INSTITUTE FOR DEFENSE ANALYSES

MDO Workshop: Test Design Challenges in Defense Testing

John T Haman, Project Leader

Kelly Avery
John T Haman
Thomas Johnson
Curtis Miller
Dhruv Patel
Han Yi (JHU-APL)

OED Draft

July 2024

IDA Product ID 3002855

INSTITUTE FOR DEFENSE ANALYSES
730 East Glebe Road
Alexandria, Virginia 22305

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 BD-9-229990, "Methods Develop," for the Office of the Director, Operational Test and Evaluation. 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 Dr. Heather M. Wojton and consisted of Dr. Breeana Anderson and Dr. Kelly Avery from the Operational Evaluation Division.

For more information:
Dr. John T Haman, Project Leader
jhaman@ida.org • 703-845-2132

Dr. Heather M. Wojton, Director, Operational Evaluation Division
hwojton@ida.org • (703) 845-6811

Rigorous Analysis | Trusted Expertise | Service to the Nation

INSTITUTE FOR DEFENSE ANALYSES

IDA Product ID 3002855

MDO Workshop: Test Design Challenges in Defense Testing

John T Haman, Project Leader

Kelly Avery
John T Haman
Thomas Johnson
Curtis Miller
Dhruv Patel
Han Yi (JHU-APL)

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 environments 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 briefing 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.

This material was presented at the Multi-Domain Operations Workshop in Alexandria, VA on 17 July 2024.

Test Design Challenges in Defense Testing

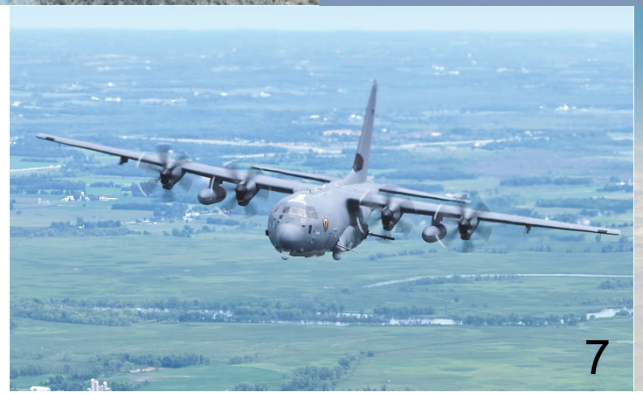
Drs. Kelly Avery, Curtis Miller, Thomas Johnson, Dhruv Patel, Han Yi (JHU-APL)
Presenter: Dr. John Haman

July 17, 2024

Institute for Defense Analyses
730 East Glebe Road ● Alexandria, Virginia 22305

Testing in DoD

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



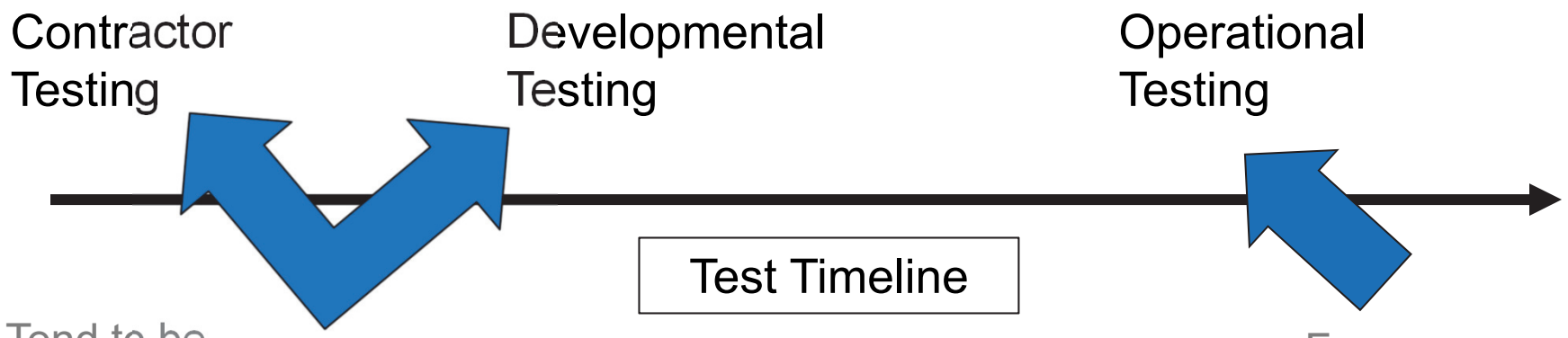
See Image Attributions slide for source information.

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

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

See Image Attributions slide
for source information.

DoD test paradigm



Tend to be requirements driven

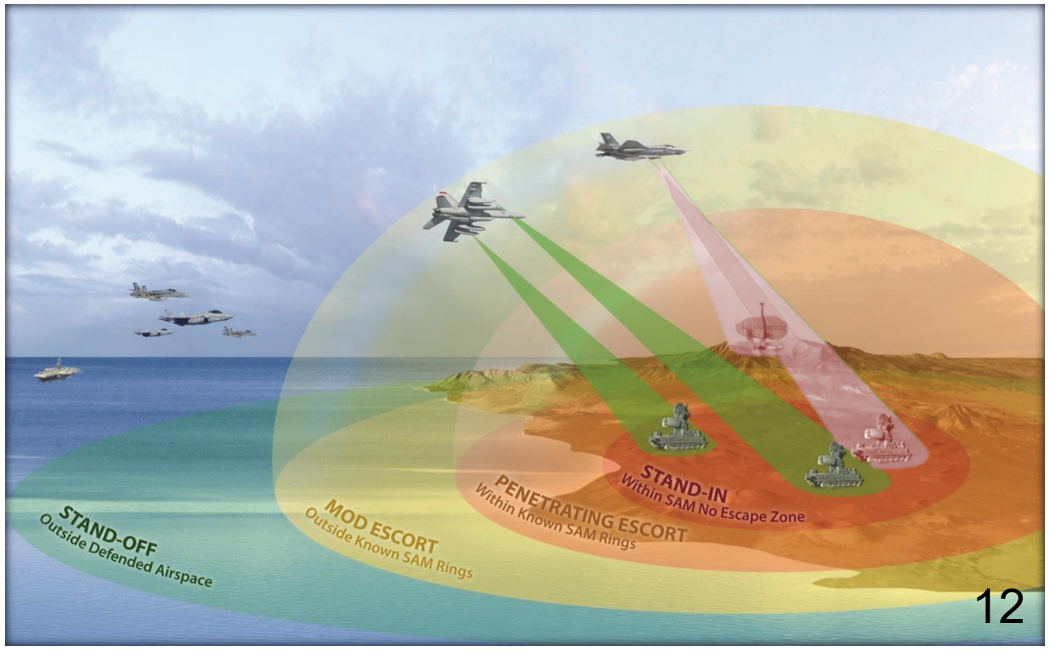
Focuses on mission success

Requirements documents are often missing important mission considerations

See Image Attributions slide for source information.

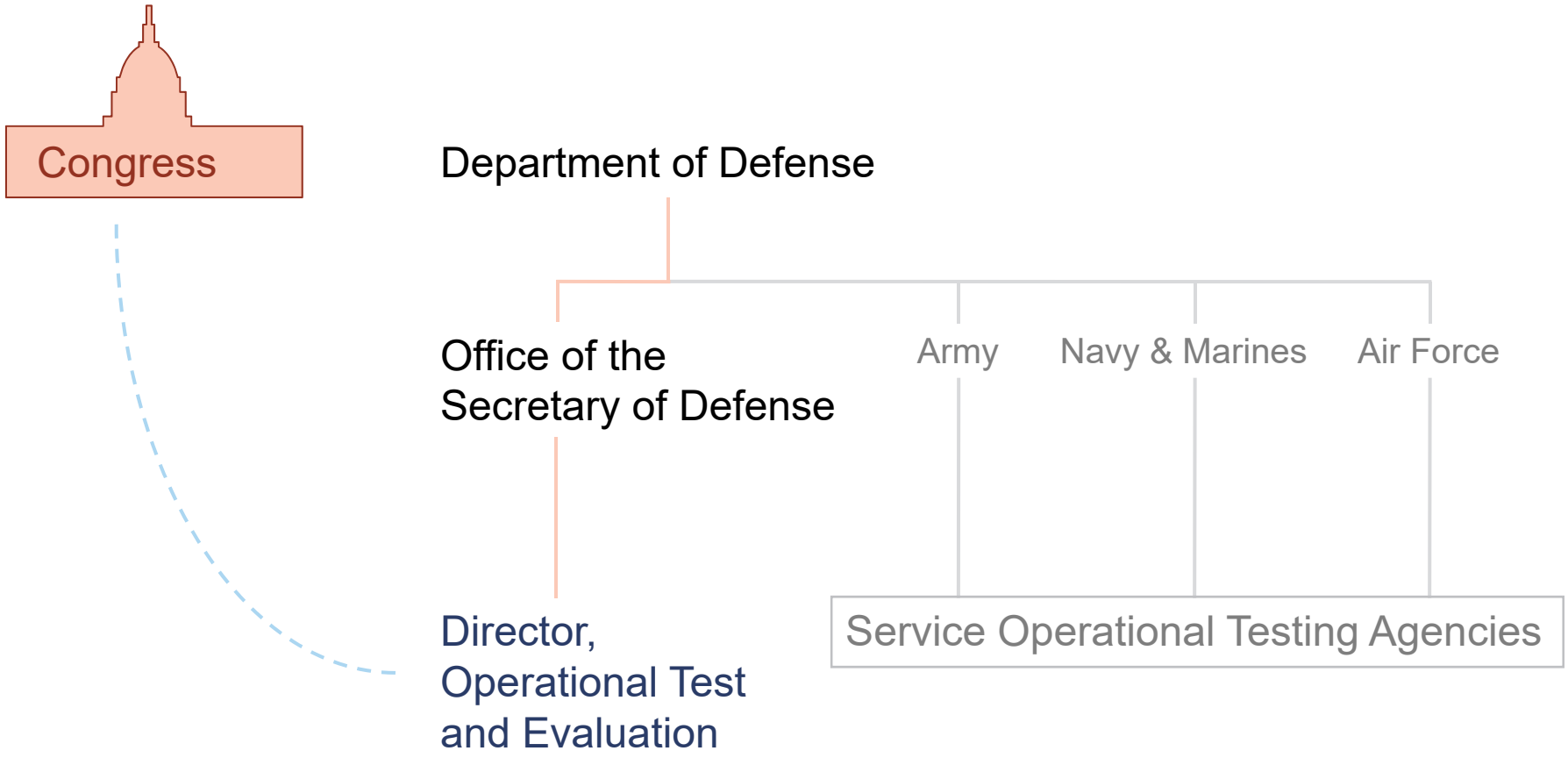
Goal of operational test: evaluate operational effectiveness, suitability, and survivability/lethality

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

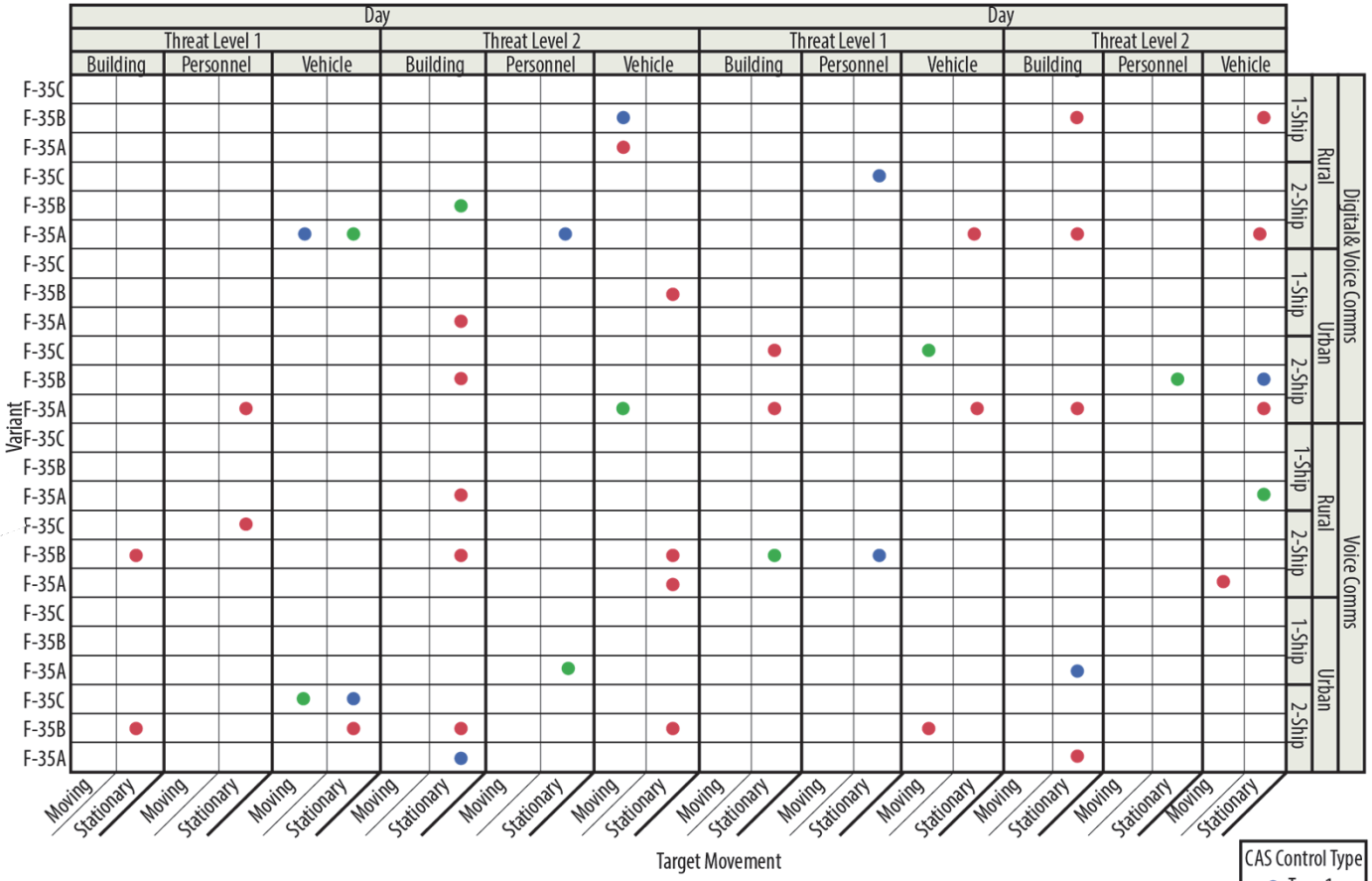


See Image Attributions slide
for source information.

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



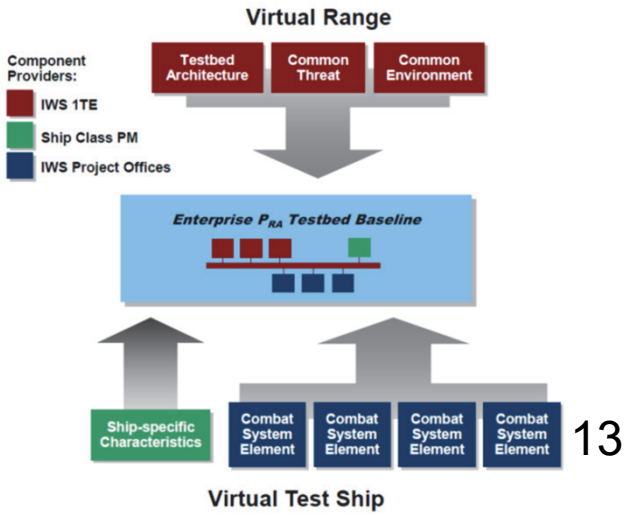
Classical Design of Experiments techniques are commonly used to efficiently collect data from live test events



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

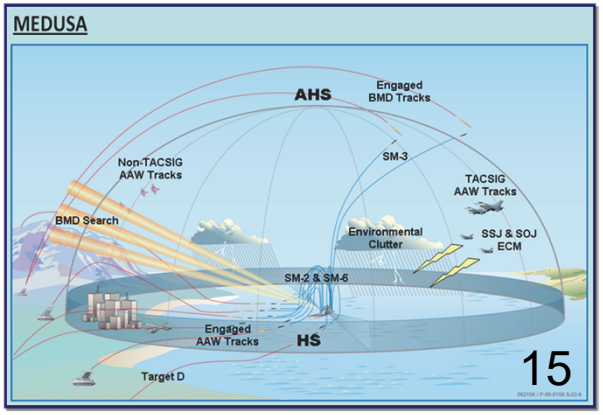
Modeling & Simulation (M&S) Validation

Operational testers commonly encounter four types of simulations



Computer/Software in the Loop:
e.g., Probability of Raid Annihilation Test Bed

Hardware in the Loop:
e.g., Environment Centric Weapons Analysis Facility



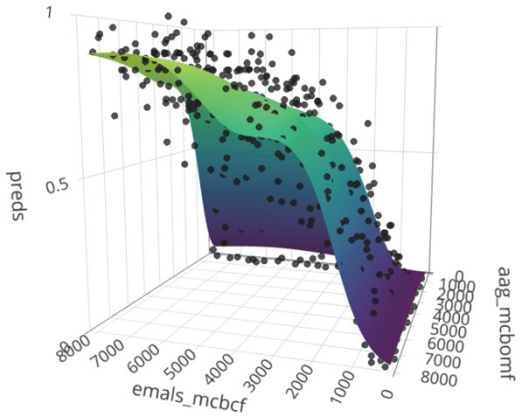
Digital Simulations:
e.g., Multi-Target Effectiveness Determined under Simulation by Aegis

Operator in the Loop:
e.g., F-35 Joint Simulation Environment

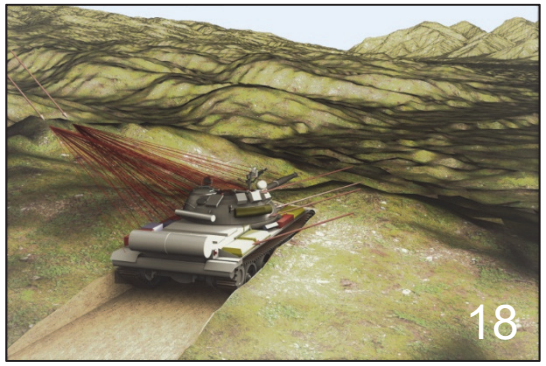
See Image Attributions slide for source information.

Operational testers can use M&S for all types of evaluations (in all warfare areas)

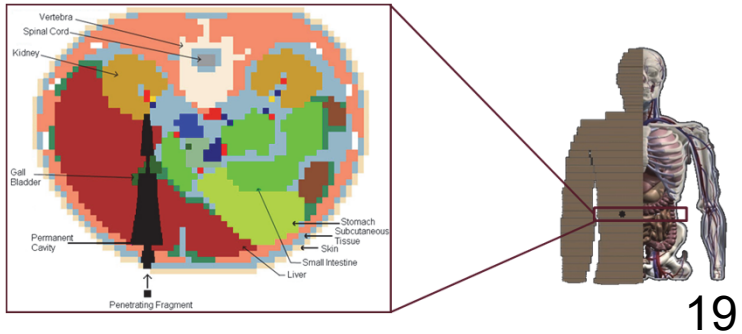
Effectiveness
e.g., Integrated Flight Simulation (for JAGM, SDB II, etc)



Suitability
e.g., CVN 78 reliability model



Survivability
e.g., AJEM



Lethality
e.g., ORCA

See Image Attributions slide for source information.

Some models can be used for multiple types of evaluation.

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

- 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).

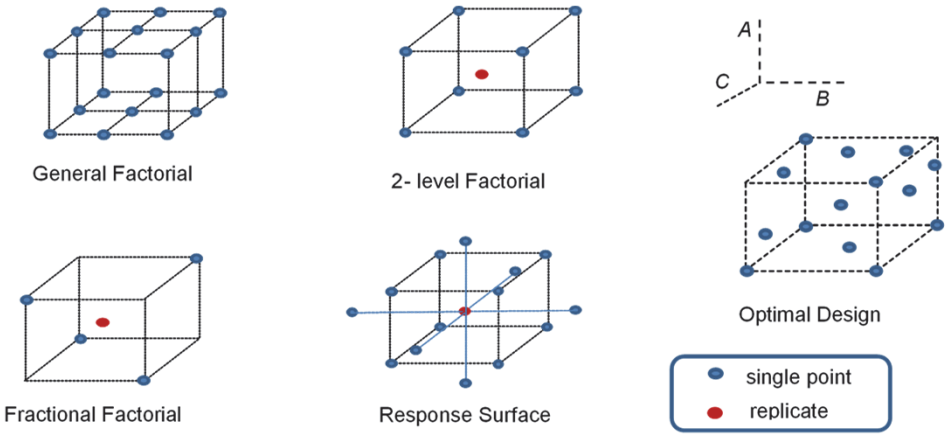
- The VV&A process should provide a quantitative understanding of how accurate an M&S capability is and identify limitations of the M&S across the factor space of interest.

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

Test designs that facilitate this goal:

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

Classical DOE still works!



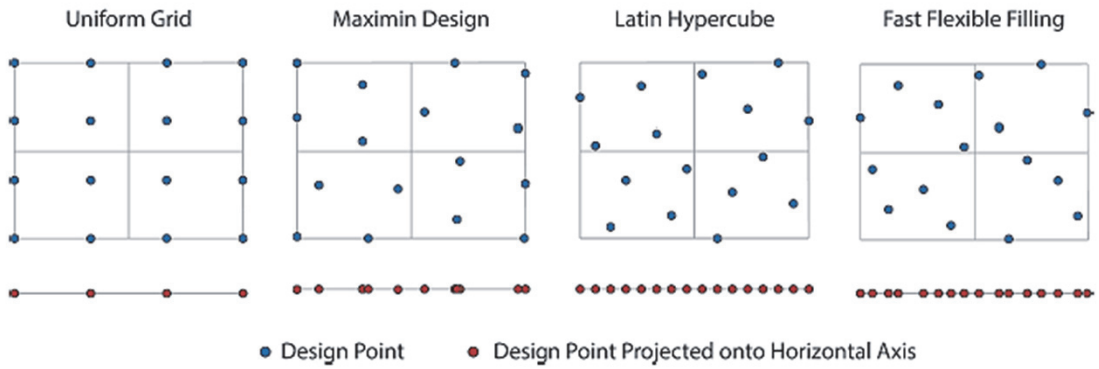
These design and analysis techniques are already widely understood and used

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

Test designs that facilitate this goal:

- Fill the M&S space
- Support building a statistical emulator (meta-model)
- Consider all input parameters (even those not able to be varied in live test)

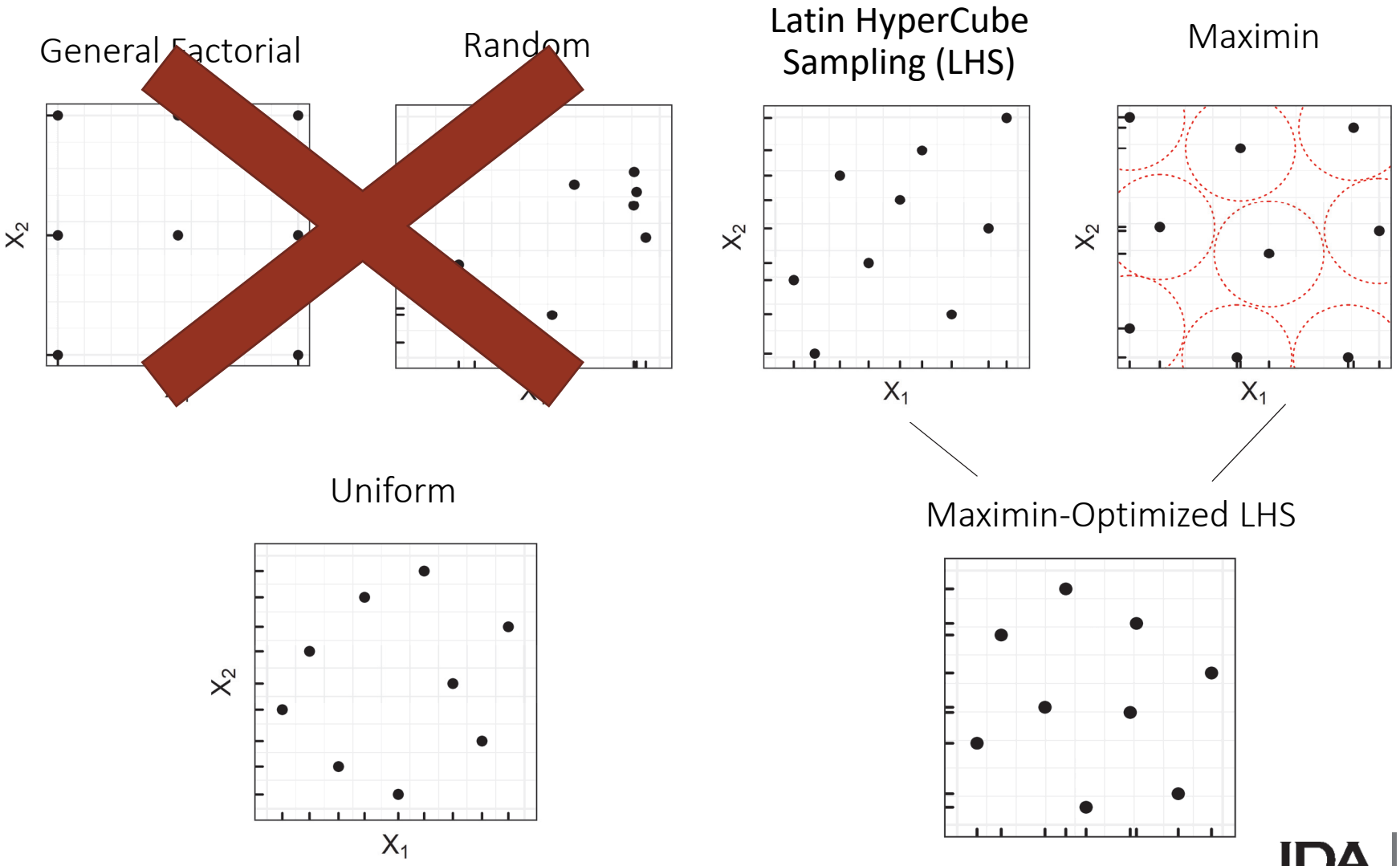
Space-Filling Designs are recommended



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

Space-Filling Designs (SFDs)

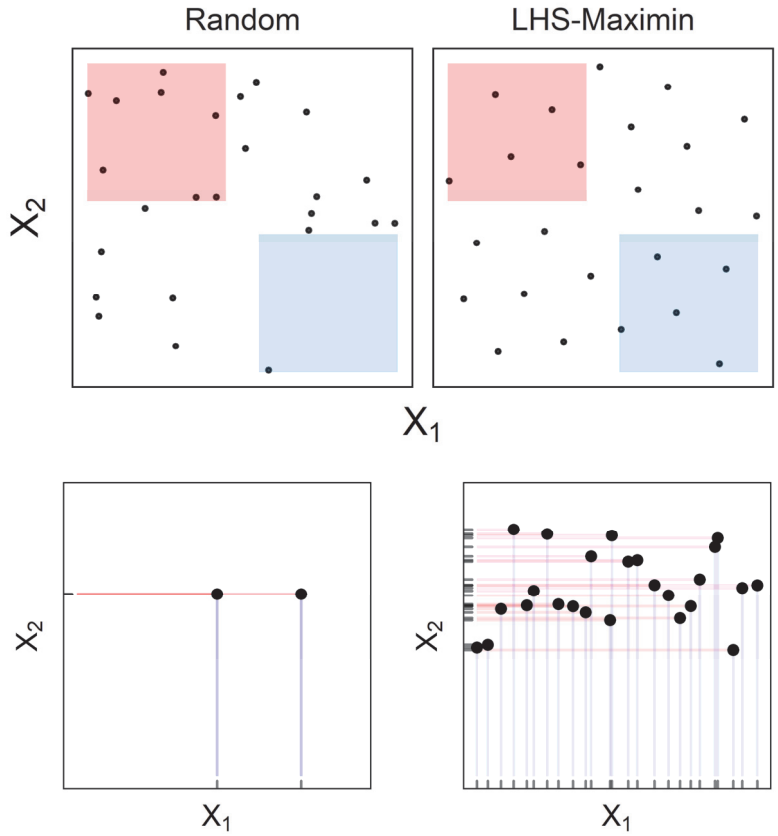
There are several common types of SFD in the literature, some of which can be useful for M&S validation



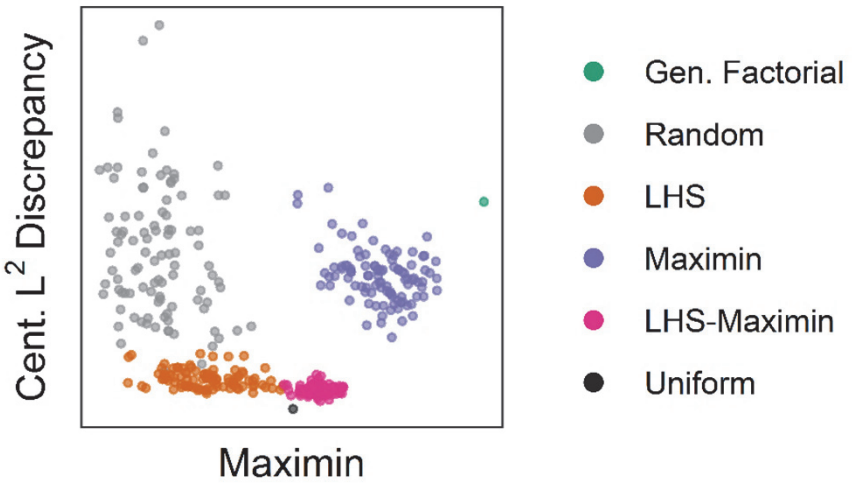
Just like with classical DOE, there are quantitative ways to evaluate a specific design

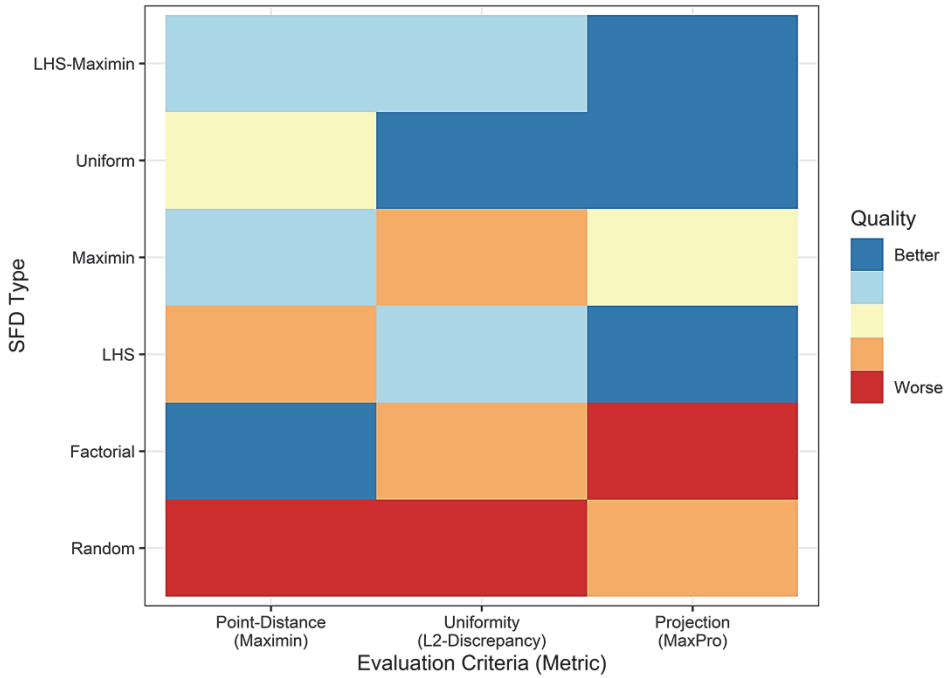
Many criteria exist, but it is particularly important that an SFD satisfy the following three criteria in order to be useful:

- 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]



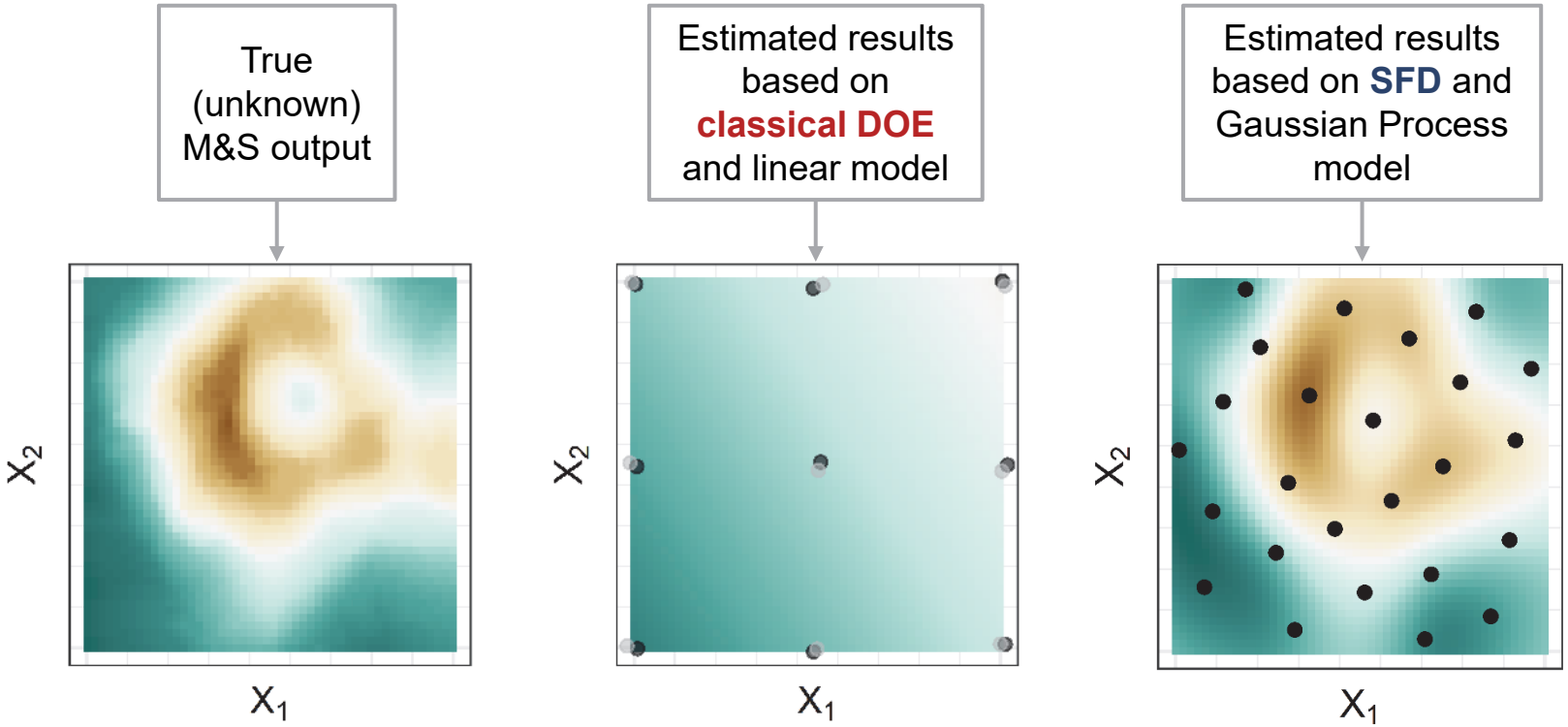
These criteria can be used to compare classes of designs and make general recommendations





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 and Opportunities for Space-Filling Designs in T&E

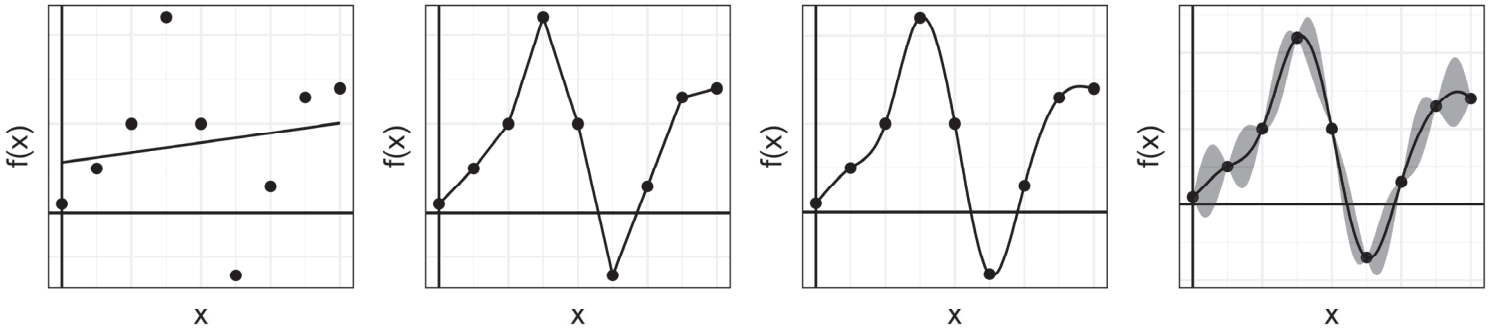
Challenges and Opportunities for Space-Filling Designs in T&E

1. Data Analysis Methods
2. Federated Models
3. Sample Size Determination
4. Categorical Variables

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



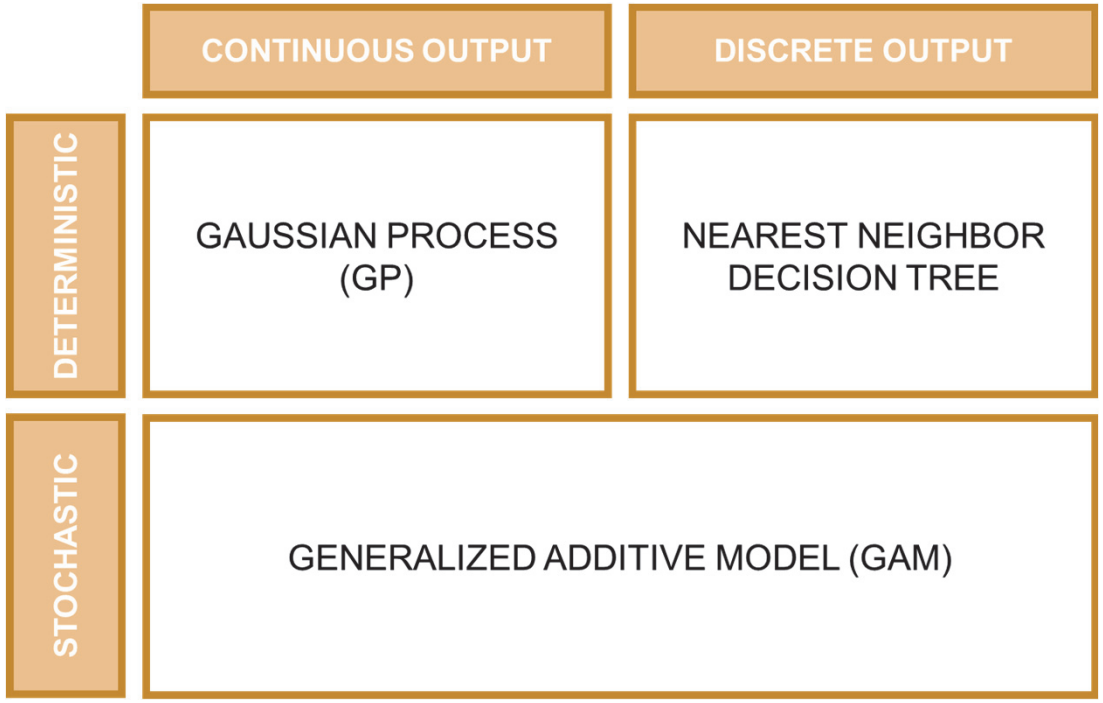
Linear Regression

Basic Interpolators and Splines

Gaussian Process Regression

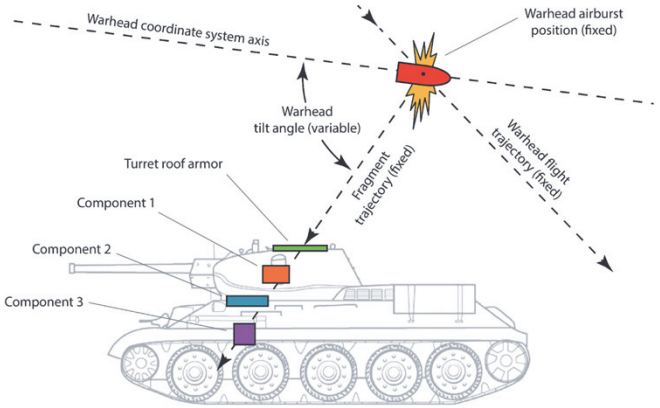
We recommend different analysis procedures based on the M&S output

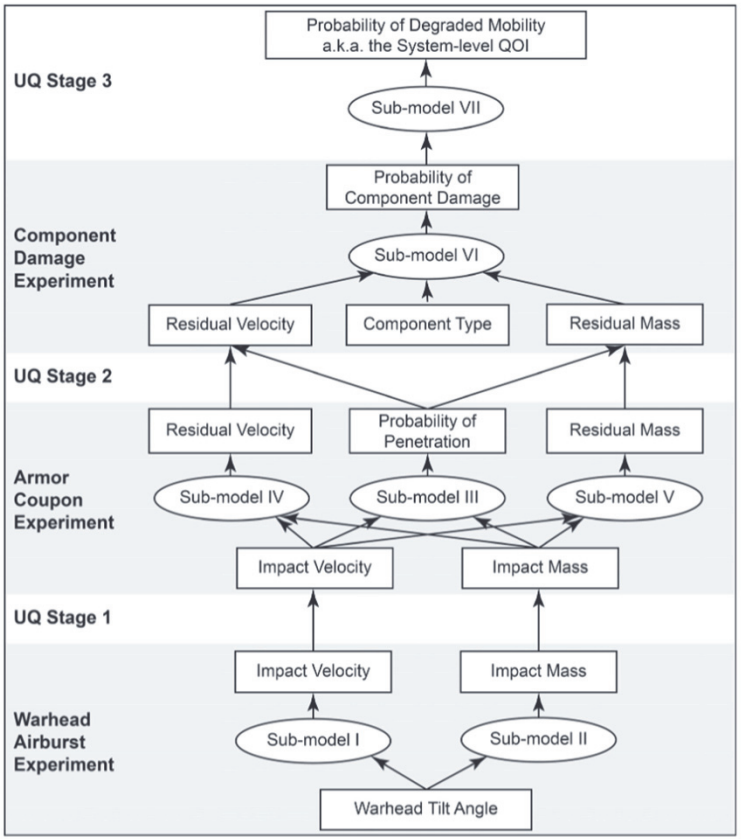
- *Deterministic* models produce the same exact results for a particular set of inputs
- *Stochastic* models possess some level of inherent randomness; the same inputs do not necessarily produce the same result



Source: *Metamodeling Techniques for Verification and Validation of Modeling and Simulation Data*, C.Miller and J. Haman, testscience.org

Which methods and designs are best for federated models?





Warhead Airburst Experiment

1. Fragment velocity model
2. Fragment mass model

Armor Coupon Experiment

3. Probability of penetration model
4. Residual velocity model
5. Residual mass model

Component Damage Experiment

6. Probability of component damage model

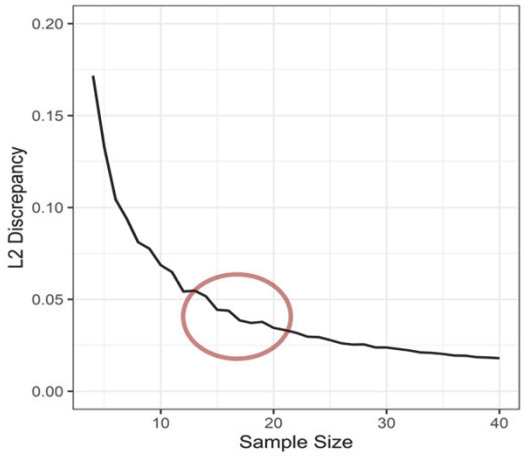
Look for *Uncertainty Quantification for Ground Vehicle Vulnerability Simulation* to be published in **Quality Engineering**

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

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:



ECWAF Simulates Torpedo Performance

We are interested in running a simulation in the ECWAF where we vary several factors:

- **Range** (continuous)
- **Targeting error** (continuous)
- **Environment** (categorical: open ocean or littoral)
- **Target type** (categorical: SSN, SSK)

In each run of the simulation, the outcome is either **hit** or **miss** (*binary*)

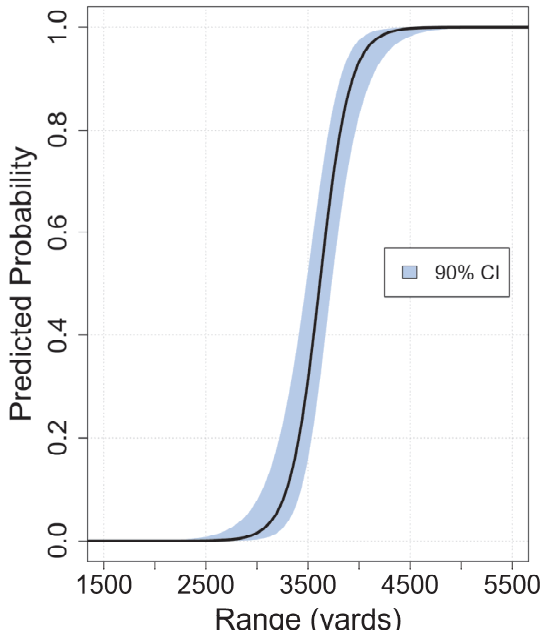
Our primary interest is the **probability of a hit** given the conditions

ECWAF – Environment Centric Weapons Analysis Facility; SSN – nuclear-powered submarine, SSK – diesel-electric submarine

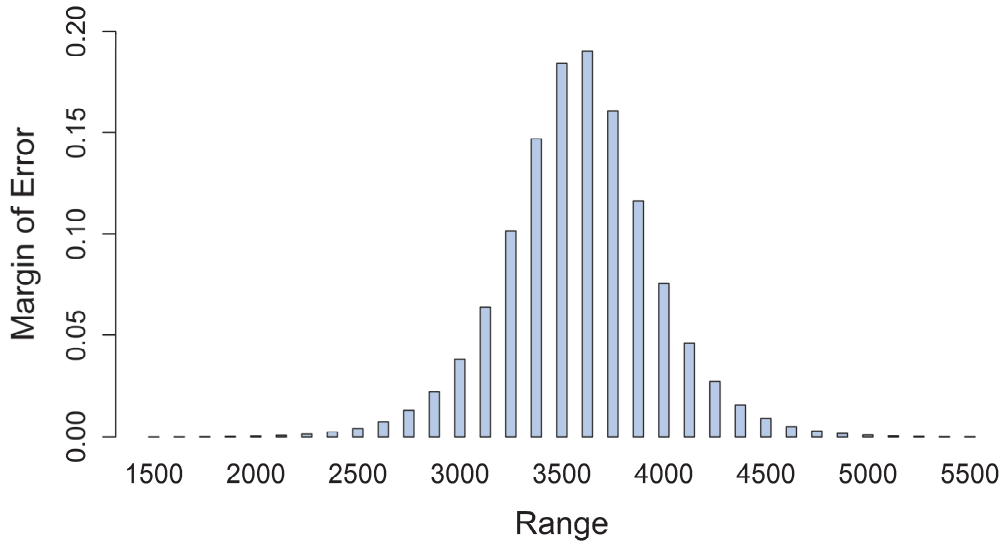
See Image Attributions slide for source information.

Tests can be sized for the a specific margin of error

Based on MOE and model discrimination, researchers can make more rigorous sample size determinations (for binary outcomes)



Estimating the margin of error **requires knowing the model**



Derived equation to calculate a **theoretical upper bound** to the margin of error at each point

Look for *Determining the Necessary Number of Runs in Computer Simulations with Binary Outcomes* at JSM in Portland this year!

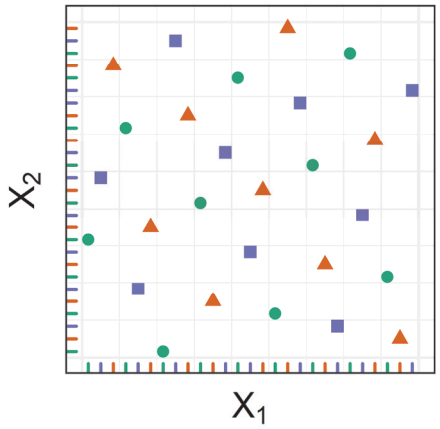
Operational and live fire tests often involve variables that are categorical or binary rather than continuous

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?

Conclusions and Future Work

Summary

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

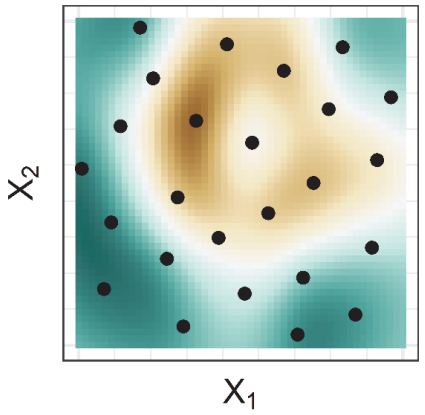
- Otherwise we risk making inaccurate predictions about system performance and drawing inaccurate conclusions in reports

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

IDA's test science group continues to conduct research, publish case studies, and provide trainings



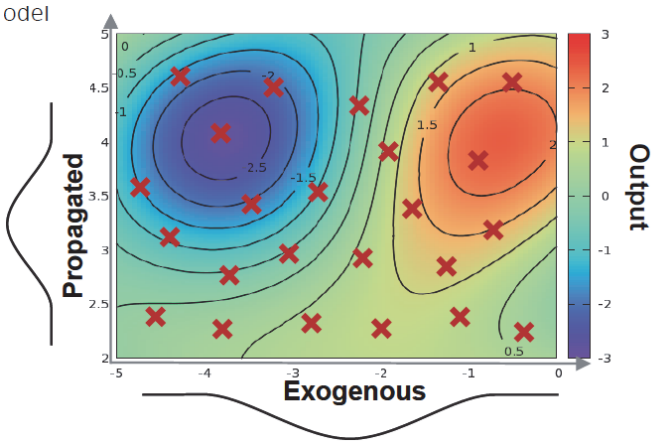
M&S Calibration

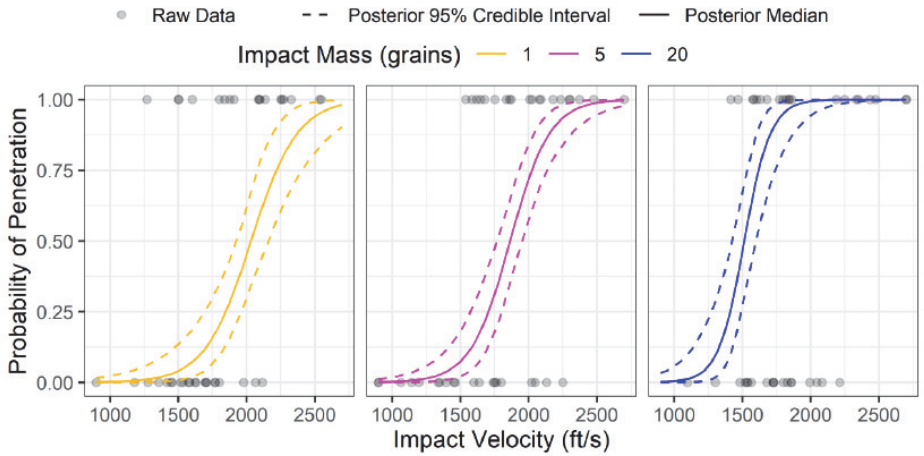
Uncertainty Propagation

Space-Filling Designs

Metamodeling

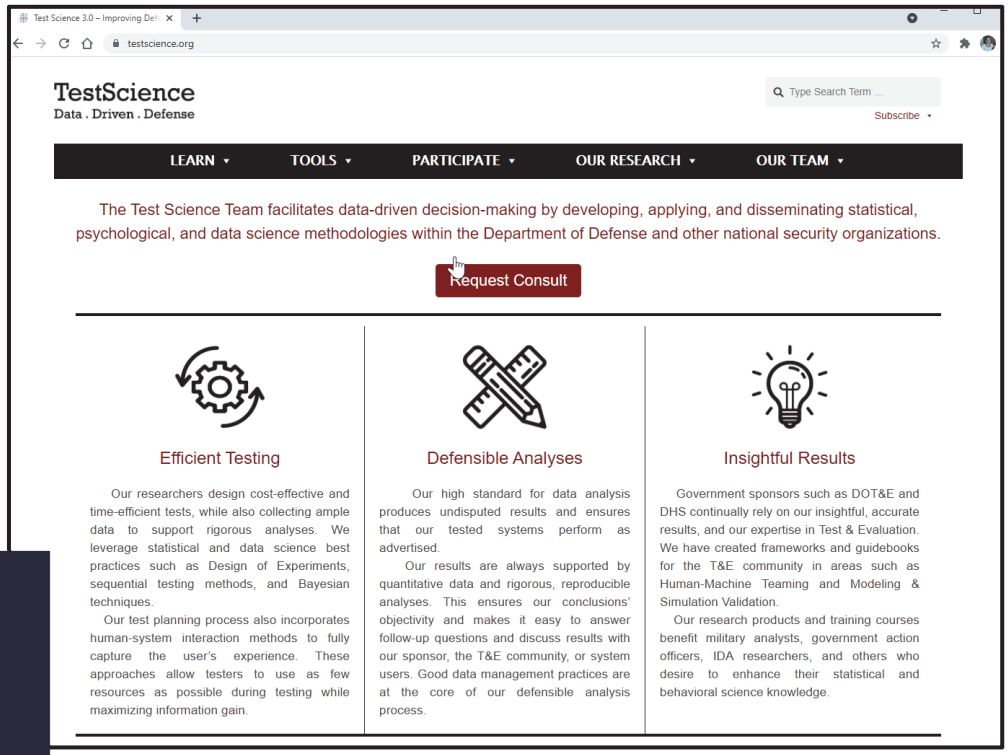
Federated Models V&V





Contact Info and Resources

John Haman
jhaman@ida.org



TestScience
Data . Driven . Defense

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/

Image Attributions

1. Director, Operational Test and Evaluation, FY 2016 Annual Report, p. 197. (https://www.dote.osd.mil/Portals/97/pub/reports/FY2016/navy/2016aav.pdf?ver=2019-08-22-105302-133)
2. *Virginia*-class submarine North Carolina (SSN-777). Photo # DCS07-901-395 by John Whalen, courtesy of northropgrumman.com. (https://www.navsource.org/archives/08/08777.htm)
3. Pedro Aragão, CC BY-SA 3.0 GFDL, via Wikimedia Commons. (https://commons.wikimedia.org/wiki/File:Boeing_E-3C_Sentry,_United_States_-_US_Air_Force_(USAF)_JP38894.jpg)
4. Director, Operational Test and Evaluation, FY 2013 Annual Report, p. 267. (https://www.dote.osd.mil/Portals/97/pub/reports/FY2013/af/2013cv-22.pdf?ver=2019-08-22-111345-720)
5. U.S. Air Force photo by Tech. Sgt. Paul Dean. (https://www.homestead.afrc.af.mil/News/Photos/igphoto/2000492195/mediaid/982765/)
6. U.S. Navy photo by Andy Wolfe. (https://en.wikipedia.org/wiki/Lockheed_Martin_F-35_Lightning_II#/media/File:CF-1_flight_test.jpg)
7. U.S. Air Force photo by Master Sgt. Christopher Boitz. (https://www.af.mil/About-Us/Fact-Sheets/Display/Article/467756/ac-130j-ghostrider/)
8. U.S. Air Force photo by Christopher Okula. (https://en.wikipedia.org/wiki/Boeing_KC-46_Pegasus#/media/File:KC-46_Pegasus_prepares_to_refuel_C-17_(cropped).jpg)
9. Director, Operational Test and Evaluation, FY 2022 Annual Report, p. 65. (https://www.dote.osd.mil/Portals/97/pub/reports/FY2022/dod/2022pki.pdf?ver=uxp7uqwOBmNz94TQr2W2_A%3D%3D)
10. U.S. Navy Photo by Photographer's Mate Airman Inez Lawson. (https://en.wikipedia.org/wiki/Command_and_control#/media/File:CIC-USS-CarlVinson-2001.jpg)
11. U.S. Air Force photo by Tech. Sgt. Vincent Mouzon. (https://www.dvidshub.net/image/1142890/operational-test-launch)
12. U.S. Navy Program Executive Office for Tactical Aircraft Programs - briefing on Next Generation Jammer. (https://fullafterburner.weebly.com/aerospace/boeing-fa-18-advanced-super-hornet-rebound-of-a-striker)
13. PEO Integrated Warfare Systems, "Modeling and Simulation for Enterprise Test and Evaluation," May 9, 2007, p. 6. (https://www.slideserve.com/bernad/navy-pra-testbed-development-for-operational-test-evaluation)
14. U.S. Navy, NUWC Division Newport UCTOC, "Code 85 Weapons Analysis Facility (WAF) Technical Engineering Services - Pre-Solicitation Conference," June 12, 2014, p. 13. (https://www.navsea.navy.mil/Portals/103/Documents/NUWC_Newport/ReadingRoom/Code85_WAF.pdf)
15. PEO Integrated Warfare Systems, "Integrated Combat Systems (IWS 1.0)," NDIA 2011 Integrated Warfare Systems Conference; p. 14. (https://ndiastorage.blob.core.usgovcloudapi.net/ndia/2011/PEO/Maniquis.pdf)
16. U.S. Air Force photo (no further attribution). (https://www.aftc.af.mil/News/Photos/igphoto/2003329775/)
17. Director, Operational Test and Evaluation, FY 2018 Annual Report, p. 83. (https://www.dote.osd.mil/Portals/97/pub/reports/FY2018/army/2018jagm.pdf?ver=2019-08-21-155807-057)
18. U.S. Army photo illustration by DEVCOM DAC. (https://www.army.mil/article/247323/dods_accredited_model_for_survivability_vulnerability_and_lethality_continues_to_roll_out_improvements)
19. Adapted from Stephen P. Swann, U.S. Army Research Laboratory, RDECOM, "Advancements in Personnel Incapacitation Methodologies for Multiple Cartridge Projectiles (MPCs)," May 19, 2010, p. 7. (https://ndia.dtic.mil/wp-content/uploads/2010/armament/WednesdayLandmarkBStephenSwann.pdf)

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
XX-07-2024	Final	

4. TITLE AND SUBTITLE

MDO Workshop: Test Design Challenges in Defense Testing

5a. CONTRACT NUMBER

HQ0034-19-D-0001

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

John Haman

5d. PROJECT NUMBER

BD-9-229990

5e. TASK NUMBER

2299(90)

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Institute for Defense Analyses
730 East Glebe Road
Alexandria, Virginia 22305

8. PERFORMING ORGANIZATION REPORT NUMBER

3002855

3002855

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Director, Operational Test and Evaluation
1700 Defense Pentagon
Washington, DC 20301

10. SPONSOR/MONITOR'S ACRONYM(S)

DOT&E

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Distribution Statement A. Approved for public release: distribution is unlimited.

13. SUPPLEMENTARY NOTES

Project Leader: John Haman

14. ABSTRACT

This slide deck describes operational testing, the role of experimental design in operational testing, and some specific types of designs -- space-filling designs (SFD) -- which are particularly applicable to the DOD's test planning activities in modeling and simulation. We then survey some of the practical problems that make the application of SFDs to defense testing tenuous, as well as some research projects IDA conducts to ameliorate the challenges.

15. SUBJECT TERMS

operational testing, design of experiments, DOE, space-filling designs, SFD, modeling and simulation

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			John Haman
Unclassified	Unclassified	Unclassified	SAR	39	19b. TELEPHONE NUMBER *(Include area code)* 703-845-2132