INSTITUTE FOR DEFENSE ANALYSES

Test Design Challenges in Defense Testing

Rebecca Medlin, Project Leader

Kelly Avery
Curtis Miller

July 2021

Approved for Public Release.
Distribution Unlimited.

IDA Document NS D-22723

Log: 2021-000241

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

The Institute for Defense Analyses is a nonprofit corporation that operates three Federally Funded Research and Development Centers. Its mission is to answer the most challenging U.S. security and science policy questions with objective analysis, leveraging extraordinary scientific, technical, and analytic expertise.

About This Publication
This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-19-D-0001, Task C9082 "Cross-Divisional Statistics and Data Science Working Group". The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Acknowledgments
The IDA Technical Review Committee was chaired by Mr. Robert R. Soule and consisted of John Haman, Han Yi, and Rebecca Medlin from the Operational Evaluation Division.

For more information:
Rebecca Medlin, Project Leader
rmedlin@ida.org • 703-845-6731

Robert R. Soule, Director, Operational Evaluation Division
rsoule@ida.org • (703) 845-2482

Rigorous Analysis | Trusted Expertise | Service to the Nation

INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS D-22723

Test Design Challenges in Defense Testing

Rebecca Medlin, Project Leader

Kelly Avery
Curtis Miller

Executive Summary

All systems undergo operational testing before fielding or full-rate production. While contractor and developmental testing tends to be requirements-driven, operational testing focuses on mission success. The goal is to evaluate operational effectiveness and suitability in the context of a realistic environment with representative users.

In modern defense testing, modeling and simulation (M&S) capabilities are often critical to fully characterizing a system's capabilities. The complexity of modern military systems and the environment in which they operate means that live testing is often expensive or even impossible; certain threats or combat scenarios simply cannot be reproduced on test ranges. M&S tools are undeniably valuable but, to ensure that they produce trustworthy results, their behavior and accuracy must be well understood in relation to their intended use.

Although classical experimental design techniques have been widely adopted across the defense community for planning live tests, gold-standard computer experiment techniques from the academic literature – such as those that use space-filling designs and Gaussian process emulators – are underused. Space-filling design techniques can significantly lower the risk of mis-estimating the response surface of the model of interest by placing samples throughout the parameter space to better capture local deviations from linearity.

Defense testing poses unique demands, such as a heavy reliance on categorical factors and binary outcomes, the mandate to judge the adequacy of sample size, extreme constraints in test conditions, and non-deterministic M&S outputs. There is currently no consensus on how to incorporate these demands into the existing academic framework for M&S. In addition, Gaussian processes can be more challenging than traditional statistical analysis techniques to implement and explain.

This brief will first provide an overview of operational testing and discuss example defense applications of, and key differences between, classical and space-filling designs. It will then present several challenges (and possible solutions) associated with implementing space-filling designs and associated analyses in the defense community.

IDA

Test Design Challenges in Defense Testing

Kelly Avery, Curtis Miller, Han Yi

July 2021

Institute for Defense Analyses
4850 Mark Center Drive ● Alexandria, Virginia 22311-1882

Testing in DoD

All military systems undergo operational testing before fielding or full-rate production…

...Even the ones you don't normally think of

- Biometrics systems
- Personnel management systems
- Logistics and readiness systems
- Command & control systems
- Pilot trainers

DoD test paradigm

Contractor Testing

Developmental Testing

Operational Testing

Test Timeline

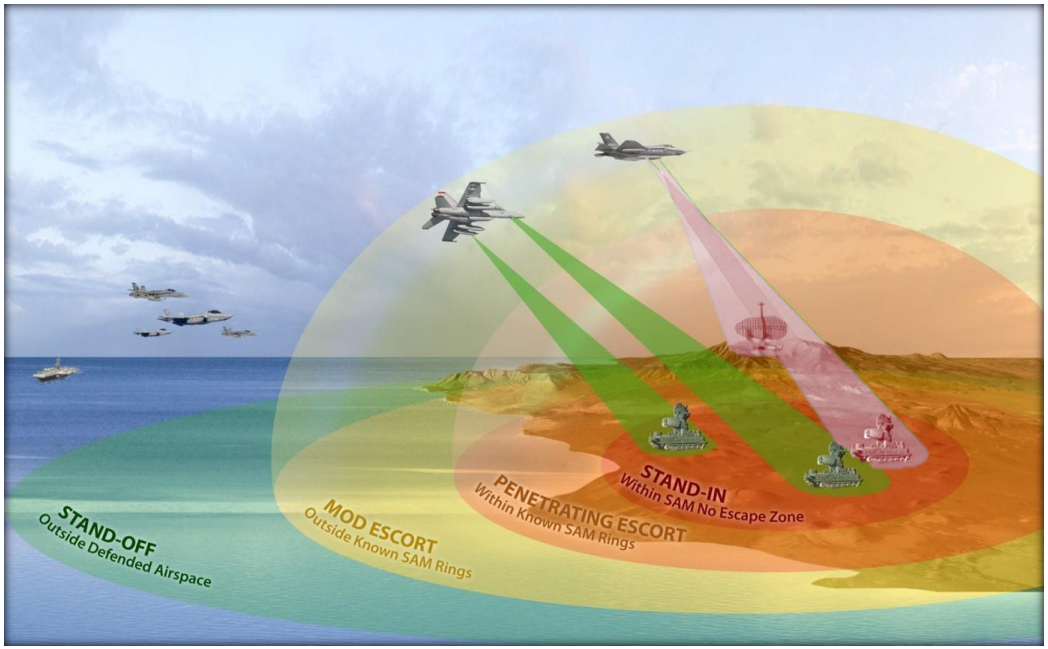
Tend to be requirements driven

Focuses on mission success

Requirements documents are often missing important mission considerations

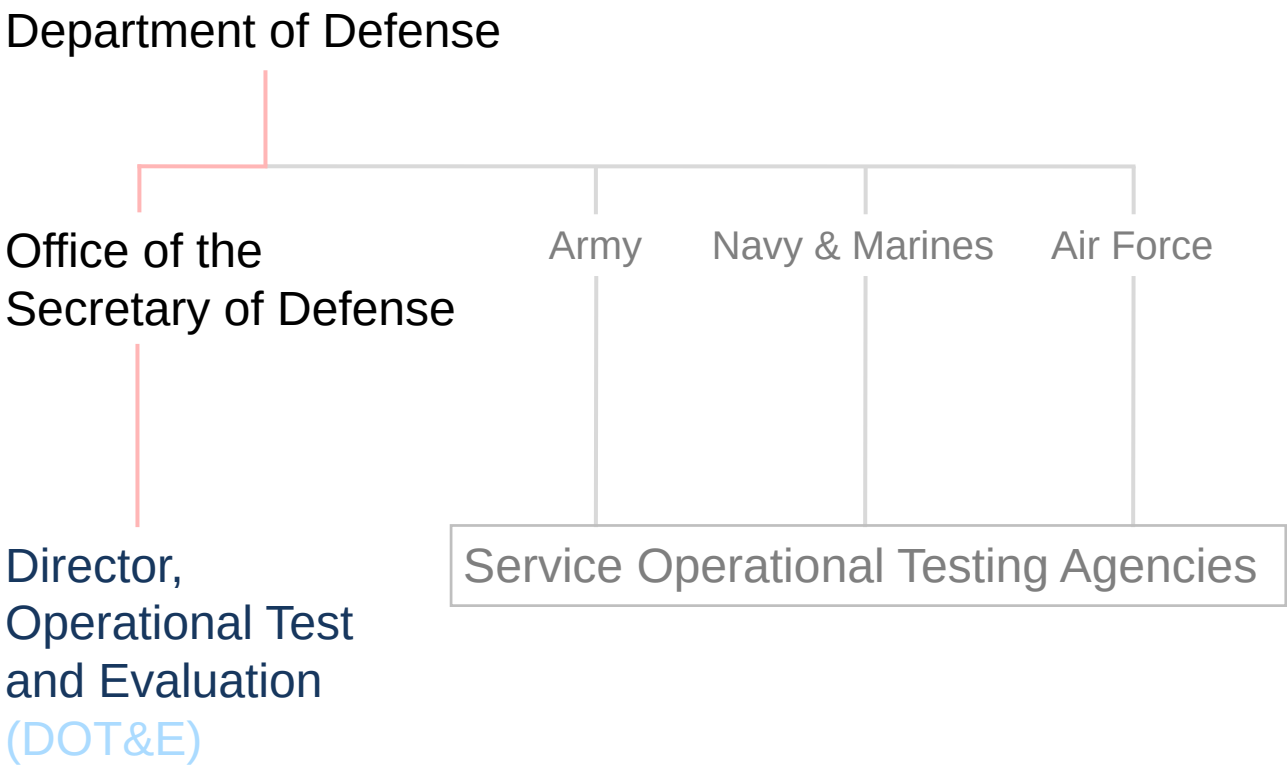
Goal of operational test: evaluate operational effectiveness, suitability, and survivability/lethality

- Operational Environment
- Representative Users
- "Real" Threats
- Conducting Missions

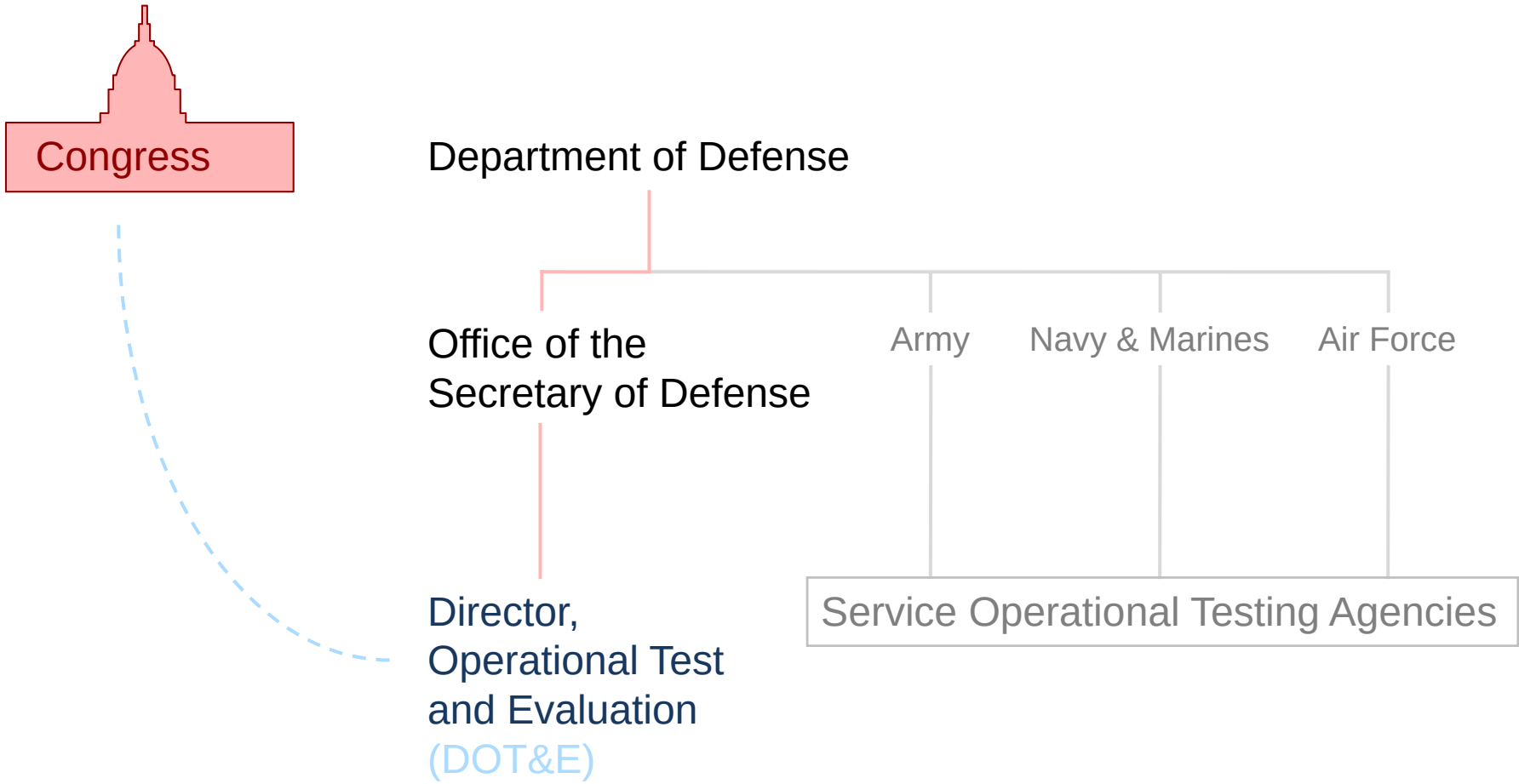


Congress established DOT&E separate from the Services' operational testing agencies

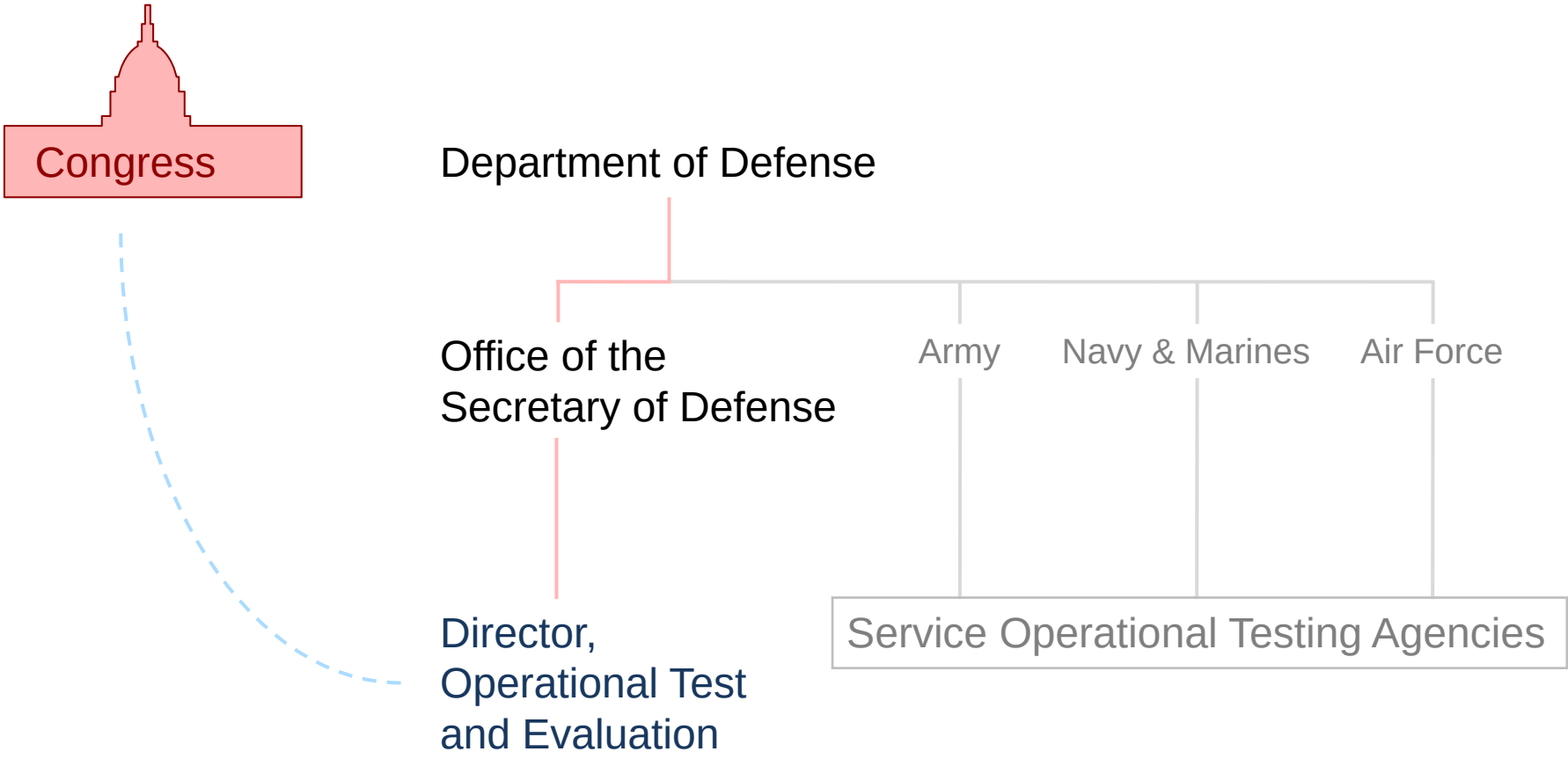
Congress established DOT&E separate from the Services' operational testing agencies



Congress established DOT&E separate from the Services' operational testing agencies

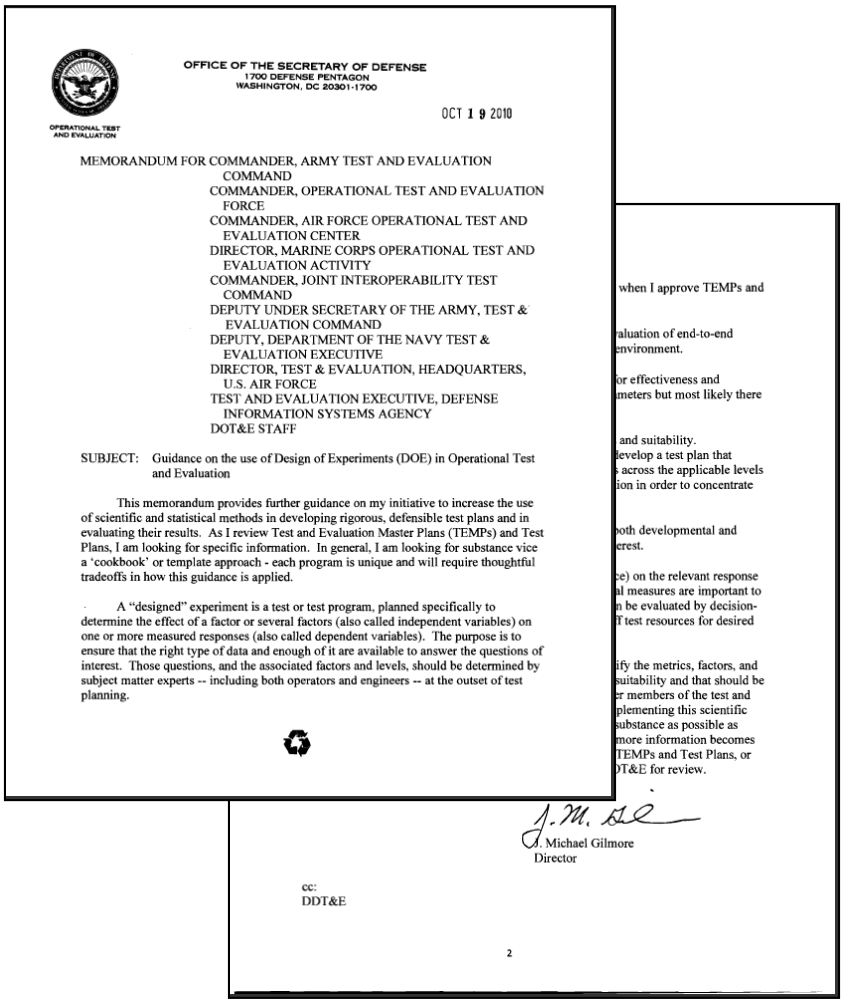


Congress established DOT&E separate from the Services' operational testing agencies



IDA

DOT&E sets policy and guidance for conducting operational testing; statistical rigor has been a point of emphasis over the past several years



OFFICE OF THE SECRETARY OF DEFENSE
1700 DEFENSE PENTAGON
WASHINGTON, DC 20301-1700

OPERATIONAL TEST AND EVALUATION

OCT 19 2010

MEMORANDUM FOR COMMANDER, ARMY TEST AND EVALUATION COMMAND
COMMANDER, OPERATIONAL TEST AND EVALUATION FORCE
COMMANDER, AIR FORCE OPERATIONAL TEST AND EVALUATION CENTER
DIRECTOR, MARINE CORPS OPERATIONAL TEST AND EVALUATION ACTIVITY
COMMANDER, JOINT INTEROPERABILITY TEST COMMAND
DEPUTY UNDER SECRETARY OF THE ARMY, TEST & EVALUATION COMMAND
DEPUTY, DEPARTMENT OF THE NAVY TEST & EVALUATION EXECUTIVE
DIRECTOR, TEST & EVALUATION, HEADQUARTERS, U.S. AIR FORCE
TEST AND EVALUATION EXECUTIVE, DEFENSE INFORMATION SYSTEMS AGENCY
DOT&E STAFF

SUBJECT: Guidance on the use of Design of Experiments (DOE) in Operational Test and Evaluation

This memorandum provides further guidance on my initiative to increase the use of scientific and statistical methods in developing rigorous, defensible test plans and in evaluating their results. As I review Test and Evaluation Master Plans (TEMPs) and Test Plans, I am looking for specific information. In general, I am looking for substance vice a 'cookbook' or template approach - each program is unique and will require thoughtful tradeoffs in how this guidance is applied.

A "designed" experiment is a test or test program, planned specifically to determine the effect of a factor or several factors (also called independent variables) on one or more measured responses (also called dependent variables). The purpose is to ensure that the right type of data and enough of it are available to answer the questions of interest. Those questions, and the associated factors and levels, should be determined by subject matter experts -- including both operators and engineers -- at the outset of test planning.

when I approve TEMPs and

aluation of end-to-end environment.

or effectiveness and meters but most likely there

and suitability. levelop a test plan that across the applicable levels ion in order to concentrate

both developmental and erest.

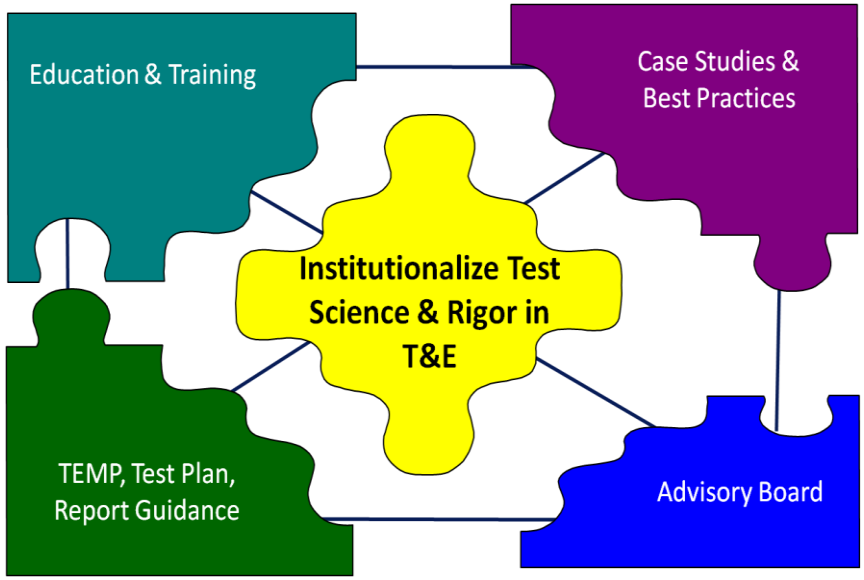
e) on the relevant response al measures are important to n be evaluated by decision- T test resources for desired

ify the metrics, factors, and suitability and that should be r members of the test and plementing this scientific substance as possible as more information becomes TEMPs and Test Plans, or OT&E for review.

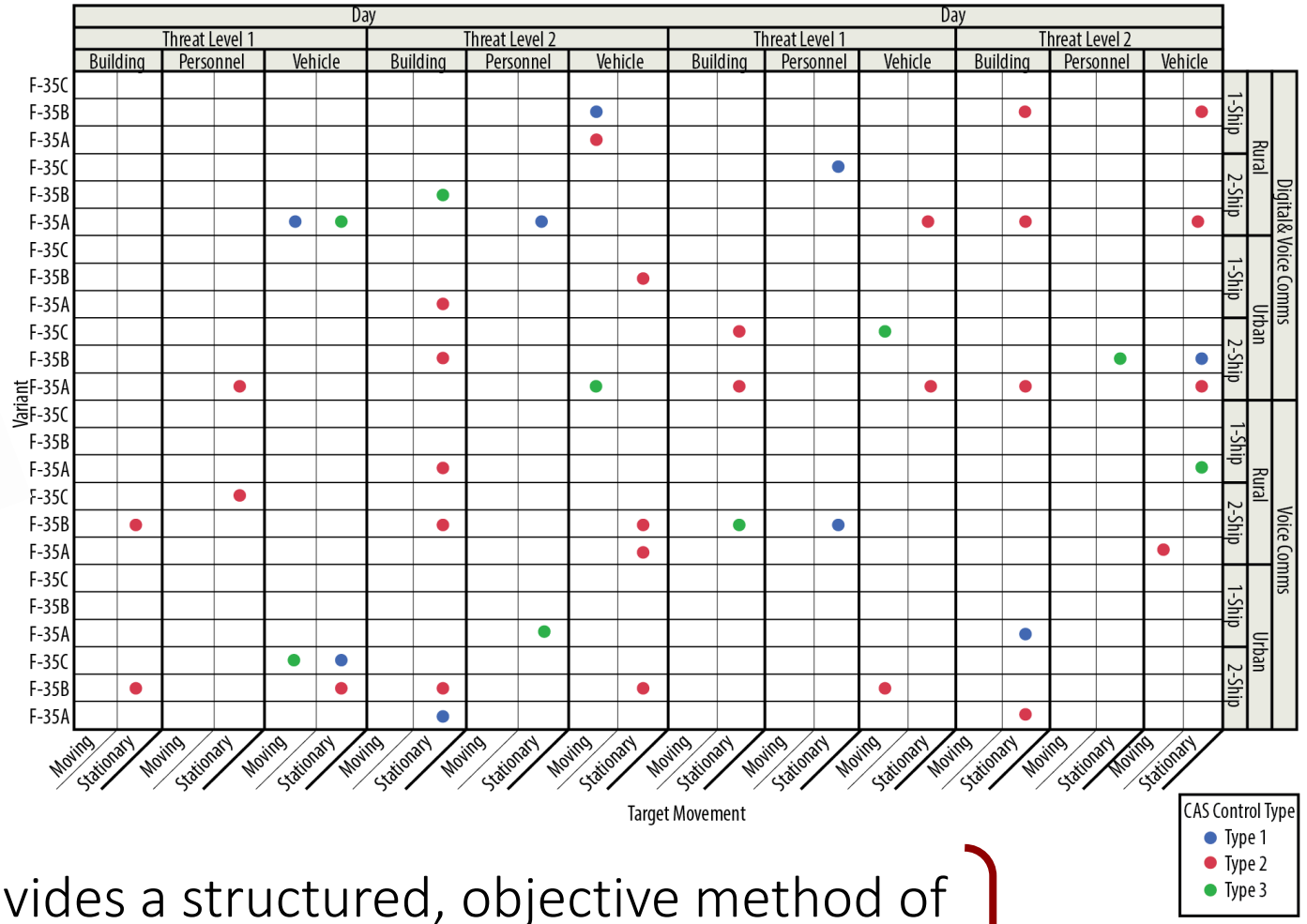
J. Michael Gilmore
Director

cc:
DDT&E

2



Classical Design of Experiments techniques are commonly used to efficiently collect data from live test events



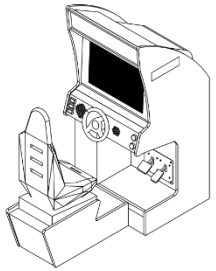
DOE provides a structured, objective method of choosing test points and assessing risk

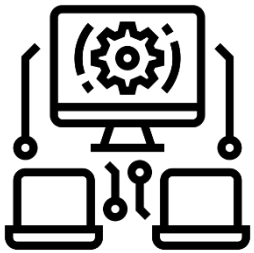
Modeling & Simulation (M&S) Validation

Evaluations of systems increasingly rely on M&S to supplement the data collected from live test events

Examples of M&S used in T&E:

- Digital computer models
- Hardware-in-the-loop simulators
- Threat emulators





Before results from M&S are used in DOT&E reports, we want confidence that these numbers mean something!

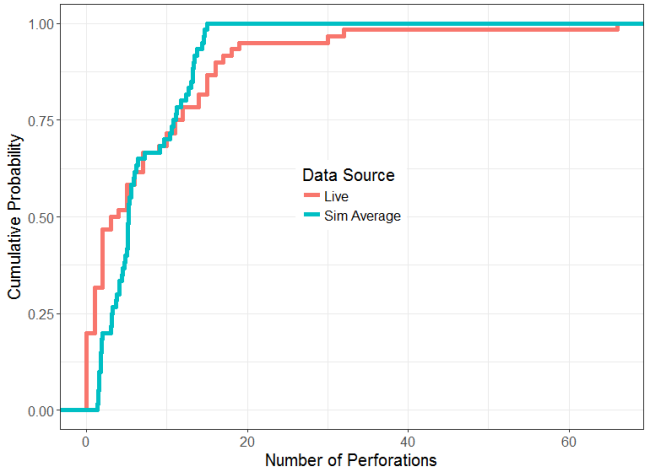
The process of establishing credibility and trust in a model is called Verification, Validation, and Accreditation (VV&A).

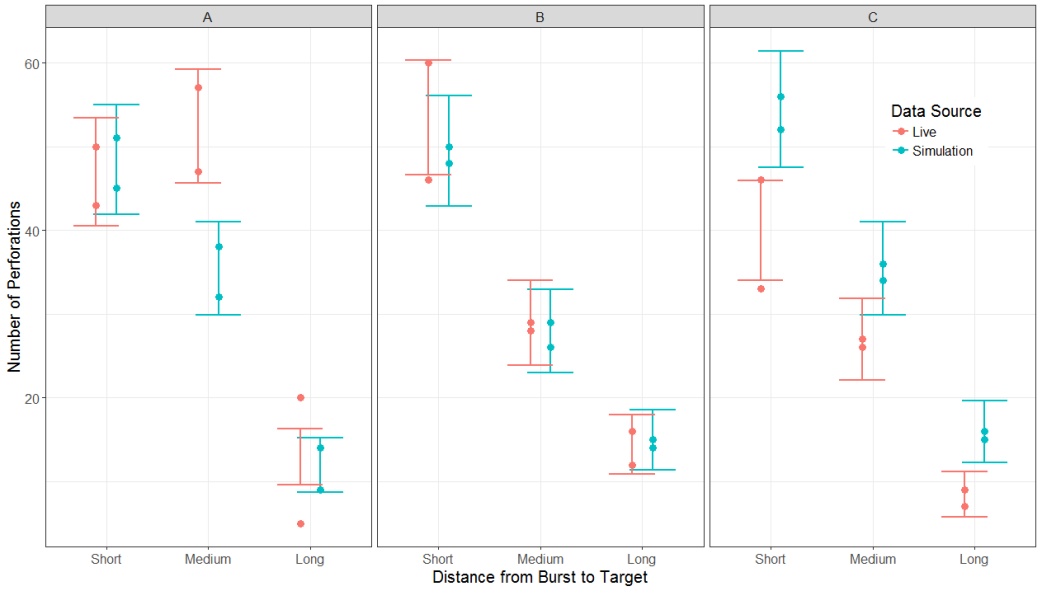
The VV&A process should provide a quantitative understanding of how accurate an M&S capability is and identify limitations of the M&S across the factor space of interest.

Goal 1: Determine whether the M&S output "matches" live data

- Basic statistical tests can reveal overarching differences between M&S and live data

- Regression analyses can quantitatively determine under which conditions the M&S output is (in)consistent with live results



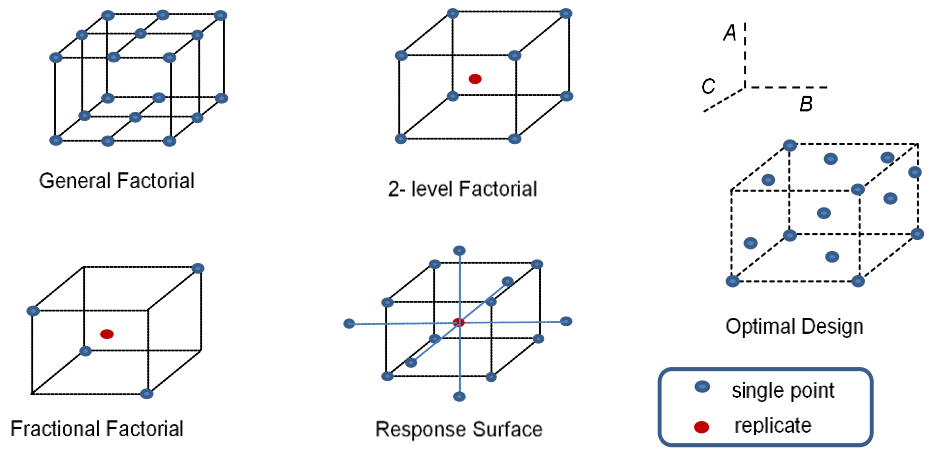


Test design(s) for M&S should facilitate this comparison with available live test data

Design Properties:

- Match the live test points (possibly with replicates)
- Support building a statistical model
- High power to detect differences between live event and M&S

Classical DOE is recommended

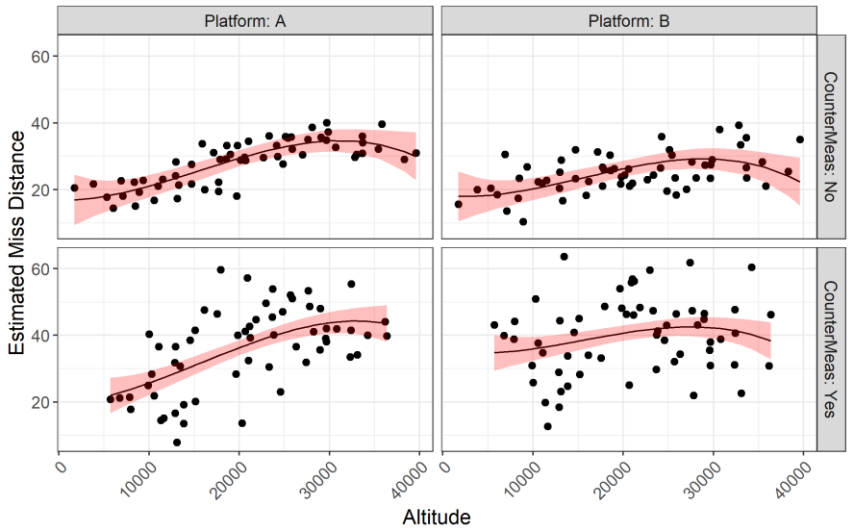


These design and analysis techniques are already widely understood and used

Goal 2: Explore and evaluate the behavior of the model itself; be able to make predictions and quantify uncertainty across the entire space

- Variation and sensitivity analyses can be used to test for model robustness, scope test designs, and identify risk areas.

- Emulators (meta-models) can be used to predict the output of the simulation under both tested and untested conditions
  - In some cases, may also serve as a surrogate for the M&S itself, which can save time and cut costs by avoiding the need to re-run the simulation over and over

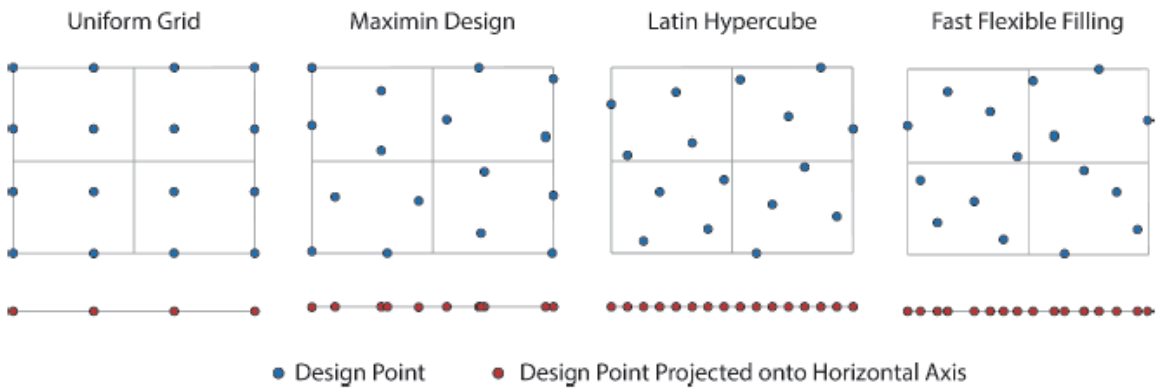


Test design(s) for M&S should also facilitate this thorough understanding of the model itself

Design Properties

- Fill the M&S space
- Consider all input parameters (even those not able to be varied in live test)
- Consider the type of output expected (e.g., nonlinearities, curvature) when deciding on the number of points

Space-Filling Designs are recommended



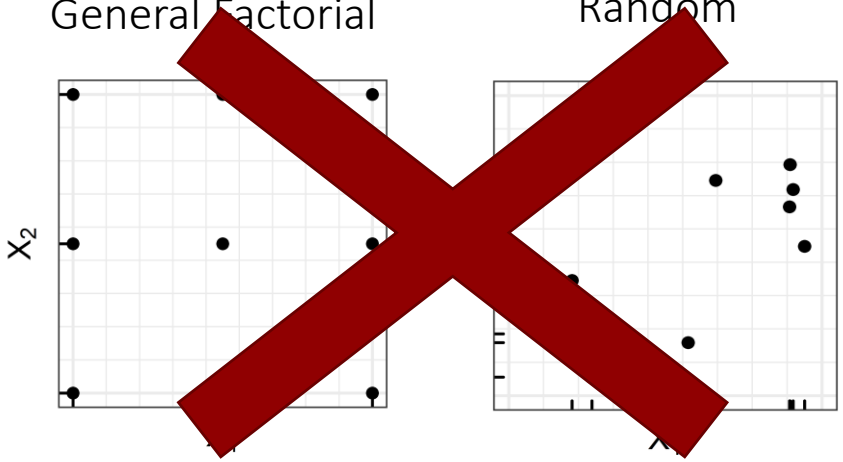
Space-Filling Designs and associated analyses are lesser known and infrequently used in the T&E community

Space-Filling Designs (SFDs)

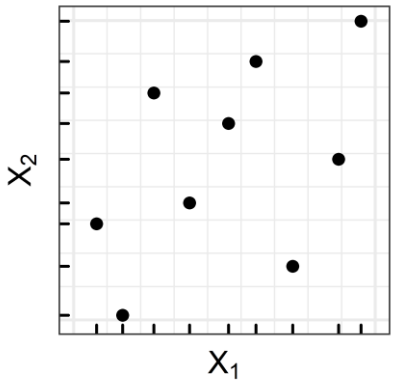
There are several common types of SFD in the literature, some of which can be useful for M&S validation

General Factorial

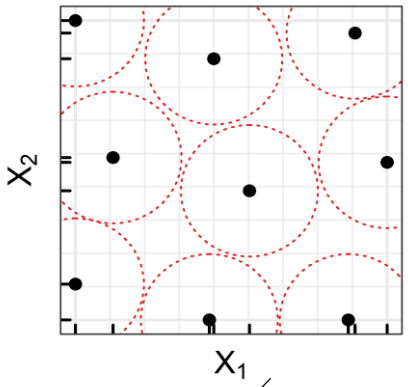
Random



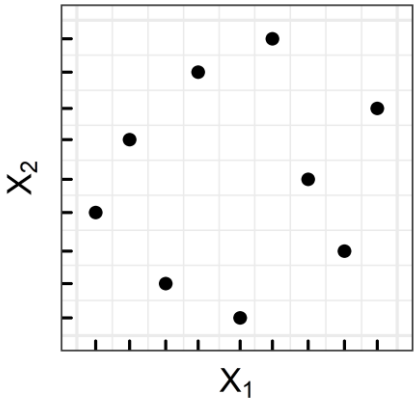
Latin HyperCube Sampling (LHS)



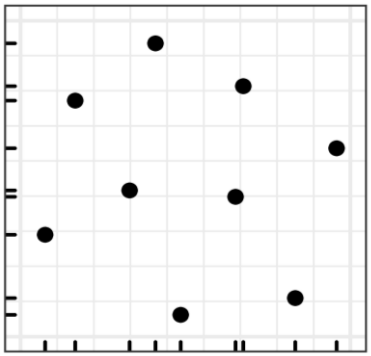
Maximin



Uniform



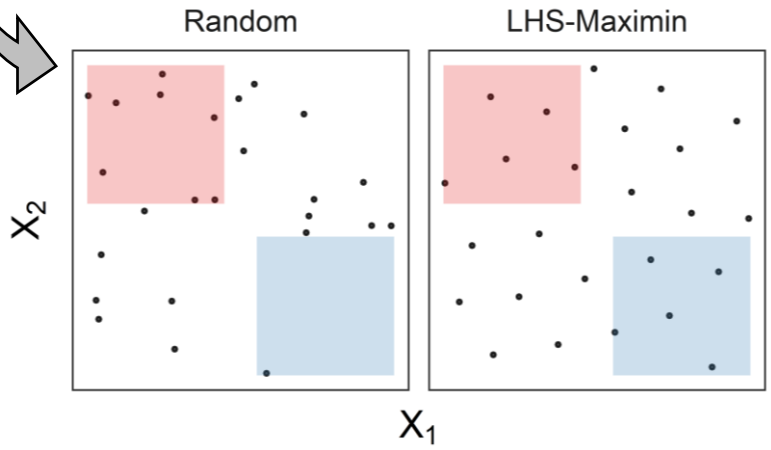
Maximin-Optimized LHS

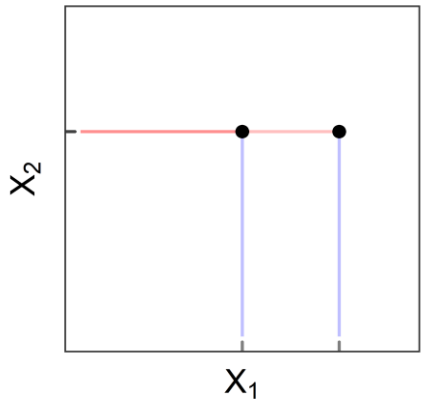


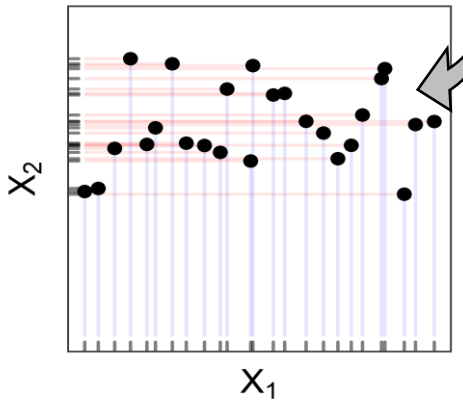
Just like with classical DOE, there are quantitative ways to evaluate a specific design

Many criteria exist, but it is particularly important that an SFD satisfy the following three criteria in order to be useful:

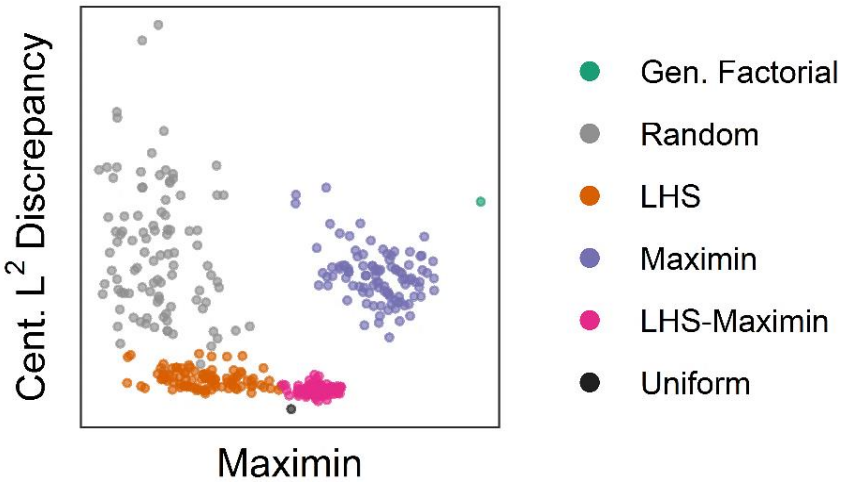
- Point-distance: Samples are placed as far apart from each other as possible. [Maximin]
- Uniformity: All regions of the design space are equally well represented. [Center L^2 Discrepancy]
- Projection: The design is robust to variables being collapsed. [MaxPro]

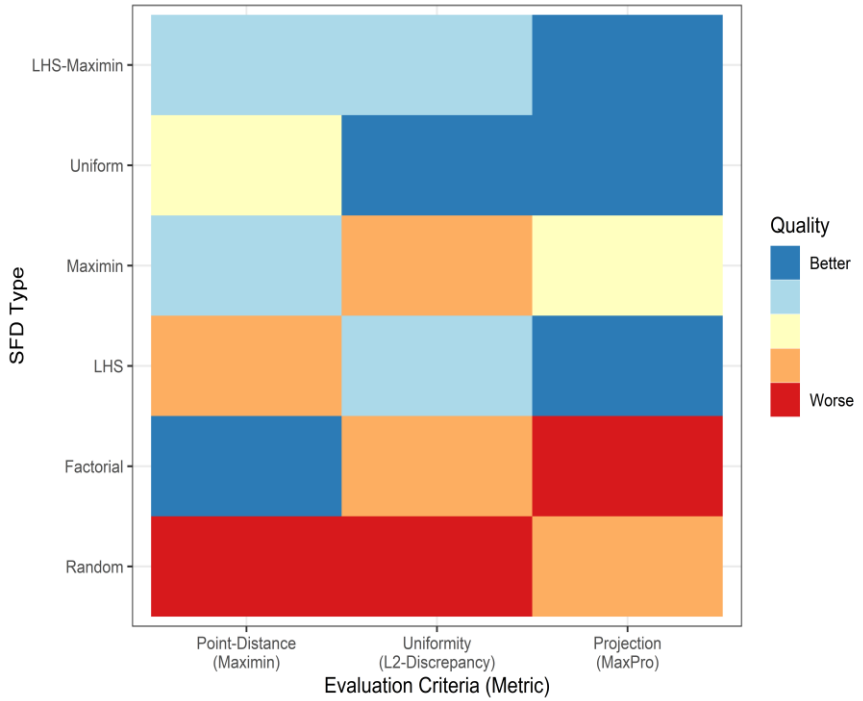






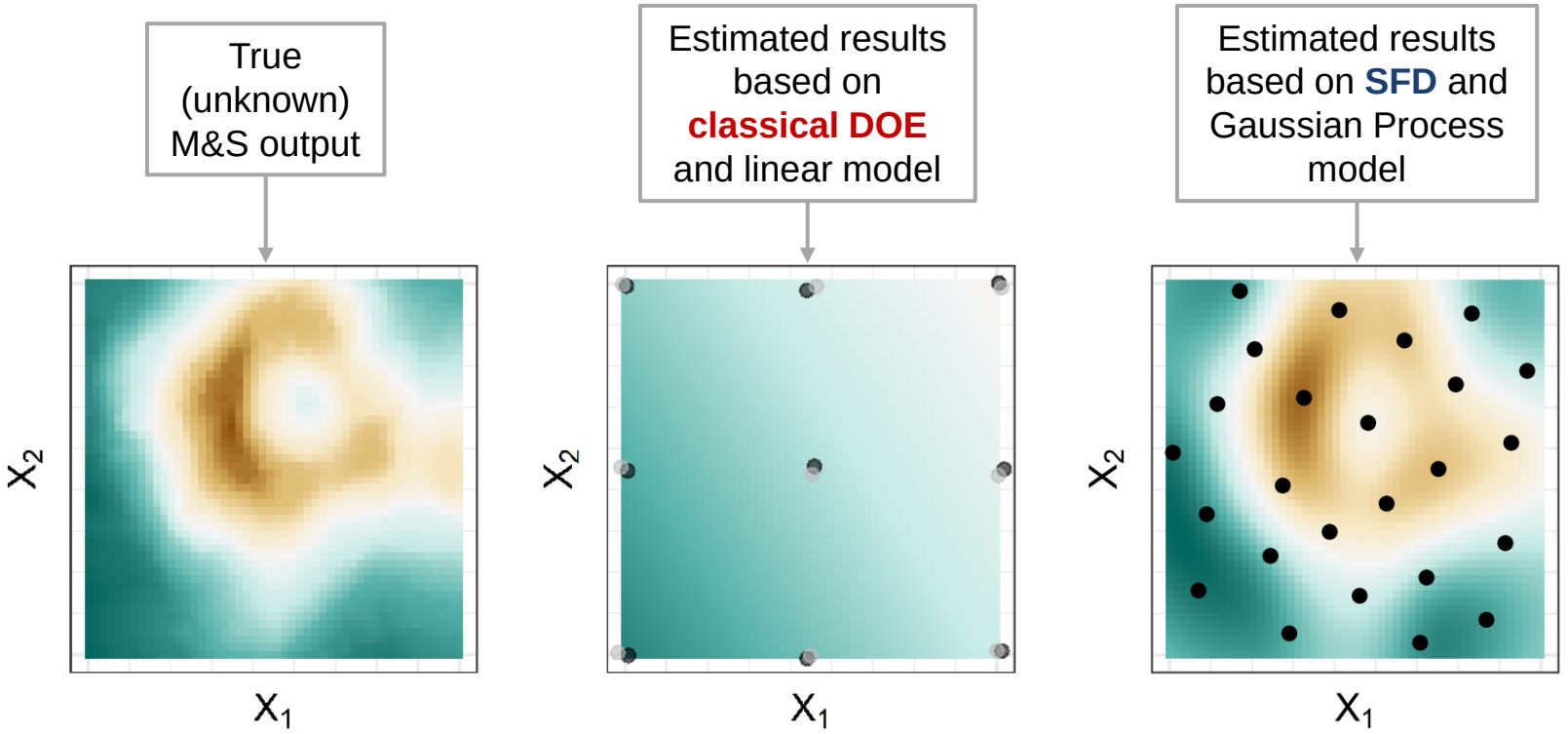
These criteria can be used to compare classes of designs and make general recommendations





Recommendations: LHS-Maximin or Uniform

SFDs capture M&S behavior more effectively than classical DOE without requiring additional test resources



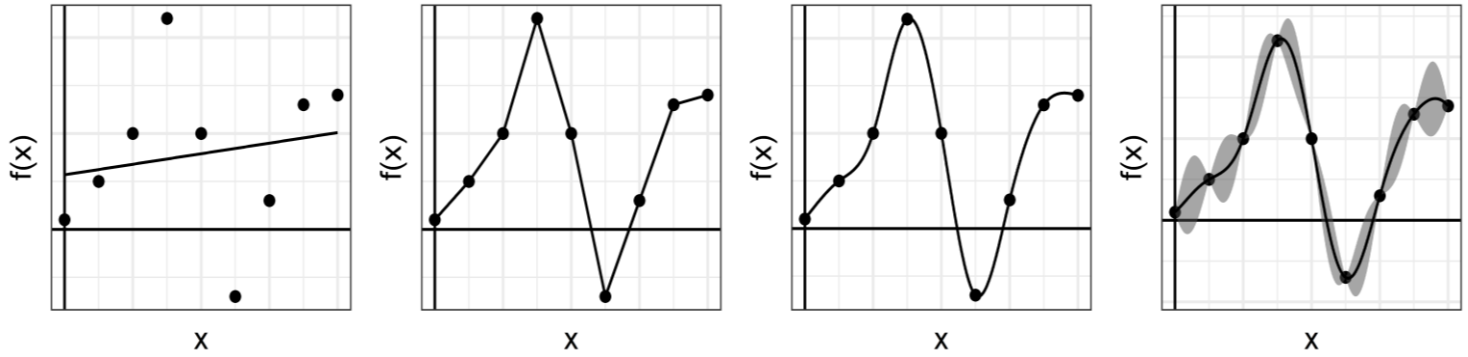
Failing to understand M&S behavior means DOT&E may include inaccurate predictions about system performance in their reports

Challenges

Statistical methods for analyzing data collected from SFD are different from those used with classical DOE

Goals: Interpolate across a complex space

Quantify uncertainty at observed and unobserved points



Linear Regression

Basic Interpolators and Splines

Gaussian Process Regression

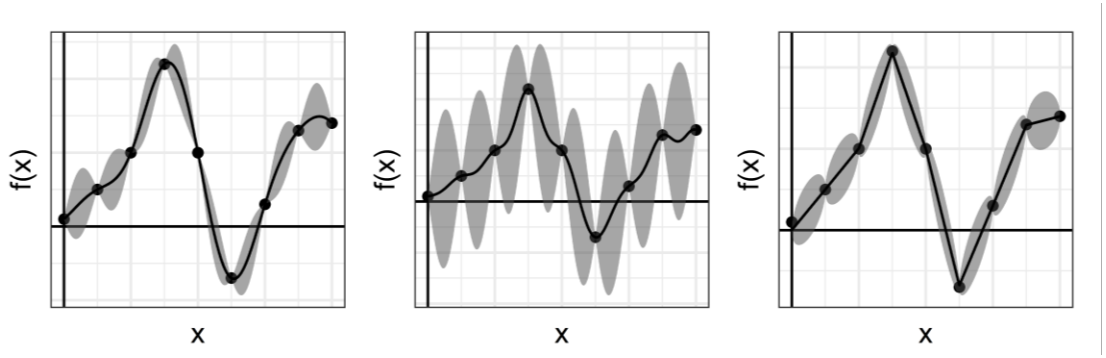
Gaussian Process (GP) regression can be challenging to implement and explain!

GP regression refers to probability models that describe randomly generated functions that follow a Gaussian distribution

$$Y \sim \mathcal{N}_n(0, \Sigma_n)$$

Uncertainty Quantification can be achieved by assigning a probability of matching the true response function to multiple interpolation functions.

- The properties of the uncertainty estimates from GP are determined by two parameters: the mean function and the covariance function
- Practitioners must have some belief about what the mean and covariance functions could be; estimating parameters may be difficult



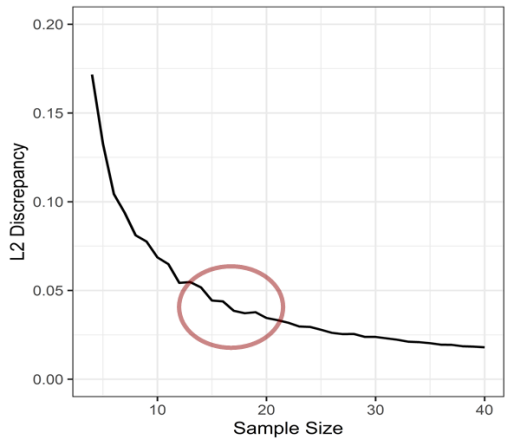
What tools and trainings are best for getting the T&E community equipped to implement GP Regression?

The SFD literature lacks rigorous, quantitative methods for sample size determinization

Power and confidence are not appropriate metrics for SFD (or any other techniques for deterministic outputs)

How do we determine what sample size is "adequate" for a Space-Filling Design?

- Rule of thumb: 10*# of dimensions
- Scree plots:



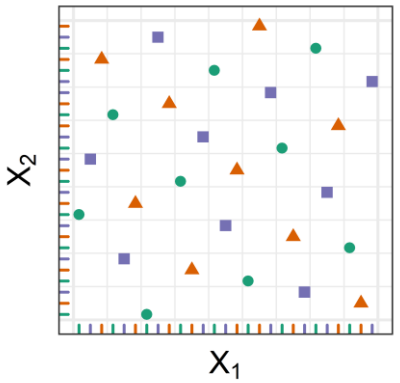
Operational and live fire tests often involve variables that are categorical or binary rather than continuous

Most SFD and Gaussian Process methods are limited to continuous inputs and outputs…

What are the best design and analysis techniques for M&S with categorical or binary inputs/outputs?

Some options available for categorical inputs:

- Fast Flexible Filling
- Sliced Latin Hypercube Sampling



How to analyze binary outputs is less clear…

- Logistic Regression?
- Generalized Additive Models?

Many of the M&S capabilities used in T&E are not deterministic (and live data are definitely not!)

Gaussian Process regression and other computer experiment techniques assume the output is either deterministic or has a small amount of randomness (e.g., from Monte Carlo draws)

How do we handle test designs for non-deterministic M&S output with high levels of noise? Or for comparing any M&S to highly stochastic live data?

SFD with replicates?

Hybrid designs (e.g., SFD overlaid with an optimal design)?

Classical DOE?

Other Challenges

Disallowed combinations and constraints on test conditions

Limited validation sets

Sequential testing approaches for SFD

Conclusions and Future Work

Summary

Test designs should support the capturing of M&S behavior and the building of a statistical emulator (meta-model)

- Otherwise we risk making inaccurate predictions about system performance and drawing inaccurate conclusions in reports

Space-Filling Designs and associated analysis techniques are often the most effective and efficient way to characterize M&S output

- Supports predictions across tested and untested conditions
- Ability to quantify uncertainty
- Potential to save time and money by not having to re-run the M&S itself

However, SFD and GPs are currently underused, and the T&E paradigm presents unique challenges to implementation

Future Work

IDA is actively researching when and how to implement SFD within the M&S VV&A process

- Case Studies
- Trainings
- Simulation Studies
- R packages

References

- Ba, Shan, William R. Myers, and William A. Brenneman. 2015. "Optimal Sliced Latin Hypercube Designs." *Technometrics* 57 (4): 479–87. https://doi.org/10.1080/00401706.2014.957867.
- Damblin, G., M. Couplet, and B. Iooss. 2013. "Numerical Studies of Space-Filling Designs: Optimization of Latin Hypercube Samples and Subprojection Properties." *Journal of Simulation* 7 (4): 276–89. https://doi.org/10.1057/jos.2013.16.
- Fang, Kai-Tai, Dennis K. J. Lin, Peter Winker, and Yong Zhang. 2000. "Uniform Design: Theory and Application." *Technometrics* 42 (3): 237–48.
- Gramacy, Robert B. 2020. *Surrogates: Gaussian Process Modeling, Design, and Optimization for the Applied Sciences*. CRC Press.
- Joseph, V. Roshan, Evren Gul, and Shan Ba. 2015. "Maximum Projection Designs for Computer Experiments." *Biometrika* 102 (2): 371–80. https://doi.org/10.1093/biomet/asv002.
- Lekivetz, Ryan, and Bradley Jones. 2019. "Fast Flexible Space-Filling Designs with Nominal Factors for Nonrectangular Regions." *Quality and Reliability Engineering International* 35 (2): 677–84. https://doi.org/10.1002/qre.2429.
- Loeppky, Jason L., Jerome Sacks, and William J. Welch. 2009. "Choosing the Sample Size of a Computer Experiment: A Practical Guide." *Technometrics* 51 (4): 366–76.
- Montgomery, Douglas C. 2017. *Design and Analysis of Experiments*. John Wiley & Sons.
- National Research Council. 1998. *Statistics, Testing, and Defense Acquisition: New Approaches and Methodological Improvements*. National Academies Press.
- Santner, Thomas J., Brian J. Williams, and William I. Notz. 2018. *The Design and Analysis of Computer Experiments*. 1st ed. New York City, New York, USA: Springer. https://doi.org/10.1007/978-1-4939-8847-1.
- Wojton, Heather, Kelly Avery, Laura Freeman, Samuel Parry, Gregory Whittier, Thomas Johnson, and Andrew Flack. 2019. "Handbook on Statistical Design & Analysis Techniques for Modeling & Simulation Validation." Available at https://testscience.org/research-on-emerging-directions/

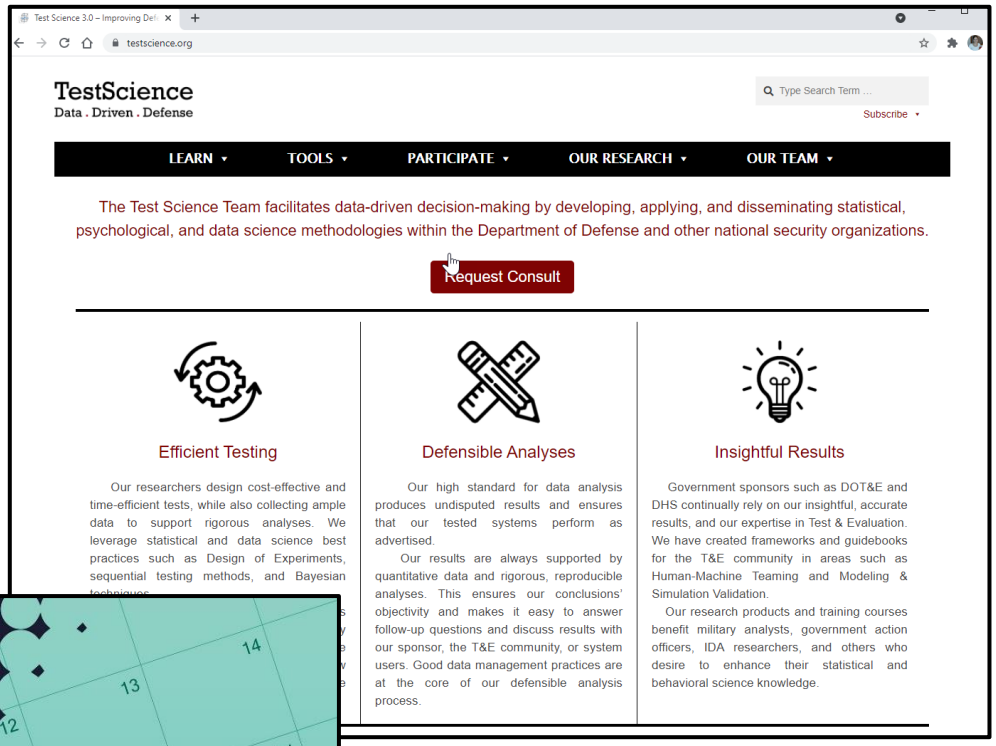
Contact Info and Resources

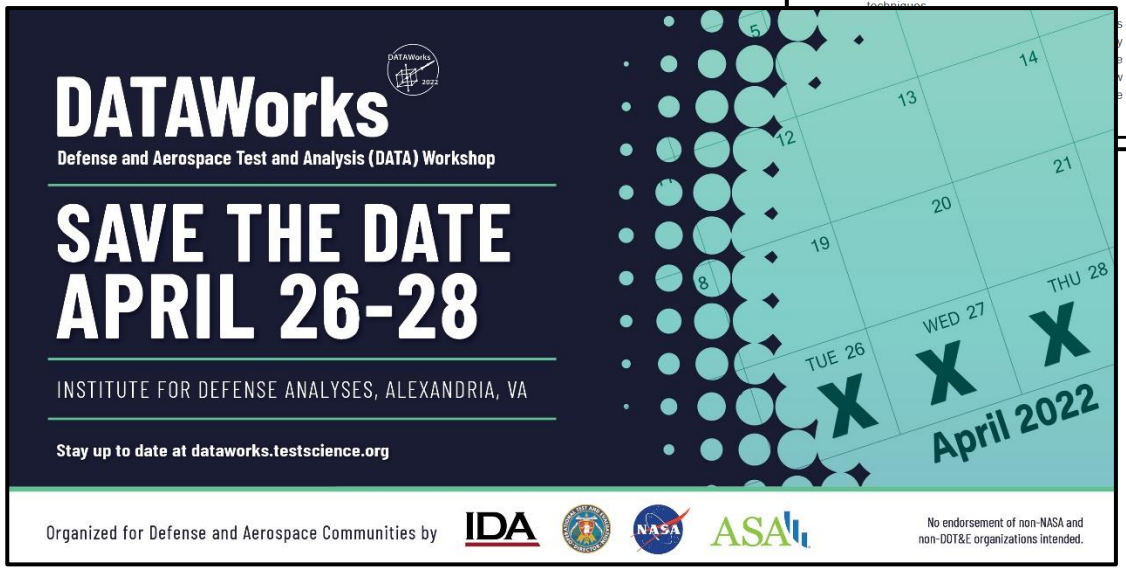
Kelly Avery

kavery@ida.org

Curtis Miller

cmiller@ida.org





TestScience

Data . Driven . Defense

BACKUP

	Space-Filling Designs (SFD)	Classical DOE
Purpose	Leverages **unique properties** of computer models (low noise and complex factor space)	Assumes live testing conditions (large amount of noise and few controllable factors)
Approach	Fills the **entire M&S factor space** efficiently; no replication	Biased toward the edge of the factor space; often includes replication
Analysis	Performs best with statistical emulators (e.g., **Gaussian Process** models)	Geared toward using simple linear models

Useful for understanding M&S behavior (Goal 2 of validation) when M&S is not highly stochastic

Useful for planning live test events and supporting comparisons between live and M&S data (Goal 1 of Validation)

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
07-2021	IDA Publication	

4. TITLE ANDSUBTITLE

Test Design Challenges in Defense Testing

5a. CONTRACT NUMBER
Separate Contract

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Kelly M. Avery (OED); Curtis G. Miller (OED);

5d. PROJECT NUMBER

5e. TASK NUMBER
C9082

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

8. PERFORMING ORGANIZATION REPORT NUMBER
NS-D-22723
H 2021-000241

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

10. SPONSOR/MONITOR'S ACRONYM(S)
IDA

11. SPONSOR/MONITOR'S REPORT NUMBER

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release. Distribution Unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Before the DoD acquires any major new capability, that system must undergo realistic testing in its intended environment with military users. Evaluations of these systems increasingly rely on computer models and simulations (M&S) to supplement live testing in cases where testing against realistic threats is impossible, unsafe, or prohibitively expensive. It is thus crucial to thoroughly validate these M&S tools using rigorous data collection and analysis strategies to ensure their output adequately represents reality.

While classical experimental design techniques have been widely adopted across the defense community for planning live tests, gold standard computer experiment techniques from the academic literature (e.g. space filling designs and Gaussian process emulators) are much less commonly used for M&S studies. Defense testing poses unique demands, such as a heavy reliance on categorical factors and binary outcomes, the mandate to judge the adequacy of sample size, extreme constraints in test conditions, and non-deterministic M&S outputs. There is currently no consensus on how to incorporate these demands into the existing academic framework for M&S. This brief will first provide an overview of operational testing and discuss example defense applications of, and key differences between, classical and space-filling designs. It will then present several challenges (and possible solutions) associated with implementing space-filling designs and associated analyses in the defense community.

15. SUBJECT TERMS
Design of Experiments (DOE); Modeling and Simulation (M&S); Operational Test (OT); Space Filling Designs; Statistics

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Rebecca Medlin (OED)
Unclassified	Unclassified	Unclassified	Unlimited	44	**19b. TELEPHONE NUMBER** *(include area code)* (703) 845-6731