational conditions, older test strategies provide data inadequate to support characterization of system capabilities.

Additionally, the current fiscal environment demands that testing be scrutinized to ensure that all data collected provide necessary and useful information in the most efficient manner. In my experience, DOE has provided a rigorous and robust methodology for quantifying the right amount of testing. Gauging the right amount of testing is more than simply determining the number of test points; equally important is the placement of those points across the operational envelope. The placement of the points is the most important aspect of determining whether the testing will be adequate to support the goals of the analysis.

The mandate to apply DOE to operational test and evaluation is not new. In 2009, the Operational Test Agencies (OTAs), in collaboration with DOT&E, endorsed the use of DOE methods in DoD testing. Furthermore, several National Academy studies have advocated that the testing community take full advantage of the benefits available from the use of state-of-the-art statistical methodologies. More recently, Dr. J. Michael Gilmore, the Director, OT&E, outlined his expectations for the application of DOE in OT&E in 2010. In 2012, the Deputy Assistant Secretary of Defense for Developmental Test and Evaluation (DASD (DT&E)) endorsed a scientific approach to testing by including DOE methods in the Scientific Test and Analysis Techniques (STAT) T&E Implementation Plan. STAT rightfully includes DOE and emphasizes the need for corresponding analysis techniques that DOT&E has regularly emphasized as well. Over the last few years, the testing community has developed best practices in applying the methods that I summarize briefly here. Although much progress has been made, many areas can improve, particularly the area of data analysis.

Rigorous methodologies in both test design *and* analysis are crucial for T&E. A sound DOE is not beneficial unless we employ the appropriate corresponding analysis techniques. In the past several years, DOT&E has identified many best practices in the design and execution of operational tests. The OT&E community has come a long way with respect to ensuring rigorous OT&E test plans. However, we have fallen short on capitalizing on the test design during data analysis and using the full STAT toolbox. The best characterization of performance across the operational space will result from employing empirical statistical models (e.g. regression techniques, analysis of variance (ANOVA)). These empirical models allow us to maximize the information from every data point, resulting in efficient tests and defensible analyses.

In addition to outlining some of these important methods, this editorial concludes with a discussion of the need for institutionalizing these rigorous methodologies across all DoD testing. I provide a summary of efforts underway to make this goal possible.

Best Practices

For the past several years, operational testers across all of the Services have been expanding their use of experimental design techniques to scope operational tests to support the characterization of mission capabilities across the operational space. Over time, many best practices have been identified. Following these best practices has resulted in test plans that are both efficient and capable of determining performance across a variety of operational conditions.

Clear test goals

First and foremost, it is essential to have clear goals for testing. If the goals are not right at the beginning of the test design process, then no statistical technique or tool can salvage the results. In operational testing, the goal is most often to characterize system capability across a variety of operationally relevant conditions. Even a comparison test between a new system and a legacy system is a subset of this goal, since we seek to characterize the new system's performance across multiple operating conditions relative to the old. Test goals should not be limited to verifying requirements under limited sets of conditions. Given the choice between expanding the number of test conditions or replicating a single set of conditions, I would recommend choosing an expanded set of test conditions in almost every case. Statistical models allow us to draw conclusions about the reproducibility of test outcomes even in the absence of replication. The misconception that we need large numbers of replications under a given condition to make statistical claims about system performance under those conditions stems from the lack of understanding about the power of statistical analysis methodologies. We need to embrace the power (both statistically and literally) of statistical analysis techniques.

Quantify the justification for the test design

Testers need to provide clear, analytically based justification for all designs. Every test design requires the quantification of acceptable risks and a determination of what differences in performance (effect size) need to be captured. Single-hypothesis statistical tests and their corresponding statistical power calculations are generally inappropriate for sizing operational tests because they do not provide the ability to characterize performance across the operational envelope, nor do they provide insights on the placement of test points within the operational envelope. Additionally, quantified risk estimates need to be anchored in terms of what changes in performance are operationally meaningful and in the expected scatter or variability of the data across the test conditions. Leveraging existing system and developmental test data can provide defensible justification for these challenging aspects of test planning. Operational test designs have the greatest chance of succeeding if testers can leverage all existing data (particularly developmental test data and legacy system data) on the system and its intended employment.

Select continuous metrics

One of the most important decisions we can make in terms of test efficiency is to select continuous metrics where possible, since they provide the maximum information from a given test size. Continuous metrics can enable 50 percent (and likely greater) reductions in test size over comparable pass/fail metrics for similar test goals.

Include all relevant factors

It is important to include all relevant factors (cast as continuous where possible) formally in the test design. By selecting relevant test factors and forcing purposeful control of those factors we can ensure that the operational test covers conditions the system will encounter once fielded.

Leveraging developmental test data is essential for narrowing the list of relevant factors and mitigating the risk of excluding important factors. Omitting known important factors from the test design results in holes in our knowledge of system performance. When resources are highly constrained we should leverage advanced design techniques coupled with developmental testing to ensure we can incorporate as many factors in the test design as possible.

Use the full statistics toolset

Finally, we should use all of the statistical tools (measures of merit) at our disposal to ensure that testing is adequate, including:

- Statistical confidence and power – we should ensure that power calculations, which are typically used to determine the size and scope of a test, are consistent with test goals. Since we are interested in determining how system performance differs across the operational envelope, power calculations based on a single hypothesis test are not useful. Additionally, power curves illustrate the analytical trade-space between resources and risk.
- Scaled prediction variance and other measures of variance in the operational space – These measures can be particularly useful in determining the placement of test points across the operational space.
- Correlation and alias matrices – These measures ensure that once a test is complete, the data will enable the evaluator to diagnose the root cause of changes in performance across the operational envelope, including synergistic (interaction) effects. These measures ensure we avoid confounding effects so that we do not wrongly attribute system performance problems to incorrect causes.

Areas for Improvement

While significant progress has been made in recent years, there is still work to be done in ensuring that the scientific community's full toolset is utilized in T&E. There has been varying degrees of quality in the application of DOE and the best practices identified above. In addition to implementing these best practices, DOT&E has highlighted several areas where further improvement could be realized by almost all test programs:

- Statistical analysis methods to analyze test data
- Advanced analysis methods to meet unique T&E challenges, and
- Well designed surveys in OT&E.

Statistical analysis method to analyze test data

Although most organizations in the OT&E community are now using statistical rigor to develop test designs, they are not always following up with the same rigor in the analysis of the data. The worst case of this is where a test is designed using DOE techniques to cover the important operational conditions efficiently, yet the data analysis is limited to reporting a single average (mean) across the test conditions. Comprehensive statistical analyses will take advantage of the efficiencies and increased information provided by a rigorous experimental design. We must employ standard statistical tools, such as regression analysis techniques, that utilize all of the factors that affect system performance (meaning the "recordable variables" that were not controlled in the test design, as well as the factors that were.) Additionally, we must improve our capabilities to verify these empirical statistical models to ensure they accurately reflect the data.

Example: Characterization in OT&E

The importance of using statistical analysis methods commensurate with the test design cannot be overstated. The following in-depth example highlights the differences between a typical analysis of the past and one that enables true characterization of system performance using regression analysis techniques. As part of this example, I illustrate the crucial use of confidence intervals. Confidence intervals, as well as prediction and tolerance intervals, are statistical tools that provide decision-makers with ranges on expected future performance. Confidence intervals also provide a measure of the level of knowledge we have of the system's performance. Wide confidence intervals indicate that we are less certain about how well the system will perform in the future. Narrow confidence intervals tell decision-makers that future system performance will be close to what we observed in the test.

Figure 1 shows an example of a robust characterization analysis using regression techniques. For this example, the system's goal is to maintain a lock on a moving target. If the system is able to successfully maintain the track for the desired period of time, the test trial is scored as a success; if the system drops the track at any point, then the test trial is scored as a failure. The purpose of our test was, therefore, to characterize the probability of maintaining track across all the operating conditions. The factors that drive the probability that the system is able to successfully maintain track include:

- Time of day (day/night)
- Target size (small/large)
- Target speed (slow/fast).

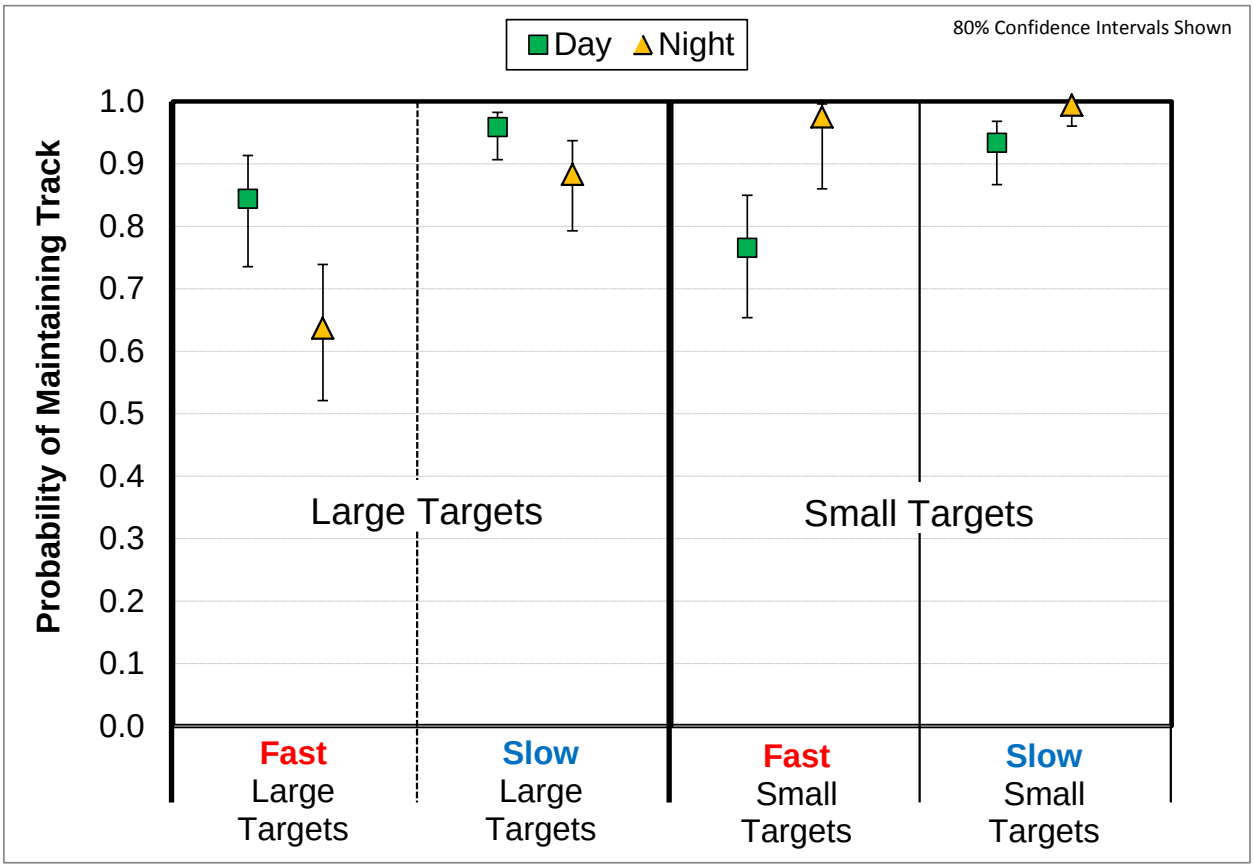

Figure 1: A Powerful Analysis Approach: Logistic Regression Enables a Full Characterization of System Performance

All tests are limited in scope to some degree, so if we hope to ensure robust characterization with minimal data, our analysis requires that the information be linked together through an empirical statistical model. In Figure 1, the predicted probabilities of maintaining a track and corresponding confidence intervals are the output of a logistic regression analysis. Using logistic regression and model selection techniques allows testers to distill the most important results from the data. In the analysis of the tracking system, the logistic regression allowed analysts to look for complex interactions across the operational space. One interaction identified as significant was the time of day by target size. This interaction manifests itself as the change to the effect of time of day for large and small targets. Notice that for the night-time results (orange triangles), performance against large targets is lower than the performance against small targets. This is true for both speed conditions, but the effect is even more pronounced when large targets are moving fast at night (second data point from the left in Figure 1). This interaction effect reveals that performance is extremely low in a specific set of conditions: large fast targets at night.

Compare the robust characterization shown in Figure 1 to a traditional analysis that calculates only one overall proportion (number of successes divided by total trials) or that selectively calculates proportions by test condition. Figure 2 shows these two analyses: an overall roll-up proportion across all conditions (far right) is common for many test reports; others may go a step

further and estimate averages across common conditions (e.g., all trials against large targets). The latter analysis results are shown as the first six points in Figure 2.

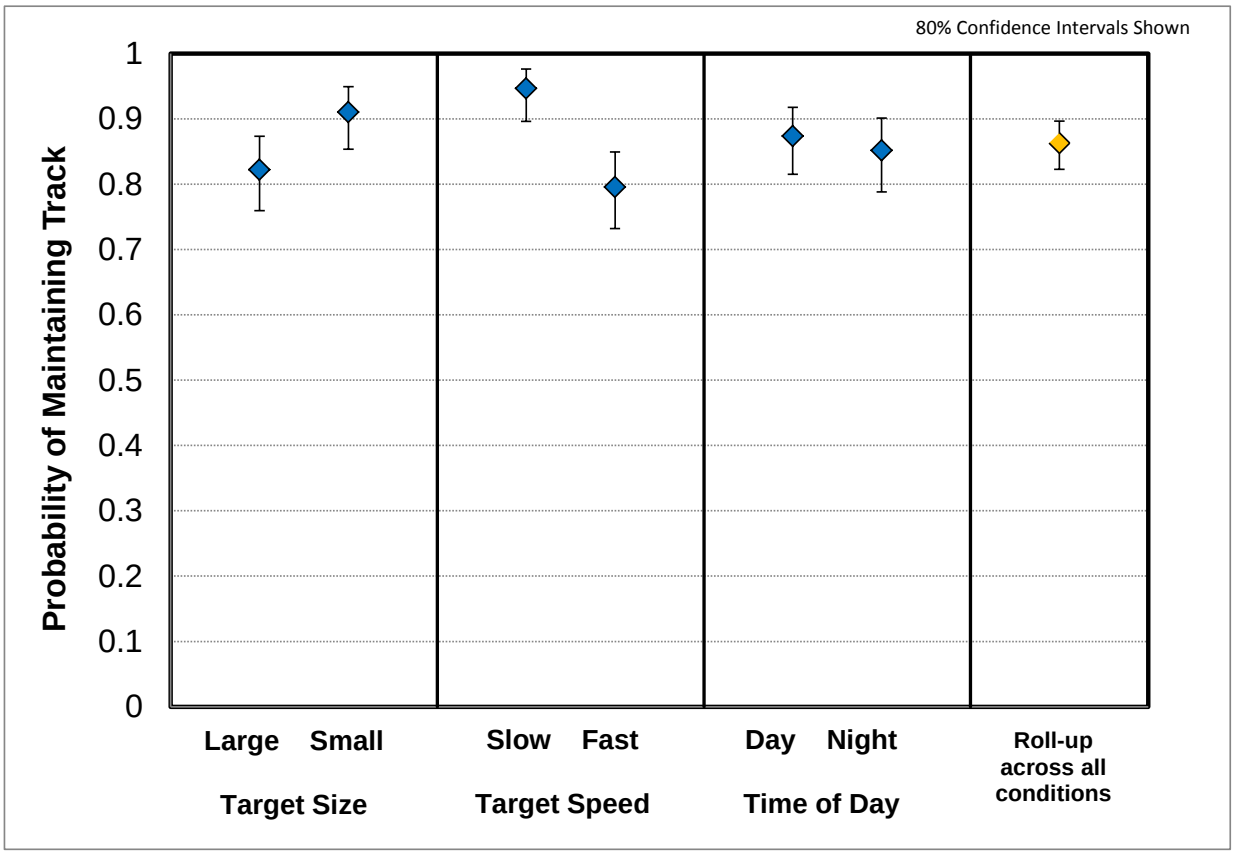

Figure 2: A Poor Analysis Approach: Calculating Average Performance Across Each Condition, or a Global Average

It is clear from this analysis that the overall roll-up calculation provides little insight into system performance across the operational space. Also, while the analysis shown in Figure 2 was able to identify that performance drops to some degree against fast targets (relative to slow targets), it fails to identify an important performance degradation that occurs against large, fast moving targets at night. An analysis based on averaging over common conditions (as shown in Figure 2) might conclude that performance was better than 0.75 in all cases; but this is clearly an incomplete conclusion, as revealed by Figure 1. Performance degrades due to the synergy between three factors; without the logistic regression analysis, conclusions about degraded performance are lost and the warfighter would be misinformed about the capability of the system. Additionally, the analysis shown in Figure 1 enables developers to target future system changes to ensure that large fast targets at night can be tracked equally as well as others.

To be clear, this example did not employ the best practice of using a continuous response variable. The narrow confidence bounds shown in Figures 1 and 2 for the tracking

characterization required 168 data points.[1] Although a large test was possible in this case, most operational tests are not resourced to collect this magnitude of data.

Advanced analysis methods to meet unique T&E challenges

Many tests are complicated by data that require more than the "standard" or "simple" analysis methods. In these cases, we should not avoid the opportunity to employ advanced methods simply because they are challenging or less well understood. We should continue efforts to employ these advanced statistical tools where appropriate, and will continue to encourage the use of and train the community on these methods. Some examples include:

- Bayesian approaches (especially in a reliability context) that allow us to leverage information from multiple phases of test while ensuring the results still reflect the operational reliability.
- Censored data analysis that allows us to incorporate information from continuous measures in cases where traditional pass/fail metrics would have been the only option.
- Generalized linear models and mixed models that allow flexible analysis methodologies that truly reflect the character of the data.

Well Designed surveys in OT&E

Surveys capitalize on the thoughts and experience of the system operators to derive essential information for the evaluation of systems. However, their use in OT&E has not always reflected the best practices of the survey community. The resulting data have had limited utility in evaluations. Data from well written surveys are useful for (a) diagnosing why certain performance goals were not met (e.g., training, system design), and (b) empirically measuring human system integration (HSI) components such as workload and usability. Workload and usability ratings can also form the basis of a robust comparison between new and legacy systems. In nearly every case, data from well written and well administered surveys aid the evaluator in assessing effectiveness and suitability.

One of the most common mistakes I have observed in surveys is the inclusion of questions that ask whether the user thought the system's performance was effective, accurate, timely, or precise enough to complete the mission. Accurate measurement of performance, effectiveness, and situation awareness requires knowledge of ground truth for the test, which operators and maintainers typically do not have. Surveys are measures of thoughts that are highly affected by context and are therefore relative, whereas requirements and performance are absolute, and are better measured by the tester.

A substantial body of scientific research exists on this topic. The following are some of the best practices, highlighted by the survey community, that we should consider when writing and administering surveys:

[1] The above-described tracking test might have used the duration of track time as a reasonable continuous metric in place of the pass/fail metric and thereby would have supported a rigorous analysis with a fraction of the required test size. In this case, the test was not constructed in a manner to enable accurate time measurements; the large data set was therefore required to ensure full characterization of performance.

a. Neutrality in the questions: The goal of the survey is to obtain the respondent's thoughts. Phrasing questions in a manner that leads a respondent towards the tester's opinions will reduce the likelihood that the respondent provides unbiased answers.
b. Knowledge liability: Do not ask questions the respondent cannot answer (e.g., did the system provide accurate tracking information?).
c. User friendly: Reduce the effort the respondent must put forward by making questions brief and clear. Also, make sure that the order of the questions is logical to the respondent.
d. Singularity: Address only one topic in a question.
e. Minimal length: The perceived length of a survey and the actual time it takes to complete it affects data accuracy. Ask the minimum number of questions needed for the goal of the test.
f. Confidentiality: When respondents believe that their data will be kept confidential, they are more likely to provide their true thoughts. Names and other personally identifiable information should be kept separate from the actual survey.

Institutionalizing STAT in T&E

DOT&E continues to observe significant benefits from applying STAT methods in OT&E. Not only does DOE provide a rigorous and defensible methodology for scoping operational tests to support robust characterization of system performance, but it also provides a methodological approach to developing tests. The logical framework helps guide test planning discussions. While characterizing system performance in OT&E is essential, the real benefits of STAT in T&E occur when applied early in the acquisition process when system design can be modified and improved. Industry players and Government agencies have traditionally applied DOE in testing that is more closely related to DoD developmental testing. The results have been improved system safety, optimized performance, and rapid detection of performance shortfalls. The DoD acquisition community as a whole would see more comprehensive benefits from applying a scientific approach to testing if DOE techniques were applied in contractor and developmental testing routinely. DASD (DT&E)) has endorsed the use of scientific approach to testing, noting that a scientific approach will generate defensible test and evaluation (T&E) strategy and improve the level of knowledge garnered from testing. Additionally, applying these methodologies across all phases of DoD testing will facility integrated learning about system performance.

The application of STAT should expand beyond DASD(DT&E), DOT&E, and oversight programs to the entire T&E community. They should be applied in contractor testing, early and late developmental testing, and in operational testing. Currently, pockets of excellence exist in these communities, but the application is far from wide-spread or standard. In order to facilitate the institutionalization of these methods we need to provide education and training to the existing T&E workforce, hire-in people with the capability to be immediate practitioners of DOE, and develop best practices and methods that adapt existing methodologies to DoD-unique applications.

Several forums can aid us in this endeavor. Existing communities sponsor workshops, conferences, and forums that T&E professionals need to reengage in once the fiscal climate enables us to. Forums, provided they are structured well, enable the community to exchange ideas and further the implementation and institutionalization of these methods across the Services and testing communities (CT, DT, OT). Some recent conferences worth noting include:

- Military Operations Research Symposium (MORS) Test and Analysis Workshop – MORS attempted an innovated approach this past year hosting a fully virtual workshop over Defense Connect Online (DCO) with training ranging from introductory to advanced. Virtual forums such as this provide valuable training in a fiscally constrained environment.
- ITEA and NDIA T&E Annual Symposiums – of particular note, at this year's ITEA symposium, there were several tracks dedicated to STAT.
- Conferences in the academic community – T&E professionals should take a more active role in statistical and operations research communities. Attending conferences such as the Joint Statistical Meetings, Fall Technical Conference, the Conference on Statistical Practice, and INFORMS will ensure that the T&E community is capturing best practices in these fields.

We can also do more to leverage expertise across the DoD and our sister Government agencies such as NASA, DHS, and NIST. In the upcoming year, DOT&E will be working on developing a cross-agency forum for exchanging knowledge and best practices across the DoD and other agencies to ensure we are capitalizing on the best talent within the Federal Government.

We also need to actively engage with science, technology, engineering, and mathematics (STEM) programs within universities to generate a pipeline for future qualified T&E professionals. DOT&E and TRMC continue to co-sponsor research and students to help build this pipeline and generate interest in T&E-unique challenges.

Finally, DOT&E and USD(AT&L) are reworking policy and guidance at the OSD level to promote the use of STAT methods. The newly released 5000.02 calls for universal employment of scientific approaches in T&E. In the DT&E Section (Enclosure 4), the guidance calls for:

"The TEMP will use scientific test and analysis techniques to design an effective and efficient test program that will produce the required data to characterize system behavior across an appropriately selected set of factors and conditions."

The Operational and Live Fire T&E (Enclosure 5) expands upon this guidance:

"Every TEMP will include a table of independent variables (or "conditions," "parameters," "factors," etc.) that may have a significant effect on operational performance . Starting at Milestone B, the updated table of variables will include the anticipated effects on operational performance, the range of applicable values (or "levels," "settings," etc.), the overall priority of understanding the effects of the variable, and the intended method of controlling the variable during test (uncontrolled variation, hold constant, or controlled systematic test design)."

DOT&E also publishes a TEMP Guide, which provides more detailed guidance on the use of STAT methods in OT&E. The TEMP Guide can be found on DOT&E's website: http:// http://www.dote.osd.mil/temp-guidebook/.

Summary

The T&E community is moving in the right direction by applying scientific approaches to T&E. The benefits are being realized in OT&E. However, analysis methods need to improve to capitalize on the power of statistically-based test design. Without robust analysis methods, the full benefits of employing DOE and executing efficient test designs are lost and our test reports will miss crucial information about system performance that decision-makers and the warfighter need to know. Additionally, we need to institutionalize the application of these methods across the T&E community, applying the best practices identified in this editorial.

References

Department of Defense. (2012). Scientific Test and Analysis Techniques in Test and Evaluation Implementation Plan. Washington, D.C.

Gilmore, J.M. (2010) Guidance on the use of Design of Experiments (DOE) in Operational Test and Evaluation. Washington, DC.