ITEA Journal 2011; 32: 237–240

Rigor and Objectivity in T&E: An Update

J. Michael Gilmore, Ph.D.
Director, Operational Test and Evaluation,
Office of the Secretary of Defense, Washington, D.C.

The Director of Operational Test and Evaluation (OT&E) began four Test and Evaluation (T&E) initiatives after his confirmation by Congress in fall 2009. Underlying his four initiatives were the need for rigorous and objective T&E. Since his original initiatives the Director has advocated for the use of statistically designed experiments as a methodology for increasing the rigor of test planning resulting in efficient tests yielding statistically defensible results. Additionally, he continues to emphasize the need for reliable systems and reliability growth plans and accordingly defensible reliability growth models in T&E.

I began my term as the Director of Operational Test & Evaluation (DOT&E) with four initiatives to increase scientific rigor in T&E. I published those initiatives in the June 2010, *ITEA Journal*, and I am happy to use this opportunity to provide an update. During the past year, I have seen several success stories as well as areas for improvement. I would like to commend ITEA for the theme of this journal, "The Rigor of the Scientific Method." And I appreciate the many articles others have authored on applying rigorous and objective scientific approaches to their specific test challenges.

J. Michael Gilmore, Ph.D.

In my initiatives I recognized that design of experiments (DOE) is an active academic discipline devoted to the study of scientifically proven methodologies for constructing and executing efficient, scientifically defensible tests. In the past year I have observed many of the benefits of using DOE in my review of test and evaluation master plans (TEMPs) and operational test plans. First, DOE requires the tester (and/or evaluator; I will make no distinction here) to provide a clear definition of the question we are trying to answer through T&E. DOE then enables the tester to ensure that the data collected will be adequate to answer the question. DOE provides the tester a large selection of strategies for efficiently spanning the operational test environment and analyzing the data. DOE provides the tester with a methodology for quantifying the risk of any proposed test (the statistical power) and statistical confidence associated with the test results. Finally, DOE provides the tester with methods for developing and analyzing sequences of tests. Before testing, DOE enables decision makers to clearly see the tradeoffs between test resources and risk. During testing, DOE enables testers to use early results to strengthen and refine subsequent tests. After testing, DOE gives decision makers a framework for understanding and weighing the importance of the results.

In October 2010, I outlined the specific elements of DOE that I am looking for when I review TEMPs and test plans. These elements are:

- The goal of the experiment. This should reflect evaluation of end-to-end mission effectiveness in an operationally realistic environment.
- Quantitative mission-oriented response variables for effectiveness and suitability. (These could be key performance parameters but most likely there will be others.)
- Factors that affect those measures of effectiveness and suitability. Systematically, in a rigorous and structured way, develop a test plan that provides good breadth of coverage of those factors across the applicable levels of the factors, taking into account known information in order to concentrate on the factors of most interest.
- A method for strategically varying factors across both developmental and operational testing with respect to responses of interest.

- Statistical measures of merit (power and confidence) on the relevant response variables for which it makes sense. These statistical measures are important to understand "how much testing is enough?" and can be evaluated by decision makers on a quantitative basis so they can trade off test resources for desired confidence in results.

Two recent examples of Milestone B TEMPs that have been substantially improved through the use of DOE are Joint and Allied Threat Awareness System (JATAS) and Joint Standoff Weapon (JSOW). Both programs have been able to provide a clearly justifiable test resource matrix based on DOE at this early stage in the program. The strength of DOE was demonstrated in both programs: in both TEMPs there is clear definition of how the test will be executed and under what conditions. Clarity helped foster rapid agreement between the program office, developmental testers (DT), and operational testers (OT). DOE also provided a clearly justifiable sample size through the use of well-known and well-studied named experimental designs. In both programs the selected designs will allow the programs to determine active operational factors in DT, with the goal of reducing the scope of testing in OT if certain factors are determined to be inactive. These two programs illustrate the strength that DOE provides to the test planning process. At present the test resources are justified with clear, quantitative analysis, to the benefit of all involved. I am confident that DOE will continue to prove its value for these programs, as they use the initial results to refine the plans and resource requirements for initial OT&E.

However, using statistically designed experiments in itself does not guarantee that the testing will be adequate. Expert system knowledge is needed to determine experimental goals, critical factors, and other testing decisions. It is important for the full test team to be engaged in the DOE process and that the implications of all decisions are clearly understood.

Two recent DOE approaches with chemical agent detectors yielded vastly different DOE outcomes. DOE achieved the intended goal for the Joint Chemical Agent Detector (JCAD) but not for the Navy's Improved Point Detection System–Lifecycle Replacement (IPDS-LR). Both systems are ionization mass spectrometers used to detect the presence of chemical agent vapors present in the air. T&E of both detectors relied heavily upon chemical agent chamber test events to help determine operational effectiveness. In these test events, agent concentration, temperature, and humidity were systematically varied over the design space to determine the detection performance over multiple potential operating environments for multiple agents. The experimental designs for IPDS-LR and JCAD both had high predicted power calculations, approaching close to 0.90 for some factors within the design at 1 : 2 signal to noise calculations.

After the tests were conducted and the data collected, the utility of the resultant data sets were vastly different. The JCAD chemical agent data set was robust and facilitated surface response modeling and prediction of detector performance over the entire operating envelope. By contrast, the IPDS-LR data set, despite high predicted power calculations, failed to provide sufficient data to even attempt surface response modeling. In the end only simple univariate probability of detection calculations could be done. Comparison of IPDS-LR performance against different agents was further complicated because by design the same temperature and humidity ranges were not tested among the agents. With an experimental design meant to build a response surface model, resources can be spared, because it is not necessary to test all factors at all levels. The successful response surface modeling of the JCAD data facilitated full analysis of the detector in similar environments against threat representative concentrations of all the chemical agents in question.

The primary cause of the incorrect DOE matrix for the IPSD-LR design was a misclassification of certain factors as categorical when in fact they were continuous. DOE is a mathematical tool that works off of a mathematical description of the system under test. Programs and testers must work together to create a correct mathematical description. The measures of performance, the factors (independent variables) affecting performance, and the appropriate levels (variable values) of those factors must be correctly described in order to design and execute testing that in the end will provide a robust set of data for analysis and evaluation.

The differences in the available analysis between JCAD and IPDS, when both test programs were based on the use of DOE, illustrates one of the many challenges for using DOE in a smart manner when constructing test designs.

Some of the challenges I have seen in implementing the use of rigorously designed experiments include the need for more training on the use of statistics and DOE techniques. Although there have been many successful case studies in industry, they are not readily available to the testing community. Additionally, there are many safety and cost constraints in the testing of military systems that play into the test process and that can limit the direct application of traditional DOE.

My office is working to tackle each one of these challenges. My Science Advisor, Dr. Catherine Warner, has formed a Test Science Steering Commit-

tee to tackle many of these issues. They, along with DT&E, are looking at the current education needs of the T&E workforce for both developmental and operational testing. In addition my office and the Test Resource Management Center have now sponsored a research consortium to look into the open research questions for applying DOE to T&E. This research consortium will compile a library of DOE case studies that the T&E community can reference in the future for all types of military systems.

Some worst practices that I have observed since embarking on the use of statistically designed experiments in T&E include:

- Reducing all data into a mean and standard deviation. This practice ignores the impact of the operational environment on the performance of the system and may result in surprise areas of poor system performance for the warfighter, a risk I most assuredly want to avoid. A simple mean and standard deviation does not adequately describe complex system performance. We need data-driven statistical models that provide a comprehensive picture of how performance varies across the operational envelope.
- Unnecessary replication. This is an area where DOE may be able to cut costs. In the JCAD example I described previously each of the cases was replicated 16 times. This amount of replication was unnecessary to draw conclusions. In fact, in a later study of the detector my office showed that four replications would be sufficient for properly modeling the detector performance.
- Testing systems in exercise. This practice allows for little control over the test. A key concept behind DOE is we can accurately characterize system performance by explicitly changing the factors that influence performance. An exercise is operationally realistic but the tester loses control of the test structure. It is difficult to obtain statistically defensible results. Additionally, depending on the system under test there is a risk that the system may not be used at all or only used in a limited capacity, which limits our ability to objectively assess system performance. The Common Aviation Command and Control System (CAC2S) is one example of this drawback that I have observed recently. No data was collected on one of the primary missions that CAC2S must perform due to the exercise nature of the test.
- Determining the factors that impact performance correctly in OT when they were not observed in DT. Applying DOE in OT alone is not nearly as beneficial as applying DOE across the testing continuum. My office is currently working closely with the Deputy Assistant Secretary of Defense, Developmental Test and Evaluation (DASD, DT&E) and his office to ensure that testing is rigorous and objective across the entire test program.

I would be remiss if I did not address reliability as a major component of my initiatives. I have focused my emphasis on the use of DOE in determining system effectiveness. However, it is also important to consider the need for objective and rigorous testing of suitability, especially given the current reliability challenges my office has observed and the implications it has for sustainment costs for military systems. In June 2010, my office wrote to the Deputy Under Secretary of Defense (Acquisition, Technology, and Logistics) about the state of reliability in T&E. The memorandum concluded:

- Poor reliability is a problem with major implications for cost (5 to 10 times more impact on total life cycle costs than do research development T&E).
- Systems emerging from design and development efforts are often not reliable.
- The essential issue of reliability is that it competes with achieving more operant capabilities. We must assure vendors' bids to produce reliable products that outcompete the cheaper bids that will not.
- Requiring use of ANSI/GEIA-STD-0009 is appropriate.

As a result of that memo and other collaborations with AT&L, in June 2011, the USD(AT&L) issued a directive type memorandum (DTM) on Reliability Analysis, Planning, Tracking, and Reporting. The DTM is a step in the right direction, and it will become part of 5000.02 by the close of 2011. The DTM does not mandate any specific methodologies for reliability analysis. It does require that program managers develop a reliability and maintainability program based on an appropriate reliability analysis method. Additionally, it requires that the reliability growth curve reflecting the reliability growth strategy be included in the TEMP beginning at Milestone B. I am hopeful that as a result of these policy changes, more programs will make use of the request for proposal and contract language recommended in conjunction with the ANSI/GEIA-STD-0009.

My office recently conducted a review of 257 of 353 programs on the December 2010 oversight list. We noted several trends when comparing 2010 to pre-2008, when the reliability initiatives of the Office of the Secretary of Defense began. The percentage of programs that have reliability as an element of test

strategy has increased since June 2008. Before then, 69% of programs planned to collect and report reliability data, while 90% of programs with a TEMP since June 2008 planned to collect and report reliability data. The percentage of programs that have a reliability growth strategy has also increased since June 2008. The essential purpose of a reliability growth curve is to drive a growth strategy, but only about half of the programs with growth strategies use reliability growth curves and include adequate time to implement corrective actions.

One issue that continues to persist is in the adequacy of test planning for testing reliability. Two recent examples have highlighted the need for the consideration of statistical confidence and power in reliability testing: Joint Air to Surface Standoff Missile (JASSM) and JATAS.

Whether one is testing effectiveness or suitability, statistically defensible test planning requires consideration of the question to be answered and the appropriate analysis to answer those questions. In the JASSM program, the Air Force performed reliability testing on baseline production Lots 1 through 5 and Lot 7. Reliability growth has been observed and significant increases in reliability were observed in Lot 4 and Lot 7; however, the point estimate of the reliability of Lot 5 dropped below the threshold of 85% reliability. Based on the poor test performance in Lot 5, the program office implemented corrective actions and increased emphasis on missile reliability. The results were positive in Lot 7; however, when one missile from Lot 6 was fired, it was a failure. What were the appropriate subsequent questions and analyses? Several different sample sizes were considered for the remainder of testing in Lot 6. DOT&E computed how many missile shots from Lot 6 would be necessary to conclude with 80% confidence that Lot 6 improved over Lot 5. Using a test of two proportions we were able to show that 11 shots would allow for a second Lot 6 failure while still demonstrating (assuming the other 10 shots succeeded) 80% confidence that Lot 6 is improved.

Test planning for reliability analysis commonly uses the outdated rule of thumb that the test time should total three times the requirement. This rule of thumb allows for one failure while still concluding the system meets the reliability requirement with 80% confidence. However, this method for test planning completely ignores the risk of only seeing one failure during the course of testing.

JATAS is one example of how the 3× rule of thumb has resulted in inadequate reliability testing. The resources for JATAS were allocated well before my office had any involvement with the program. The reliability growth curve for JATAS currently only allows for one failure in all testing if the system is to meet its reliability requirement with 80% confidence. However, the probability of only observing one failure is near zero, unless the system is highly reliable (i.e., true reliability is four to five times the reliability requirement). The JATAS program office acknowledged the extremely high risk the system has for failing the reliability requirement and planned for reliability development and growth testing in the laboratory as well as a dedicated reliability qualifying test. However, none of this replaces true flight hours.

There have been significant gains made in the past year in the process of developing rigorous and objective tests. However, there is still a long way to go. My office is working to find solutions to the challenges and make sure they are available for use throughout the Department. ❑

Dr. J. Michael Gilmore was sworn in as director of Operational Test and Evaluation on September 23, 2009. A Presidential appointee confirmed by the United States Senate, he serves as the senior advisor to the Secretary of Defense on operational and live fire test and evaluation of Department of Defense weapon systems. Previously, Dr. Gilmore was the assistant director for National Security at the Congressional Budget Office (CBO). Dr. Gilmore is a former Deputy Director of General Purpose Programs with the Office of the Secretary of Defense, Program Analysis and Evaluation (OSD[PA&E]). Dr. Gilmore's service with Program Analysis and Evaluation covered 11 years. Earlier, Dr. Gilmore worked at the Lawrence Livermore National Laboratory; Falcon Associates; and McDonnell Douglas Washington Studies and Analysis Group where he became manager, electronic systems company analysis. Dr. Gilmore is a graduate of Massachusetts Institute of Technology, Cambridge, Massachusetts, where he earned a bachelor of science degree in physics. He subsequently earned master of science and doctor of philosophy degrees in nuclear engineering from the University of Wisconsin, Madison, Wisconsin. E-mail: mike.gilmore@osd.mil