

## **The Science of Test Workshop** April 11-13, 2016







# **IDA** | OPERATIONAL EVALUATION DIVISION



January 2017

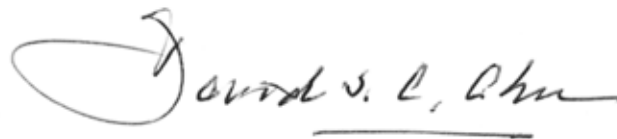
The Department of Defense Director of Operational Test and Evaluation, the National Aeronautics and Space Administration, and the Institute for Defense Analyses hosted the inaugural “Science of Test Workshop,” in connection with IDA’s 60<sup>th</sup> Anniversary celebration. The Defense and Aerospace analytical communities have long believed such a workshop would be productive – and, indeed, it proved exceptionally so.

The workshop focused on statistically rigorous approaches for testing defense and aerospace systems. While the case for rigorous statistical test and evaluation of such systems might seem obvious, reviewing the benefits reinforced what we too often assume is accepted. Rigorous statistical approaches create a defensible framework for determining how much testing is enough. They offer the analytical trade space that allows decision makers to weigh the importance of information against the cost of testing. They identify the most efficient test plans.

In the evaluation of test results, statistical methods provide robust estimates of likely performance, and they aid decision makers in understanding how confident we are in those estimates. They provide operators with a deeper understanding of the capabilities and limitations of systems under the range of possible conditions. To the extent that performance falls short of what’s desired, a statistical analysis can help guide the corrections that will give the operators what they need to accomplish the mission – and to return to tell us about it.

As with any profession, advances are constantly occurring in statistics. New techniques are discovered, and older techniques are improved. This constant evolution is why the DoD and NASA need to develop a community of interested practitioners. Building such a community was the purpose of convening the Science of Test Workshop. The Workshop supported the healthy exchange of ideas critical to capitalizing on the broad intellectual advances of the profession as a whole.

We have already started work to hold a second Science of Test Workshop (<http://workshop.testscience.org/>). I am confident that exchanges such as these will better prepare national security analysts and statisticians to continue their contribution to this country’s interests.

A handwritten signature in black ink, reading "David S. C. Chu". The signature is fluid and cursive, with a large, stylized initial "D" that loops around the first part of the name.

David S.C. Chu



OPERATIONAL EVALUATION DIVISION

---

## Contents

<b>Overview</b>	1
<b>Keynote Address</b>	3
Joint Strike Fighter (JSF) F-35	3
Littoral Combat Ship (LCS)	5
Remote Minehunting System (RMS) and Remote Multi-Mission Vehicle (RMMV)	6
Probability of Raid Annihilation (PRA) Testbed	7
Warfighter Information Network – Tactical (WIN-T)	8
New Display (Simplified)	8
Conclusion	9
<b>Rigorous Test and Evaluation for Defense, Aerospace, and National Security: Panel Session Summary</b>	11
Clear Communication	12
Answer the Right Question	13
Statistics Is a Team Sport	14
Implementing Statistical Thinking in Organizations	14
Concluding Thoughts	15
<b>Summary of the Agenda</b>	17
Workshop Short Courses	17
Mini-Tutorials	17
Breakout Sessions	18
<b>Workshop Attendee Feedback and Future Directions</b>	19
Attendance Motivation	19
Overall Event Quality and Presentation Satisfaction	19
Lessons Learned for Planning Future Events	22
Conclusions	23

---

**IDA** | OPERATIONAL EVALUATION DIVISION

---

## Overview

In April 2016, NASA, DOT&E, and IDA collaborated on a workshop designed to build a community around statistical approaches to test and evaluation in defense and aerospace. The workshop brought together practitioners, analysts, technical leadership, and statistical academics for a three-day exchange of information, with opportunities to attend world-renowned short courses, share common challenges, and learn new skill sets from a variety of tutorials.

The Department of Defense, NASA, and other government agencies acquire the world's most advanced, sophisticated, and complex systems. Rigorous testing and scientific characterization of these systems are essential for defensible decision making and successful operation. The goals of the Science of Test Workshop were to establish a professional community concerned about statistical approaches to test and evaluation and to provide an opportunity for test and evaluation practitioners to be exposed to academic experts. Analysts from across the federal government benefited from training sessions, technical presentations, and case studies showcasing best practices. Leadership perspective highlighted the importance of rigorous methods in testing and the evaluation of system capabilities.

The workshop program contained a unique mix of training opportunities, leadership perspectives, case studies, and technical tracks. These proceedings summarize the key elements of the workshop including the opening keynote by Dr. J. Michael Gilmore, Director, Operational Test and Evaluation (DOT&E), and a cross-organizational leadership panel showcasing representatives from the DoD Service Operational Test Agencies and key technical leadership within NASA.

More than 200 people participated in the workshop, representing organizations from across the DoD, NASA, the National Labs, Federally Funded Research and Development Centers (FFRDC), NIST, and academia. Of the 200 attendees, 76 provided feedback via an email survey; the feedback was overwhelmingly positive. All respondents said they would recommend the workshop to others. The final section of these proceedings summarizes the results of the survey and suggests topics for future workshops addressing the science of test in Defense and Aerospace. A recurring theme in the survey is that participants valued the focus on analytical topics.

**IDA** | OPERATIONAL EVALUATION DIVISION



## Keynote Address


The opening keynote address from Dr. J. Michael Gilmore, Director, Operational Test and Evaluation, provided his perspective on operational testing of new military systems and how rigorous statistical and analytic techniques are critical to the efforts of the test community.

Dr. Gilmore noted that the purpose of operational testing is to provide realistic and objective assessments of how systems improve mission accomplishment under realistic combat conditions. To do this, he noted, “It seems common sense to use rigorous statistical and analytic techniques. This is true in any budget environment, but it is particularly true in the environment that we have now where budgets are coming down and they are unlikely to increase substantially in the near future.” He went on to describe that we should use rigorous techniques to design operational tests, to determine how much testing needs to be done and under what circumstances, to understand which factors affect performance, and to assess whether the men and women who use these systems in the field can accomplish their missions.

Dr. Gilmore noted, “You have to explain why you’re doing what you’re doing and why it is needed. I’ve found statistical design and analytic techniques to be indispensable in that regard.” He illustrated these points through a series of examples.

### Joint Strike Fighter (JSF) F-35

In his first example, Dr. Gilmore discussed the Joint Strike Fighter (JSF), the largest program within the Department of Defense. “It is important to note,” Dr. Gilmore said, “that stealth aircraft are not invisible, just hard to detect, which means that the pilots must know when threats, particularly mobile threats, are in the vicinity in order to avoid them.”



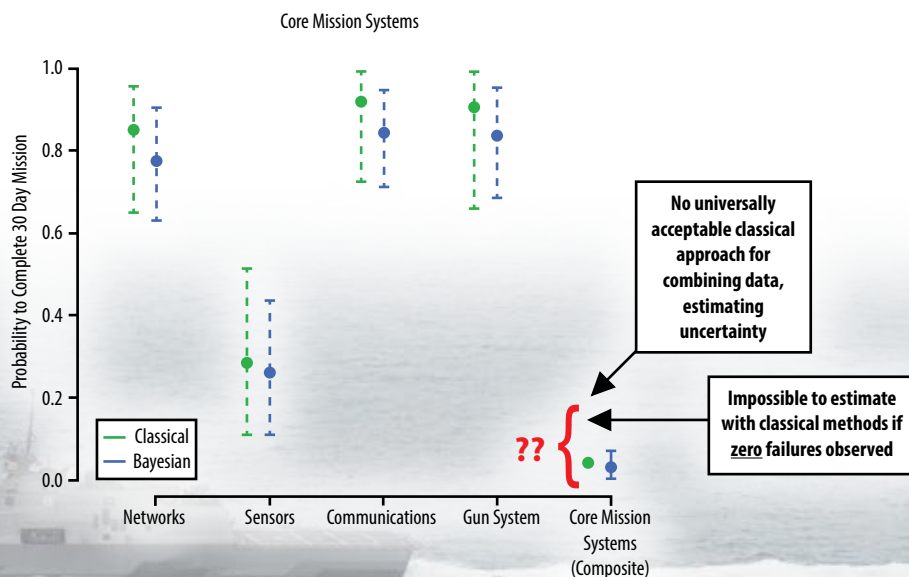
An important part of the F-35’s mission systems is the mission data file, which contains the settings that JSF sensors use to detect a threat’s integrated air defense systems. The United States Reprogramming Laboratory (USRL) is responsible for building the mission data file. Initially, the problem seemed overwhelming. It’s a multi-variable optimization problem: maximizing the F-35’s detection and correct classification of threats and minimizing the ability of threats to



## Littoral Combat Ship (LCS)

In his second example, Dr. Gilmore discussed the LCS, which is designed to operate in the shallow waters of the littorals as well as in deep water, open ocean environments. The Navy is currently procuring two variants of LCS. The *Freedom* class is a semi-planing monohull design, and the *Independence* class is trimaran design. Both variants include core mission systems that are permanently installed and separate mission packages that can be loaded onto either variant for specific missions such as mine warfare or surface warfare. The core mission systems include the ship's computing environment, sensors, communication systems, and a gun system.

The Navy would like to know whether LCS's core mission systems can complete a 30-day mission without an operational mission failure. The core mission systems are a mix of continuous use systems (computing environment, sensors, and communications) and one on-demand gun system, which makes the combining of reliability data into an overall estimate of core mission system reliability complex. The figure below shows the data from a recent test analyzed with both classical statistics and Bayesian techniques. Unfortunately, using classical statistics, there is no universally accepted method of estimating uncertainty for the composite of continuous and on-demand reliability data. IDA recommended a Bayesian approach that provides an estimate of the uncertainty. Dr. Gilmore noted that this is "an example of the application of rigorous techniques for analyzing data from testing and coming up with what I personally found to be a rather compelling result." These ships are unlikely to complete a 30-day mission without an operational mission failure, and we can be confident in this conclusion because statistical techniques allow us to estimate the uncertainty in the result.



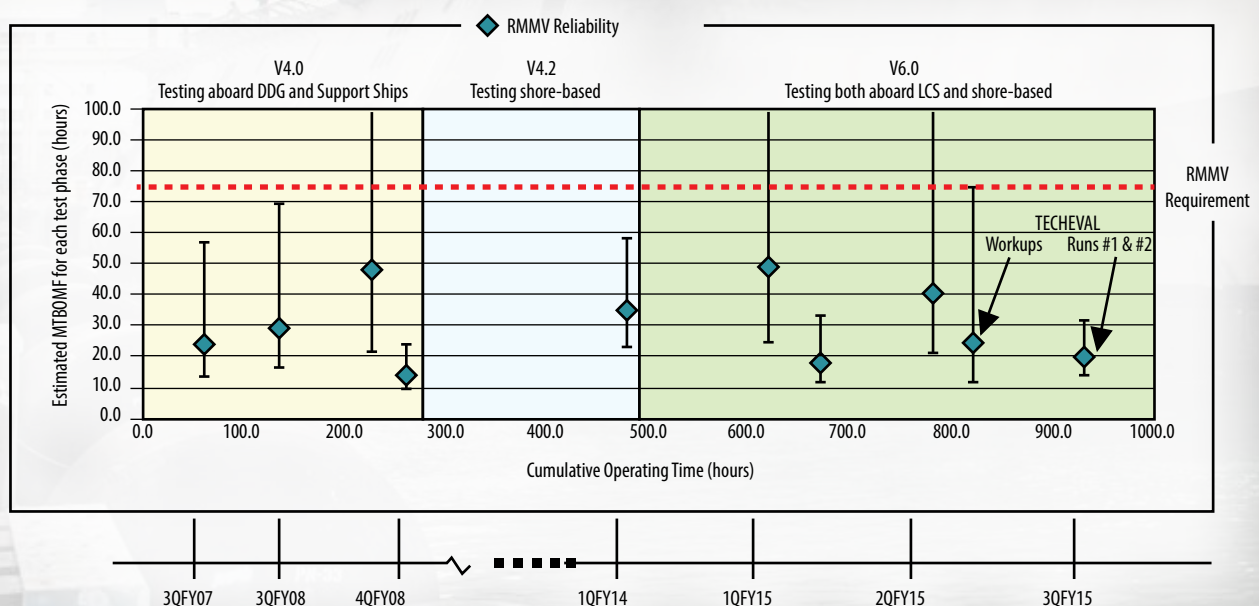


## Remote Minehunting System (RMS) and Remote Multi-Mission Vehicle (RMMV)

The third example examined the Navy's newest minehunting system, RMS, a component of the LCS's mine countermeasures mission package. The RMS comprises the RMMV and a towed sensor package that the Navy uses to hunt for mines. They have been under development for more than 10 years. Dr. Gilmore noted, "This system has had a troubled development including problems with localization errors, depth errors, and false contacts, but here I will talk about the reliability problems."

Part of the RMMV's troubled reliability history has been reliability metrics that were not operationally meaningful. The Navy assessed that the system exceeded the requirement for mean time between operational mission failures by more than a factor of 200 percent. The Navy's estimate, however, was based on the definition of operating time that included the many hours when the system was not in the water conducting operations. During this time, the system had no opportunity to fail, and inclusion of the idle time in the calculation grossly overestimated the vehicle's true reliability. This example illustrates that clear and operationally meaningful definitions for reliability are essential to assessments of military systems.

In addition, there were questions about whether RMMV reliability was improving and whether recent failures were due to the age of the equipment (wear-out failures). IDA examined the data, and Dr. Gilmore noted, "These questions are amenable to statistical analysis and that's what this chart below shows. We looked at reliability over time using the reliability definition in the contracts. We found that the growth was essentially zero with no statistically significant change in reliability for years." Dr. Gilmore continued, "Which also means it wasn't wearing out either; it was just continuing to be unreliable." In the end, the results were well below the requirements and the entire program was canceled.





## Probability of Raid Annihilation (PRA) Testbed

One of Dr. Gilmore's recent initiatives is increasing the rigor of model validation. One area where we use modeling and simulation is in operational evaluations of surface ships to defend themselves against attacks by anti-ship cruise missiles. These threats are proliferating and are becoming much more stressing to our air defense systems. In 1987, an Iraqi Exocet subsonic cruise missile hit the USS *Stark*, raising the visibility of this threat. However, Dr. Gilmore said, "Testing sometimes can get a little more realistic than people would like," illustrated by the recent accidental collision of a target drone with the USS *Chancellorsville*. Risks like these mean that we can do only a limited number of live tests, so an important part of the evaluation is modeling and simulation. Thus, the Navy developed the Probability of Raid Annihilation (PRA) Testbed that provides high-fidelity models of a ship's ability to defend against a raid of incoming cruise missiles.

Before we can use the PRA Testbed, we need to convince ourselves that it is reasonably representing what's seen in the live tests. To do this, we want to compare the PRA Testbed results to results from live testing. The PRA Testbed estimates numerous metrics, including the range at which a ship initially detects the incoming threats, when the ship launches missiles in self-defense, and missile intercept ranges. These and other metrics can be compared to live test results. Dr. Gilmore noted that historically these comparisons have relied on subject matter expertise gathered around various tables, but that the subject matter expertise judgment alone is not sufficient.

Consequently, IDA proposed that statistical analysis should add rigor. This is illustrated in the equation below, where initial detection range is examined as a function of whether the result is from a live test or the PRA Testbed (*TestType*), which threat is being examined (*TestThreat*), and an interaction term. When this equation is fitted to the data, we expect that we will see statistically significant differences between different threats, but we hope that we do not see statistically significant differences between the live and model data (i.e.,  $\beta_1$  is small and statistically insignificant). If we do not see such differences, we can say that the live results and the PRA Testbed are statistically indistinguishable.

$$\text{Initial Detection Range} = \beta_0 + \beta_1 \text{ TestType} + \beta_2 \text{ TestThreat} + \beta_3 (\text{TestType} * \text{TestThreat}) + \epsilon$$



In 1987, two Iraqi Exocets hit USS *Stark*, increasing the focus on ship self-defense.



Cost and safety restrictions limit the number of live test events.

## Warfighter Information Network – Tactical (WIN-T)

Dr. Gilmore's fifth example was WIN-T, an Army communications system using both satellite and terrestrial datalinks. It allows warfighters to exchange information in tactical situations. As WIN-T has been upgraded over the years, usability has been a key question.

The initial testing of WIN-T focused on the performance, which experienced problems with maintaining connectivity. In addition, problems with the complexity of the system were noted in the early testing, including how hard it was for soldiers to get it to work even when the software and hardware were working correctly. Dr. Gilmore noted that, "to the Army's credit, they kept working and improving the performance of the system and making it easier for the soldiers to operate. They now have a system that I was able to evaluate as effective as of last year."



**Old Display (Complex)**



**New Display (Simplified)**

The testing examined here focused on improving the man-machine interface that soldiers use to keep the system operating on the battlefield. Initially, the interface had multiple sub-menus, and, when the system went down, it could take 40 minutes to an hour to restart it. As depicted below, the original interface was complex and difficult to read. The new interface is far simpler.

### New Display (Simplified)

During testing, we wanted to understand the difficulties that soldiers may have had when using the system, which was evaluated using surveys. The Army initially constructed surveys that were complex, with nested questions and "Not Applicable" as a potential response. Dr. Gilmore observed, "If you're tired at the end of the day and you're getting sick of filling out this survey, what are you going to do? You're going to check 'Not Applicable.'"

#### **Example (poor) survey question:**

**"During movement were you able to communicate using the PoP?"**

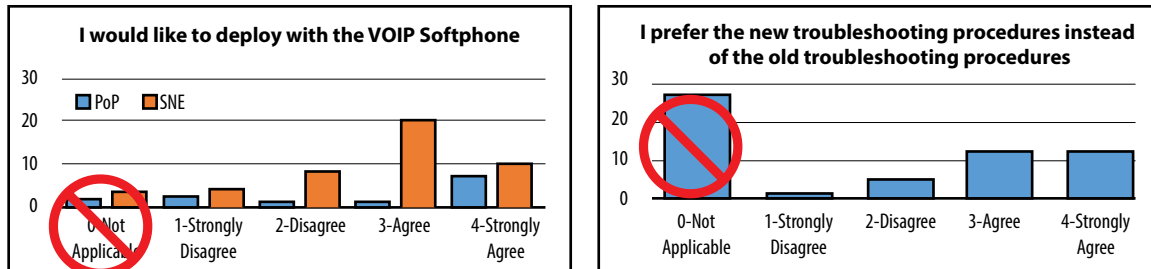
**If there were times when you could not communicate, do you know why?**

**What was the terrain? Approximately how fast were you going?**

**[...question continues...]"**



Dr. Gilmore has been pushing the test and evaluation community to incorporate survey science into testing, and IDA worked with the Army to improve the surveys. This led to the simpler surveys depicted below, which are much more meaningful, much more likely to be completed in a useful manner, with results easier to interpret. However, there is still room for improvement in these surveys; future surveys will remove the “Not Applicable” category.



## Conclusion

Dr. Gilmore’s keynote address hit many of the major points that he has emphasized as the Director, Operational Test and Evaluation. The points included analytic rigor, test designs that are efficient and cover the operational envelope, reliability, model validation, and survey design. Overall, they illustrated his desire that we apply rigorous techniques to design operational tests, to determine how much testing needs to be done and under what circumstances, to understand which factors affect performance, and to determine whether the men and women who use military systems in the field can do so effectively.



**IDA** | OPERATIONAL EVALUATION DIVISION



## Rigorous Test and Evaluation for Defense, Aerospace, and National Security: Panel Session Summary



A highlight of the workshop was the Tuesday afternoon technical leadership panel chaired by Dr. Catherine Warner, Science Advisor, DOT&E. Panelists shared stories illustrating the lessons they had learned over the years in technical leadership roles. The panelists' remarks reflected a few core themes, including the importance of clear communication about statistical methods and an emphasis on making sure that you are answering the right question. As one panelist, Mr. Roberts, eloquently noted, "You can greatly increase value and impact by doing simple analysis and answering the right question, as opposed to doing very sophisticated analysis and answering the wrong question." It also was observed that solving problems using statistical methods is really a "team sport" and that clear communication and asking the right questions are the first steps in getting all stakeholders involved in the process to ensure a successful outcome. Finally, the panelists touched on the challenges they had seen in successfully implementing "leading edge" statistical methods in organizational processes.

The panel consisted of representatives from across the Department of Defense (DoD) Test and Evaluation (T&E) communities and senior technical leadership from NASA. The panel participants included:

- Mr. Paul Roberts, NASA Engineering and Safety Center Chief Engineer at Langley
- Mr. Frank Peri, Deputy Director of the Langley Engineering Directorate, NASA
- Mr. Paul Johnson, Scientific Advisor, Marine Corps Operational Test and Evaluation Activity (MCOTEA)
- CAPT Peter Matisoo, Technical Director, Commander Operational Test Force (COTF)
- Jeff Olinger, Technical Director, Air Force Operational Test and Evaluation Command (AFOTEC)

Dr. Warner started the conversation by reflecting on the large number of participants at the workshop, contrasting their numbers and enthusiasm with the situation in 2010 when she first became the DOT&E Science Advisor, and many in the DoD were openly hostile towards

---

using statistical approaches for operational test and evaluation. Dr. Warner noted that the success stories being related throughout the workshop highlighted how people were using regression and other kinds of modern statistical techniques to solve the complex problems that face our national defense and aerospace industries.

## **Clear Communication**

All of the panelists discussed the importance of clear communication in the successful implementation of statistical methods in their experiences. Mr. Johnson recounted his experience explaining to senior leadership why the Marine Corps needed experimental data to support decisions on incorporating women in ground combat roles. His unconventional approach to discussing statistics with senior leaders involved using weighted dice and live demonstrations of statistical hypothesis testing. Mr. Johnson shared stories of rolling the dice with leaders at the highest levels of the Marine Corps to explain the importance of experimental designs and statistical thinking!

*“A four-star general has enormous decisions on his plate. He does not have the time to come and meet me half way in the explanation and understanding process. I assume it is my responsibility to go 100 percent in his direction, to learn his vernacular, and carry my message across so that he gets it.”*

His story of convincing leadership that this was the right direction to move resonated with the audience. Mr. Johnson emphasized to the technical audience:

*“It’s incumbent upon us to make sure that our message – the statistical designs – the robustness – the reason why we’re doing what we’re doing – is accessible to the decision-makers.”*

The panelists noted the importance of communication in the DoD testing, which requires involvement from many organizations. Mr. Olinger and CAPT Matisoo noted that it was important to have strong working groups that communicated clearly and frequently across organizational lines, including the requirements experts, system operators, and oversight organizations, and that all stakeholders be involved in the test planning process. The panelists were especially reflective of communication with DOT&E oversight (their panel moderator!); Mr. Olinger noted:

*“We work with DOT&E; we don’t work for them. A lot of people talked about close communication being a key piece. Per DoD instruction, DOT&E approves our plan adequacy, and so we’re working very closely on the plan that we go execute.”*

Finally, CAPT Matisoo noted the importance of communication in turning the data into information, highlighting that communication is important in both planning a test and reporting the results. He noted that you need the system operators (the customers) in the room when discussing and presenting analysis to ensure that the analysis is meaningful and the operational implications are clear,

*“Get those operators back in the room, sit down with the operators, and talk about what we’re seeing from the experiment.”*

---

## Answer the Right Question

Closely related to the need for clear communication is the need to make sure that you are trying to answer the right question. Mr. Roberts related a story about an engine valve failure they were investigating on the space shuttle program. He noted that the original question posed was “*the probability of a catastrophic failure due to another poppet valve failure.*” The answer to the question was critical for future determination of flight readiness. An interdisciplinary team was established and set off to answer the question,

*“We tested for a few weeks, generating data, did a lot of sophisticated statistical analysis on it, walked that preliminary analysis back to the program office and said this is the type of information we’re going to be able to give you and they looked at it and said I can’t make a decision from that, that doesn’t do me any good. So we were answering the question that they asked, but it wasn’t the right question. We went back and we began to get a heavy dose of statistical engineering, what we should have done right up front. We started asking questions, talking to all kinds of people, trying to find out what knowledge they were really looking for. Now we eventually found the right question and that was, what is the velocity that causes a certain depth of dent due to a projectile hitting in various orientations? From the answer to that question, they could tie it with structural analysis and come up with a critical velocity of the flow, and as long as we kept everything under that critical velocity, the probability of having a catastrophic event due to a poppet failure would be very very low. ...*

*What I want you to take away from that particular little story is, if you apply the proper statistical engineering methods, you can greatly increase the value and the impact by doing simple analysis and answering the right question, as opposed to doing very sophisticated analysis and answering the wrong question.”*

From the operational test perspective CAPT Matisoo reflected similarly on the need to understand the question the operational test is trying to answer:

*“One of the key lessons that I could give you is sitting down with whoever your operators are, in our case it’s folks who are in warfighting uniforms, but sitting down with the customer or the operator and really understand what the question you’re trying to answer is.”*

Mr. Olinger addressed the same concern of getting the right question from another perspective; he noted that in operational testing we have to balance test needs with test complexity.

*“Another challenge is level of test. We need to figure out the balance between testing at a campaign level – can we win the war – which requires a large, complex test – to how did the system do in the limited environment that it was in a smaller test?”*

Mr. Olinger highlighted that the two different questions drive different approaches to test design, and each can have different impacts on cost and the capabilities required to execute the test.

---

## Statistics Is a Team Sport

All of the technical leaders emphasized the need for an interdisciplinary team for determining the best approach for answering the questions of interest. Mr. Olinger described the AFOTEC initial test design process:

*“We bring in all the stakeholders, ... the program officers, the contractor, we bring in DOT&E, we bring in intelligence. You name it, if anyone’s got a stake in it we bring them into the meeting. Yesterday in Dr. Montgomery’s talk, the first three steps of any experiment design, if you will, he said is a team sport. So you really need to make sure you have all the people involved.”*

From a different perspective, Mr. Peri also noted the importance of fully integrating statistical engineers and statistical engineering into existing team and project structures. He related this message through a story about how NASA is moving toward composite materials and the need to incorporate statistical thinking in that process:

*“We’re making investments in composite technologies for fuel tanks with pressurized hydrogen and oxygen. And one of the benefits would perhaps be a reduction in weight, right? Getting mass out of lower earth orbit is the biggest problem, and the lighter you can make the vehicles the better chance you have of carrying more payload into orbit. So you make the vehicle lighter and use composites. Well as it turns out, the engineers, being as conservative as we are as an agency putting humans in space, were designing our prototype composite structures to the same standards that we’re doing for aluminum lithium. So part of the culture that we have to overcome is, let’s change our standards. Let’s come up with different ways to approach engineering processes that don’t use and rely on 50 plus years of conservative engineering design.”*

Mr. Peri noted that statistical methods for quantifying risk are essential for thinking about design standards differently.

## Implementing Statistical Thinking in Organizations

Dr. Warner started the panel session talking about the changes in the DoD test and evaluation workforce in terms of the attitudes and widespread implementation of statistical methods. The dramatic change has been made possible by strong organization leadership, changes to standard processes, and the hiring of analysts with statistical backgrounds, all of which are challenges for organizations.

Additionally, these methods are not as widespread in the DoD developmental test and evaluation community, as CAPT Matisoo noted:

*“Coming from a developmental test background like myself, that’s where you fly at the precise altitude and precise air speed and you can control all those conditions, that’s great for system characterization with DOE, ... but I think we have yet to convince developmental test activities of the value of design of experiments.”*



---

He emphasized that “design of experiments actually is a tremendous benefit to understanding the system performance,” and that the developmental community will start using these methods once the operational community shows them the benefits.

Mr. Peri noted the difficulties that NASA has seen in incorporating the relatively new discipline of statistical engineering into their processes:

*“When we try to look at some of these new disciplines, it’s very challenging for us to adopt that into our engineering methodologies in a wholehearted way because we don’t have a lot of different projects going on. As a research organization [NASA Langley Research Center], it’s a little bit easier for us because we do have some flexibility, and I’m sure you know Peter Parker’s work [Statistical Engineering Group] is kind of at the leading edge of the kind of work we’ve been trying to do within the center, but taking that and scaling that to a broader engineering discipline has been very challenging for us.”*

Mr. Olinger noted the challenges with manpower:

*“We really have a challenge from an Air Force perspective to get enough analysts and engineers on our teams. Our ideal preference is that we have one on each team, but sometimes that one is also designated on two other teams so they’re really spread thin, so trying to get that test design, that analytical input into our test can really be a challenge simply because of manpower.”*

Relating organizational change to clear communication, Mr. Peri also noted that it is, “really important for the senior managers to understand the benefits and try to fold those into our mainline activities.”

## **Concluding Thoughts**

The overarching comments from the panelists provided a lot of material for an engaged question-and-answer session with the analytical audience. The panel emphasized the importance of such cross-organizational dialogs to share lessons learned, training resources, and motivating examples of why incorporating statistical thinking into organizational processes is critical for the future of our organizations.

**IDA** | OPERATIONAL EVALUATION DIVISION

---

## Summary of the Agenda

The workshop program contained a mix of training opportunities, leadership perspectives, case studies, and technical tracks. The first day of the workshop consisted solely of short courses designed to provide analysts with a baseline understanding of a new methodology. The next two days of the workshop consisted of leadership perspective discussions, mini-tutorials, and breakout sessions. In addition to the keynote address from Dr. J. Michael Gilmore and the leadership panel, there were three additional keynote speakers:

- Mr. Jon Holladay, NASA Technical Fellow for Systems Engineering, delivered a lunchtime keynote about the importance of the NASA Engineering Safety Center, highlighting current areas of emphasis.
- Dr. Christine Anderson Cook, a statistician at Los Alamos National Lab, gave the technical keynote where she discussed statistical tools for communicating with engineers and subject matter experts. She emphasized bringing them into the decision making process when selecting a specific test design.
- Dr. David Brown, Deputy Assistant Secretary of Defense, Developmental Test and Evaluation, delivered lunch keynotes on the importance of Design of Experiments in Developmental Test and Evaluation.

### Workshop Short Courses

The workshop provided practitioners in Defense and Aerospace with the opportunity to take one-day short courses from experts in subjects such as reliability, experimental design, regression analysis, and Bayesian statistics. These courses were taught by faculty members who are well known in their respective fields and in many cases even wrote the textbook used to teach the course at academic institutions. Each of the faculty members also made the course materials available to practitioners in the DoD and NASA; they are available online at <http://workshop.testscience.org/workshop-archives/>. Workshop short courses were on:

- *Experimental Design* by Dr. Doug Montgomery, Arizona State University.
- *Experiences in Reliability Analysis* by Dr. Bill Meeker, Iowa State University
- *Regression Modeling* by Dr. Geoff Vining, Virginia Tech
- *Introduction to Bayesian* by Dr. Robert Gramacy, University of Chicago

### Mini-Tutorials

Training opportunities continued throughout the workshop in the form of mini-tutorials on topics relevant to the community. Mini-tutorials provided practitioners with the opportunity to learn a new skill, understand best practices, and gain awareness of common analysis methods. The mini-tutorials are also available on the workshop website.

- 
- *Bootstrapping 101* – Dr. Matt Avery, IDA
  - *Software Reliability Tools and Models* – Dr. Lance Fiodinella, University of Massachusetts
  - *Censored Data Analysis for Performance Data* – Dr. Bram Lillard, IDA
  - *Statistical Power to Support Test Adequacy Decisions* – Dr. Jim Simpson, JK Analytics
  - *Survey Design and Analysis* – Dr. Heather Wojton, Mr. Jonathan Snaveley, IDA
  - *Bayesian Data Analysis in R/STAN* – Dr. Kassandra Fronczyk, IDA, and Dr. James Brownlow, US Air Force 812TSS/ENT
  - *Presenting Complex Statistical Methodologies to Military Leadership* – Mr. Paul Johnson, MCOTEA, Dr. Jane Pinelis, CNA
  - *Sensitivity Experiments*, Dr. Tom Johnson, IDA

## Breakout Sessions

Breakout sessions completed the list of sessions to which workshop attendees had access. The breakout sessions focused on case studies that illustrated how statistical methods and concepts were used to benefit Defense and Aerospace challenges. Session themes and complete abstracts, which are available at, <http://workshop.testscience.org/workshop-archives/>, included:

- Improving Reliability Assessment
- Experimental Design Methods and Applications
- Managing Model Uncertainty for Risk Analysis and System Safety
- Experimentation Involving Human Interactions
- Statistical Engineering in Government and Industry
- Bayesian Methods for Improving Understanding
- Statistical Engineering for Improved Outcomes
- Improved System Characterizations
- Importance of Modeling and Simulation
- Emerging Methods for Cybersecurity Testing

Speakers in the breakout sessions represented DoD, NASA, industry, academia, National Labs, FFRDCs, and NIST – making for a very diverse conversation.



## Workshop Attendee Feedback and Future Directions

A workshop addressing the scientific approaches to Test and Evaluation (T&E) would not be complete without a scientific survey to gather participant feedback. Of the 200 attendees, 76 provided feedback via an email survey. The respondents represented a wide range of organizational affiliations, with Federally Funded Research and Development Centers (FFRDC) having the highest representation (34.3%), followed by Department of Defense (DoD; 24.3%), and National Aeronautics and Space Administration (NASA; 20.0%). The majority of the respondents reported roles as analysts, statisticians, practitioners, or researchers (72.0%) within their organizations. The most common method of learning about the event was by way of word of mouth (65.0%) from coworkers, followed by workplace announcements (22.5%).

### Attendance Motivation

Professional competence was the number one reason for attendance, with 53 respondents (69.7%) reporting it as a primary reason and 14 (18.4%) reporting it as a secondary reason. Personal interest followed as the second most highly rated reason, with 39 (51.3%) primary reason responses and 28 (36.8%) secondary reason responses. Networking had more secondary (42; 55.3%) than primary (25; 32.9%) reason responses, but still ranked highly as a motivating factor. Professional service and education credits were more commonly categorized as “Not a Reason” (47 and 67, or 61.8% and 88.2%, respectively). Figure 1 shows the full rating frequencies.

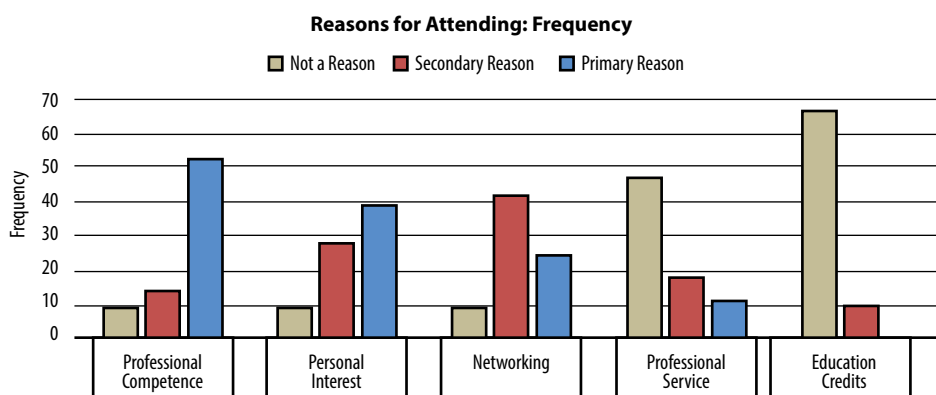


Figure 1. Reasons for Attending versus Priority

### Overall Event Quality and Presentation Satisfaction

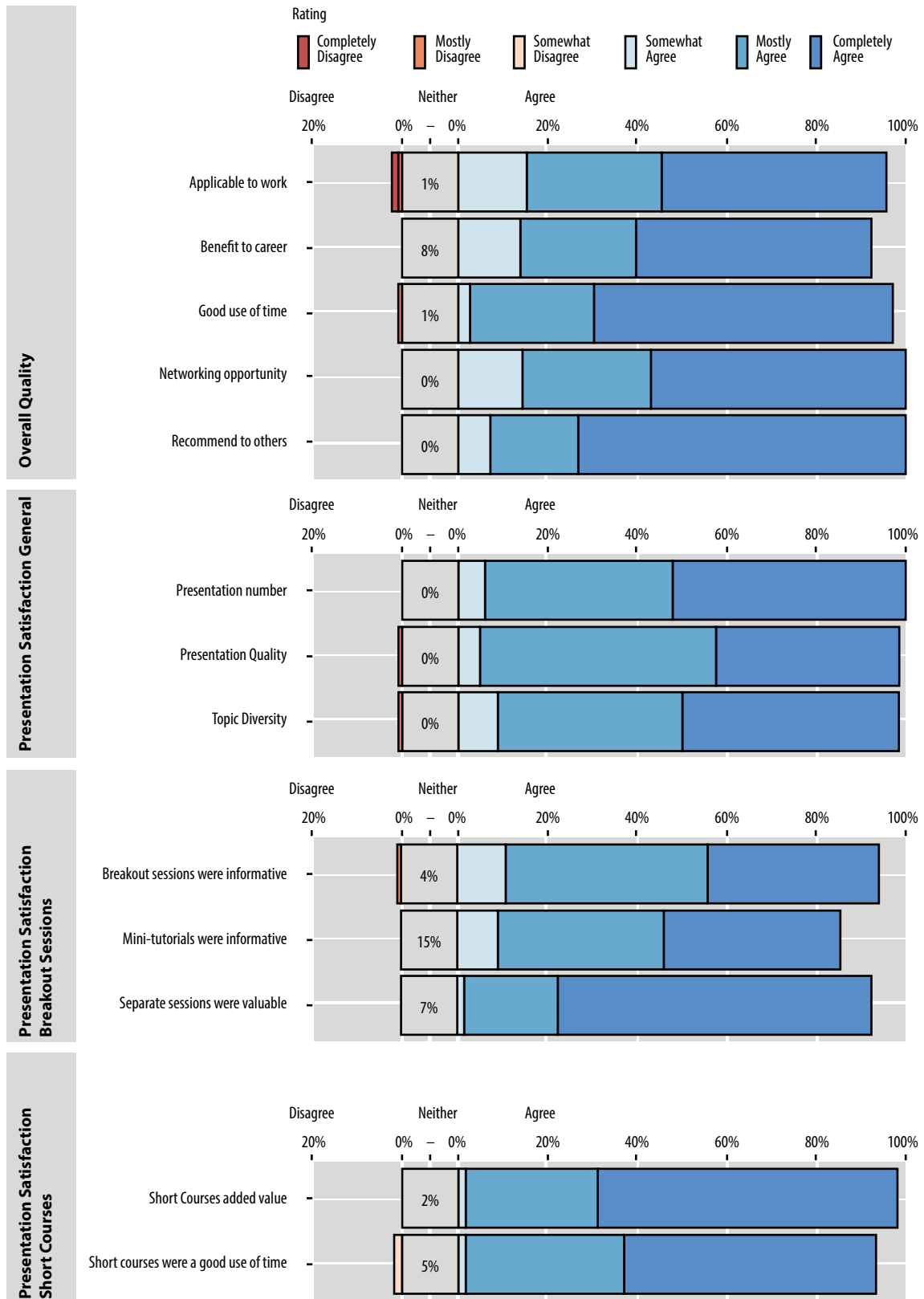
The feedback on the workshop was overwhelmingly positive. The survey posed questions about whether participants thought the workshop was a good use of time, whether they would recommend it to others, whether it was applicable to their work, whether it was a good networking opportunity, and whether they thought it would benefit their career. Each of the questions had median ratings equivalent to agree or strongly agree, indicating an

---

overwhelming majority of positive opinions about the event overall. Only one (1.4%) of 73 respondents did not agree that the event was a good use of time, and only two (2.8%) of 72 respondents indicated that the information was not applicable to their work. Figure 2 shows the percentages of responses by response category for overall quality metrics, presentation satisfaction, and session type satisfaction. Satisfaction with presentation quality, topic diversity, and number of presentations was also uniformly positive. Of the 29 general comments received in the survey, 20 people (69%) made positive statements about the event, including appreciation for the meals and logistics, stating a desire to attend the event annually, and recommending that the workshop continue its efforts to attract young analysts.

A common “negative” comment about the event was that there were too many overlapping sessions with content that attendees wanted to learn.

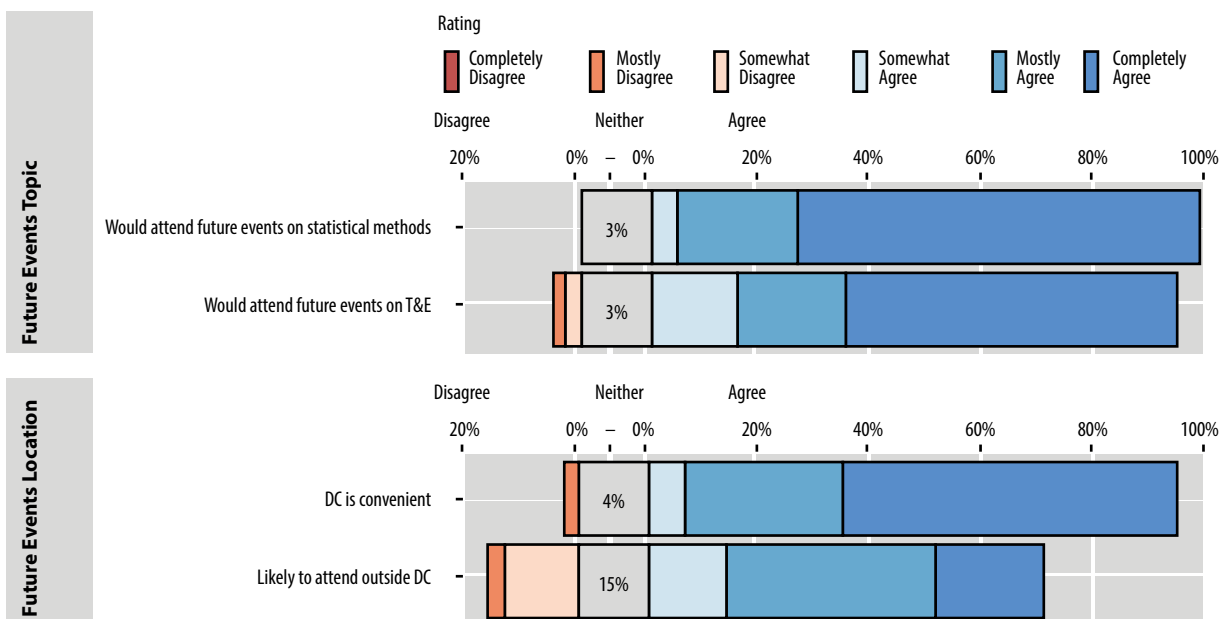
One unique aspect of the workshop was that participants could opt either to attend traditional conference-style breakout sessions or follow a tutorial track with longer mini-tutorial topics. Both the breakout sessions (95% Agree) and mini-tutorials (85% Agree) were rated as informative, and respondents recommended that the workshop continue to offer both sessions separately (93% Agree). Only one (1.4%) of the respondents was dissatisfied with the breakout sessions and completely disagreed that they were informative. Ratings were similar for short courses, with 94% agreeing that they were a good use of time. Ninety-eight percent agreed that the short courses added value to the event as a whole. Of the 33 comments on presentations, 17 (51.5%) were positive and 16 (48.5%) offered suggestions for the future. The primary suggestion involved increasing the time allowed for short courses and breakout sessions. The comments also showcased the diversity of analysts attending the event. Some respondents noted a gap in practical material and too much emphasis on theory. Other respondents, however, suggested that the workshop emphasized practical examples too much. They noted that practical examples were interesting but less compelling than quantitative presentations. Another interesting theme from the survey comments was the idea that the utility of design of experiments (DOE) is now largely accepted in the test community and the time spent “selling” the approach could be replaced with more advanced content.



**Figure 2. Event Quality Ratings**

## Lessons Learned for Planning Future Events

Participants clearly indicated that they would be interested in attending future events on both statistical methods and test and evaluation (97% and 93% Agree, respectively). However, there was greater variability in the question about future locations for the workshop. Eleven percent of respondents disagreed that DC was convenient, but 31 percent indicated that they would be unlikely to attend an event held outside the area. Several comments mentioned that scheduling the event in a less congested, cheaper area, or at least scheduling the event during a non-tourist season, would be preferable. Figure 3 shows the breakdown of responses concerning future meetings.



**Figure 3. Likelihood of Attending Future Events**

Respondents provided a substantial amount of feedback on potential improvements for future meetings and potential future topics. Additionally, a few enthusiastic respondents volunteered to help organize the next meeting. Recurring themes for improvements included:

- Improved focus of content for the keynote and leadership talks
- More time for short courses
- More advanced content
- Less theoretical and more practical content
- Larger session rooms and space for conversation
- Moderators to keep presenters on schedule

Seventeen suggestions were offered regarding future topics. As shown in Table 1, these primarily centered on general analysis techniques or specific applications of those techniques:

**Table 1. Suggestions for Future Topics**

General Analysis Techniques	Specific Applications
• Anything related to test and evaluation	• Cybersecurity
• Combining information/use of developmental testing in future evaluations	• Design of computer experiments and deterministic modeling
• Model validation	• Software reliability
• Data analysis	• High-reliability system testing
• Presentation of statistical results to general audiences	• Test and evaluation activities early in the acquisition process
• Statistical engineering	• Use of simulation in Bayesian prior construction

The comments also indicated that attendees were not interested in general acquisition strategy:

- “We get a lot on acquisition thru DAU, so I feel that would be duplicative. There are other ‘science’ workshops, but I don’t feel that they are specifically focused on science in T&E. If there are other workshops to the level of this one, please pass along the info.”
- “Please....not at all interested in general acquisition topics....let’s keep this cutting edge for analysts.”

## Conclusions

Overall, the survey reinforced that the event was much needed in the DoD and NASA analytical communities. The high response rate, enthusiastic comments, and clear desire to attend such events in the future reinforce that there will continue to be a need for a workshop that emphasizes statistically rigorous approaches to the acquisition process at all stages. Furthermore, positive responses indicate that the event was executed professionally and without major areas in need of revision. Challenges for the future include determining general technical topics of interest without focusing too heavily on specific statistical methods, and maintaining the high caliber of presentations that was achieved at this event. The test and evaluation community comprises highly skilled and specialized professionals who deserve opportunities provided by gatherings such as the Science of Test Workshop. Future events are clearly desired and will continue to enhance the community’s professionalism, competence, and sense of identity.



**IDA** | OPERATIONAL EVALUATION DIVISION





© Institute for Defense Analyses

4850 Mark Center Drive • Alexandria, VA 22311-1882

[www.ida.org](http://www.ida.org)

 @ida\_org

IDA Document NS D-8249

