INSTITUTE FOR DEFENSE ANALYSES

Statistical Methods Development Work for M&S Validation

John Haman, Project Leader

Curtis G. Miller

May 2023

IDA Document NS D-33460

Log: H 2023-000116

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-9-2299(90), "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. V. Bram Lillard and consisted of Dr. Kelly M Avery, Dr. Matthew R. Rickert, and Dr, Chad A. Brisbois- from the Operational Evaluation Division.

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

Dr. V. Bram Lillard, Director, Operational Evaluation Division
vlillard@ida.org • (703) 845-2230

Rigorous Analysis | Trusted Expertise | Service to the Nation

INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS D-33460

Statistical Methods Development Work for M&S Validation

John T Haman, Project Leader

Curtis G. Miller

Executive Summary

Modeling and simulation (M&S) validation is a critical component of ensuring that M&S is accredited for its intended use. While there are many methods, one of the most rigorous forms of validation is a data-driven comparison of simulation outputs with live test data, and statistics is the science of how to best make and collect data for these comparisons. However, technical barriers inhibit the utilization of statistical best practices in DOD testing.

Starting in 2016, Director, Operational Test and Evaluation (DOT&E) published guidance[1][2] on M&S validation for operational and live fire testing. Briefly, DOT&E expects that acquisition programs apply experimental design methodologies, including formal statistical tests, and recommends space-filling designs as one such methodology. Additionally, DOT&E recommends that programs use metamodels to understand M&S outcomes and to quantify uncertainty. However, DOT&E stopped short of recommending specific methodologies in their guidance to the Services. To address the aforementioned technical barriers and implement DOT&E's M&S guidance, IDA's Test Science group has been working on trainings, tools, and methods that show the DOD test community how to apply statistically rigorous methods to validation.

IDA has published recommendations on planning for M&S validation,[3] designing computer experiments,[4] statistically analyzing the results, and comparing live testing data to M&S outcomes. This presentation summarizes those publications, provides links, and points toward interactive tools on TestScience.org that reduce technical barriers. We break the summary of our statistical

1 "Guidance on the Validation of Models and Simulation used in Operational Test and Live Fire Assessments." DOT&E Memorandum. March 14, 2016.

2 "Clarifications on Guidance on the Validation of Models and Simulation used in Operational Test and Live Fire Assessments." DOT&E Memorandum. January 17, 2017.

3 Wojton, Heather; Avery, Kelly; Freeman, Laura; Parry, Samuel; Whittier, Gregory; Johnson, Thomas; Flack, Andrew. Handbook on Statistical Design & Analysis Techniques for Modeling & Simulation Validation. IDA Document NS D-10455. February 2019.

4 Avery, Kelly; Miller, Curtis; Yi, Han. Space Filling Designs for Modeling & Simulation. IDA Document NS D-21562. February 2021.

recommendation for M&S validation into four sections, as follows:

First, test planners should include the relevant assessments of statistical risk in test and evaluation master plans (TEMPs) and test plans. For model validation, that entails power analyses. Lack of appropriate power may result in inappropriate trust of M&S predictions or waste test resources.

Second, when designing M&S experiments, testers should consider space-filling designs, which allow for the exploration of the operational space. Specifically, sliced maximin Latin hypersquare designs and MaxPro designs are reasonable default space-filling designs for making many model validation problems.

Third, space-filling designs should be coupled with appropriate statistical metamodels. This is an important method of conducting uncertainty quantification and summarizing the M&S in its own right. Depending on the situation, we recommend either generalized additive models, Gaussian process models, or decision trees.

Finally, we recommend that comparisons between live test data and simulation outputs be made using statistical procedures, and IDA has published on statistical validation procedures for a wide variety of test situations,[5] depending on the response distribution, type of data collection, and sample size. IDA has also published methods to address the case where the response is binary and differences in factor variation between live data and M&S outputs are of interest.[6]

[5] See Table 2 in Footnote 4.

[6] Metts, Carrington; Bartis, Elliot. DATAWorks 2023: Development of a Wald-Type Statistical Test to Compare Live Test Data and M&S Predictions. IDA Document NS D-33406. February 2023.

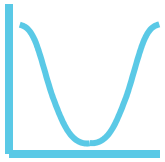
Statistical Methods Development Work for M&S Validation

Dr. Curtis Miller

May 4, 2023

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

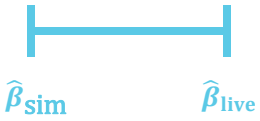
BLUF: Statistical methods help operational testing stakeholders decide if computer models should inform OT assessments



Test planning must adequately assess statistical risk

Space-filling designs for M&S assessment allow for discovery of unanticipated trends

Metamodels characterize trends through the entire operational space



M&S outputs need to be compared to live data with appropriate statistical tests

BLUF: Bottom line up front; DOE: Design of Experiments; OT: Operational testing

IDA provides analytical support to DOT&E

DOT&E: Director, Operational Test and Evaluation; IDA: Institute for Defense Analyses; OED: Operational Evaluation Division; OT: Operational Testing; USC: United States Code

Why does M&S VV&A matter to DOT&E?

Good OT&E reduces risk and uncertainty regarding the performance of systems in wartime

United States General Accounting Office

GAO

Report to the Honorable
William V. Roth and the Honorable
Charles E. Grassley, U.S. Senate

October 1997

TEST AND EVALUATION

Impact of DOD's Office of the Director of Operational Test and Evaluation

number of unknowns prior to the decision to begin full production, while program and service officials typically sought less testing and were willing to accept greater risk when making production decisions. The additional testing DOT&E advocated, often over the objections of service testers, served to meet the underlying objectives of operational testing—to reduce the uncertainty and risk that systems entering full-rate production would not fulfill their requirements.

GAO/NSIAD-98-22

https://www.gao.gov/assets/nsiad-98-22.pdf

OT&E: Operational Test and Evaluation

DOT&E assessments must be based on actual tests, not modeling and simulation alone

USC Title X § 4171

(a) (1) [SECDEF] shall provide that a [program] may not proceed beyond [LRIP] until initial [OT&E] of the program… is completed. …

(b) (1) Operational testing of a major defense acquisition program may not be conducted until [DOT&E] … approves (in writing) the adequacy of the plans …

(h) In this section, the term "operational test and evaluation" …does not include an operational assessment based exclusively on-
(1) computer modeling;
(2) simulation; or
(3) an analysis of system requirements, engineering proposals, design specifications, or any other information contained in program documents.

DOT&E: Director, Operational Test and Evaluation; LRIP: Low Rate Initial Production; OT&E: Operational Test and Evaluation; USC: United States Code; SECDEF: Secretary of Defense

DOT&E assessments must be based on actual tests, not modeling and simulation alone

M&S OUTPUTS ARE A LENS THROUGH WHICH WE INTERPRET LIVE TEST RESULTS

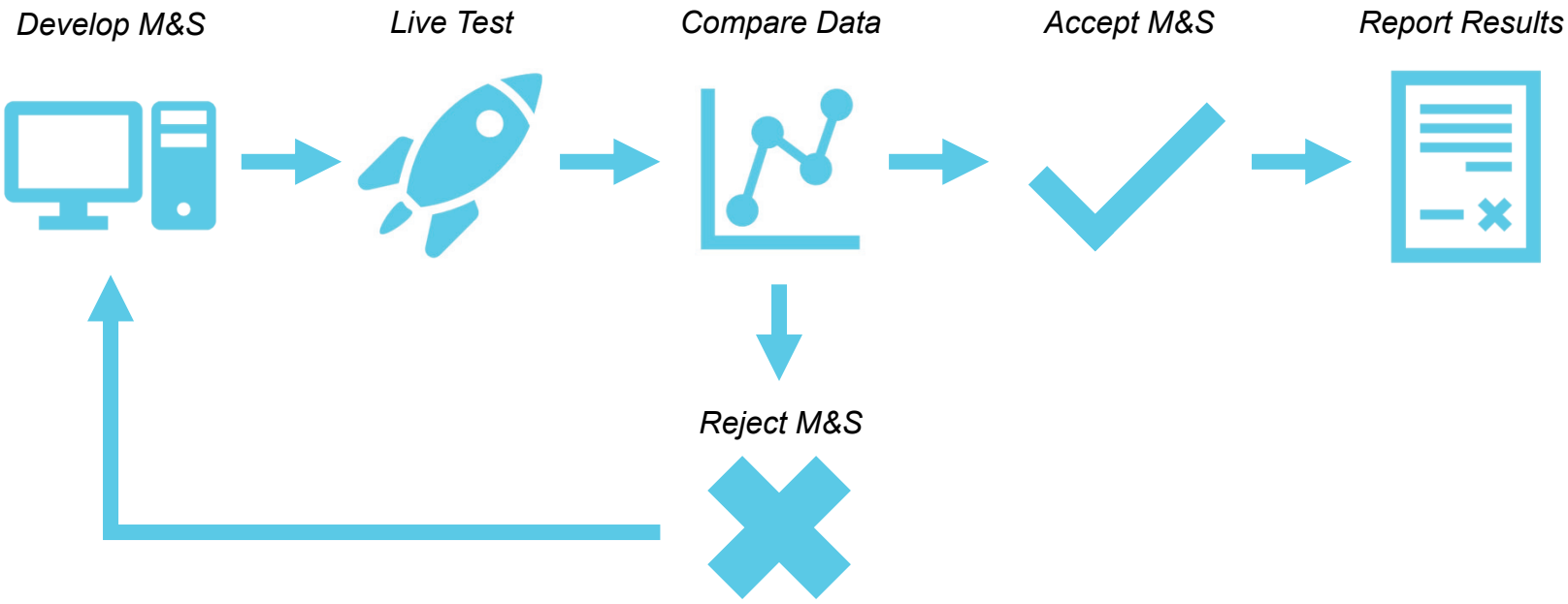
…de that a [program] may not …initial [OT&E] of the program…

…a major defense acquisition …d until [DOT&E] … approves (in …plans …

(h) In this section, the term "operational test and evaluation" …does not include an operational assessment based exclusively on-
(1) computer modeling;
(2) simulation; or
(3) an analysis of system requirements, engineering proposals, design specifications, or any other information contained in program documents.

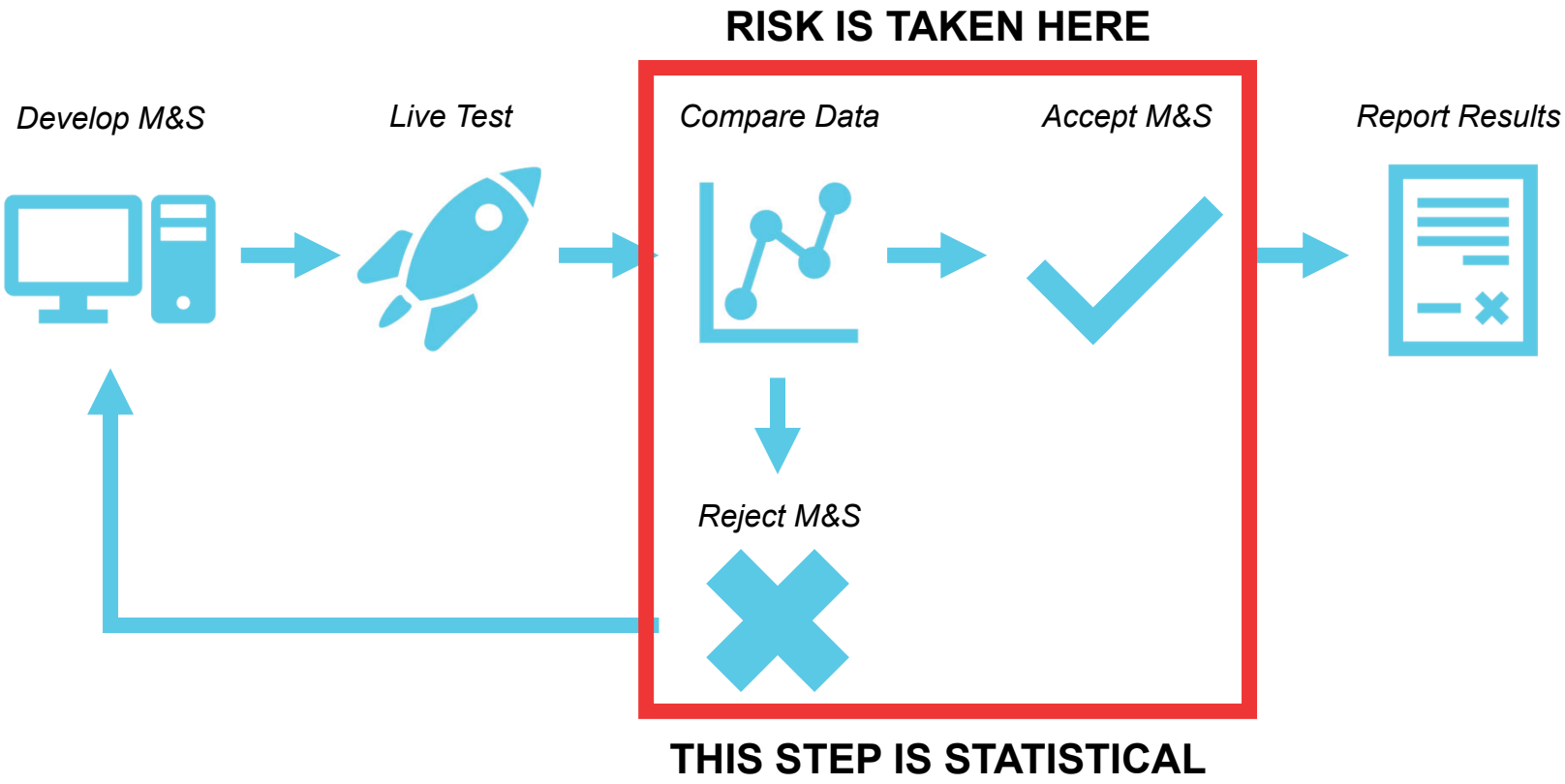
DOT&E: Director, Operational Test and Evaluation; LRIP: Low Rate Initial Production; M&S: Modeling and Simulation; OT&E: Operational Test and Evaluation; USC: United States Code; SECDEF: Secretary of Defense

If M&S outputs will be used for predicting OT results, we must compare M&S outputs to live test data



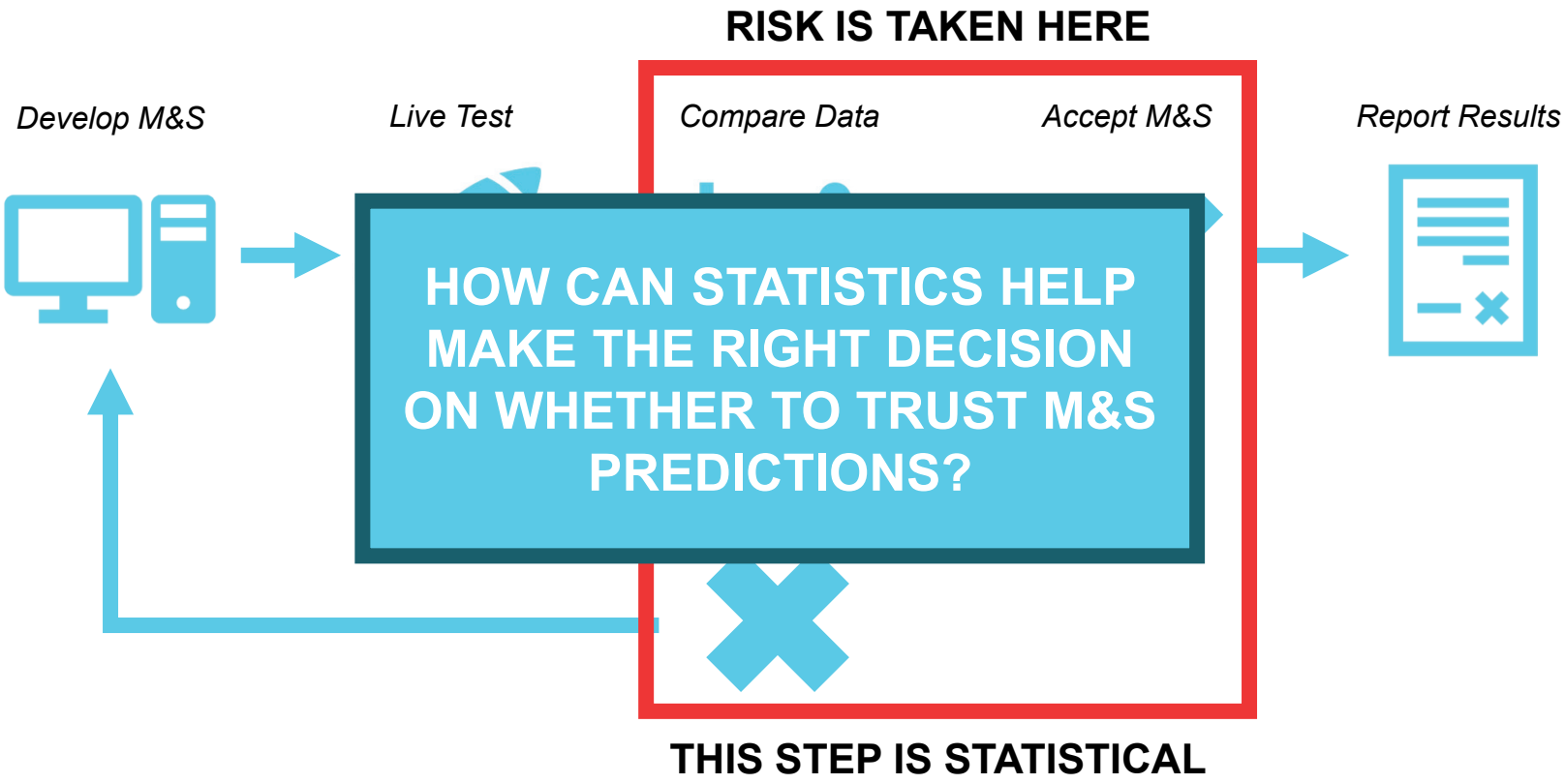
M&S: Modeling and Simulation; OT: Operational Testing

If M&S outputs will be used for predicting OT results, we must compare M&S outputs to live test data



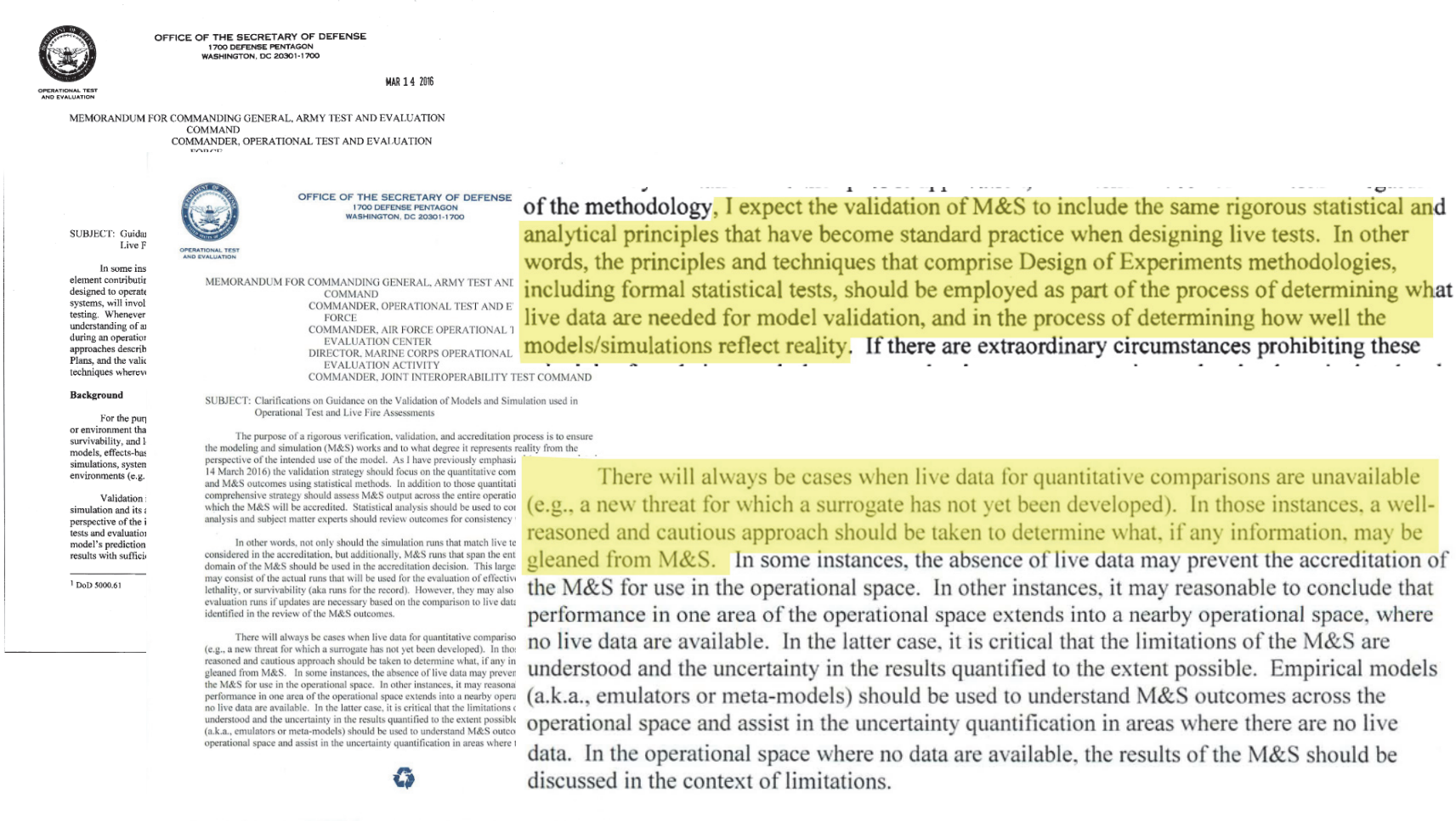
M&S: Modeling and Simulation; OT: Operational Testing

If M&S outputs will be used for predicting OT results, we must compare M&S outputs to live test data



M&S: Modeling and Simulation; OT: Operational Test; VV&A: Verification, Validation, and Accreditation

DOT&E's memos reveal their thoughts on the risk and set standards



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

MAR 14 2016

MEMORANDUM FOR COMMANDING GENERAL, ARMY TEST AND EVALUATION COMMAND
COMMANDER, OPERATIONAL TEST AND EVALUATION

SUBJECT: Guidu
Live F

In some ins
element contributir
designed to operate
systems, will invol
testing. Whenever
understanding of a
during an operatior
approaches describ
Plans, and the valic
techniques wherev

Background

For the pur
or environment tha
survivability, and l
models, effects-bas
simulations, systen
environments (e.g.

Validation
simulation and its
perspective of the i
tests and evaluatio
model's prediction
results with sufficie

[1] DoD 5000.61

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

MEMORANDUM FOR COMMANDING GENERAL, ARMY TEST ANI
COMMAND
COMMANDER, OPERATIONAL TEST AND E
FORCE
COMMANDER, AIR FORCE OPERATIONAL T
EVALUATION CENTER
DIRECTOR, MARINE CORPS OPERATIONAL
EVALUATION ACTIVITY
COMMANDER, JOINT INTEROPERABILITY TEST COMMAND

SUBJECT: Clarifications on Guidance on the Validation of Models and Simulation used in Operational Test and Live Fire Assessments

The purpose of a rigorous verification, validation, and accreditation process is to ensure the modeling and simulation (M&S) works and to what degree it represents reality from the perspective of the intended use of the model. As I have previously emphasi
14 March 2016) the validation strategy should focus on the quantitative com
and M&S outcomes using statistical methods. In addition to those quantitati
comprehensive strategy should assess M&S output across the entire operatio
which the M&S will be accredited. Statistical analysis should be used to co
analysis and subject matter experts should review outcomes for consistency

In other words, not only should the simulation runs that match live te
considered in the accreditation, but additionally, M&S runs that span the ent
domain of the M&S should be used in the accreditation decision. This large
may consist of the actual runs that will be used for the evaluation of effectiv
lethality, or survivability (aka runs for the record). However, they may also
evaluation runs if updates are necessary based on the comparison to live data
identified in the review of the M&S outcomes.

There will always be cases when live data for quantitative compariso
(e.g., a new threat for which a surrogate has not yet been developed). In tho
reasoned and cautious approach should be taken to determine what, if any in
gleaned from M&S. In some instances, the absence of live data may prever
the M&S for use in the operational space. In other instances, it may reasona
performance in one area of the operational space extends into a nearby opera
no live data are available. In the latter case, it is critical that the limitations
understood and the uncertainty in the results quantified to the extent possibl
(a.k.a., emulators or meta-models) should be used to understand M&S outco
operational space and assist in the uncertainty quantification in areas where

of the methodology, I expect the validation of M&S to include the same rigorous statistical and analytical principles that have become standard practice when designing live tests. In other words, the principles and techniques that comprise Design of Experiments methodologies, including formal statistical tests, should be employed as part of the process of determining what live data are needed for model validation, and in the process of determining how well the models/simulations reflect reality. If there are extraordinary circumstances prohibiting these

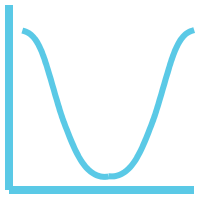
There will always be cases when live data for quantitative comparisons are unavailable (e.g., a new threat for which a surrogate has not yet been developed). In those instances, a well-reasoned and cautious approach should be taken to determine what, if any information, may be gleaned from M&S. In some instances, the absence of live data may prevent the accreditation of the M&S for use in the operational space. In other instances, it may reasonable to conclude that performance in one area of the operational space extends into a nearby operational space, where no live data are available. In the latter case, it is critical that the limitations of the M&S are understood and the uncertainty in the results quantified to the extent possible. Empirical models (a.k.a., emulators or meta-models) should be used to understand M&S outcomes across the operational space and assist in the uncertainty quantification in areas where there are no live data. In the operational space where no data are available, the results of the M&S should be discussed in the context of limitations.

DOT&E: Director, Operational Test and Evaluation

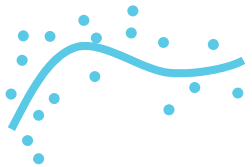
https://www.dote.osd.mil/Portals/97/pub/policies/2016/20140314_Guidance_on_Valid_of_Mod_Sim_used_in_OT_and_LF_Assess_(10601).pdf?ver=2019-08-19-144201-107

https://www.dote.osd.mil/Portals/97/pub/policies/2017/20170117_Clarification_on_Guidance_on_the_Validation_of_ModSim_used_in_OT_and_LF_Assess(15520).pdf?ver=2019-08-19-144121-123

IDA's Test Science team continues to develop methods to improve M&S VV&A



Statistical Risk Analysis



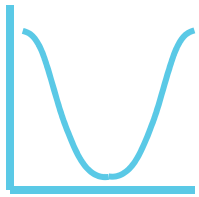
Metamodel Construction

Space-Filling DOE

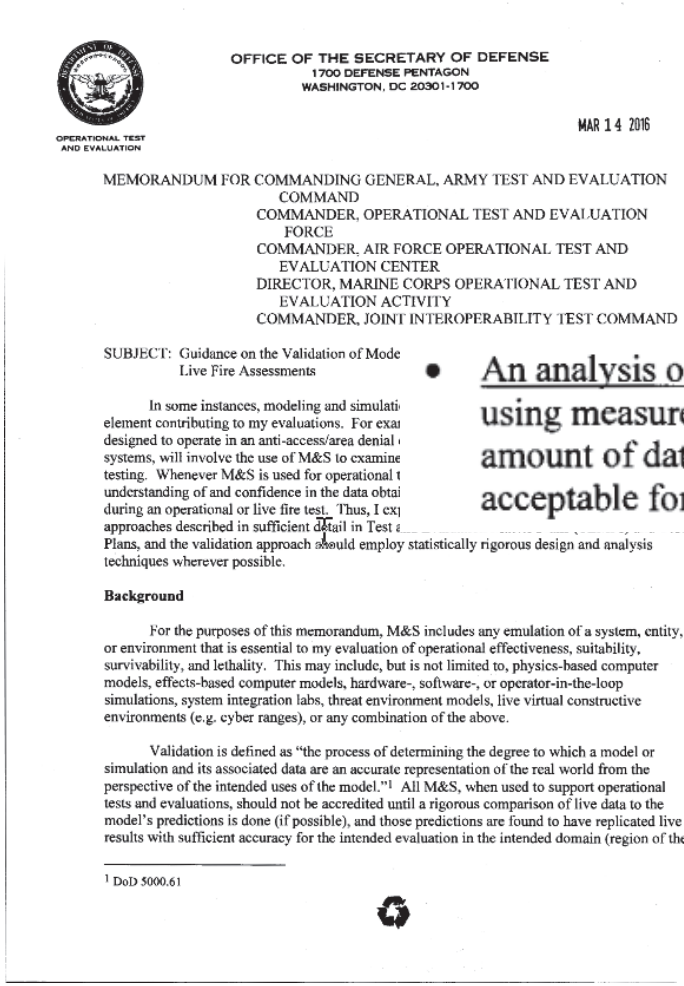
Statistical Validation

DOE: Design of Experiments; IDA: Institute for Defense Analyses; M&S: Modeling and Simulation; VV&A: Verification, Validation, and Accreditation

Are we adequately considering statistical risk?



DOT&E requires analysis of statistical risk for VV&A



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

OPERATIONAL TEST AND EVALUATION

MAR 14 2016

MEMORANDUM FOR COMMANDING GENERAL, 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

SUBJECT: Guidance on the Validation of Mode
Live Fire Assessments

In some instances, modeling and simulati
element contributing to my evaluations. For exa
designed to operate in an anti-access/area denial
systems, will involve the use of M&S to examine
testing. Whenever M&S is used for operational
understanding of and confidence in the data obtai
during an operational or live fire test. Thus, I ex
approaches described in sufficient detail in Test
Plans, and the validation approach should employ statistically rigorous design and analysis techniques wherever possible.

Background

For the purposes of this memorandum, M&S includes any emulation of a system, entity, or environment that is essential to my evaluation of operational effectiveness, suitability, survivability, and lethality. This may include, but is not limited to, physics-based computer models, effects-based computer models, hardware-, software-, or operator-in-the-loop simulations, system integration labs, threat environment models, live virtual constructive environments (e.g. cyber ranges), or any combination of the above.

Validation is defined as "the process of determining the degree to which a model or simulation and its associated data are an accurate representation of the real world from the perspective of the intended uses of the model."[1] All M&S, when used to support operational tests and evaluations, should not be accredited until a rigorous comparison of live data to the model's predictions is done (if possible), and those predictions are found to have replicated live results with sufficient accuracy for the intended evaluation in the intended domain (region of the

[1] DoD 5000.61

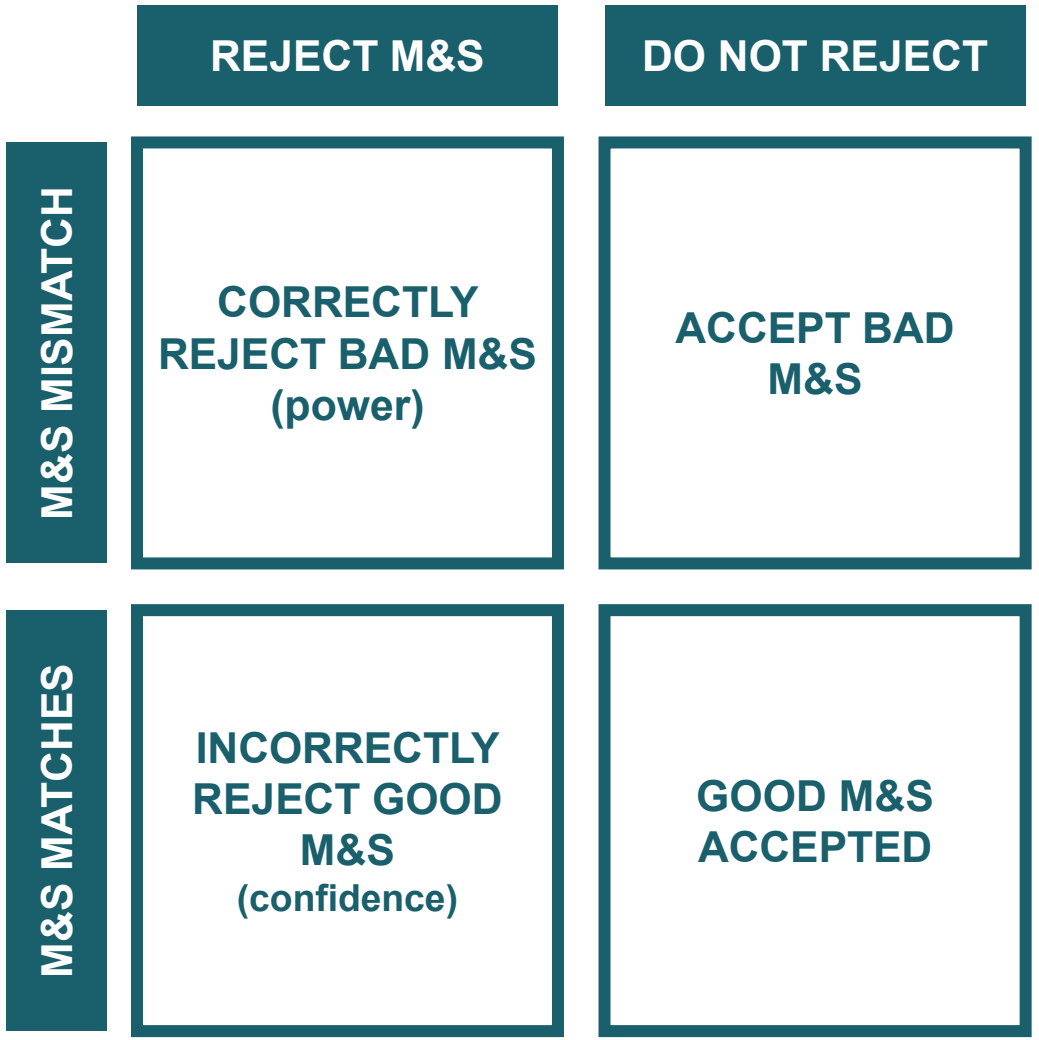
- An analysis of statistical risk. Where possible, statistical risk should be assessed using measures of merit such as power and confidence in order to scope the amount of data necessary for validation. Calculated uncertainties must be acceptable for accreditation for the intended use.

DOT&E: Director, Operational Test and Evaluation; VV&A: Verification, Validation, and Accreditation

The M&S community needs to think about statistical risk for VV&A differently than in live testing

Null Hypothesis: M&S outputs match real-world outcomes

Alternative Hypothesis: M&S outputs typically differ from real-world outcomes

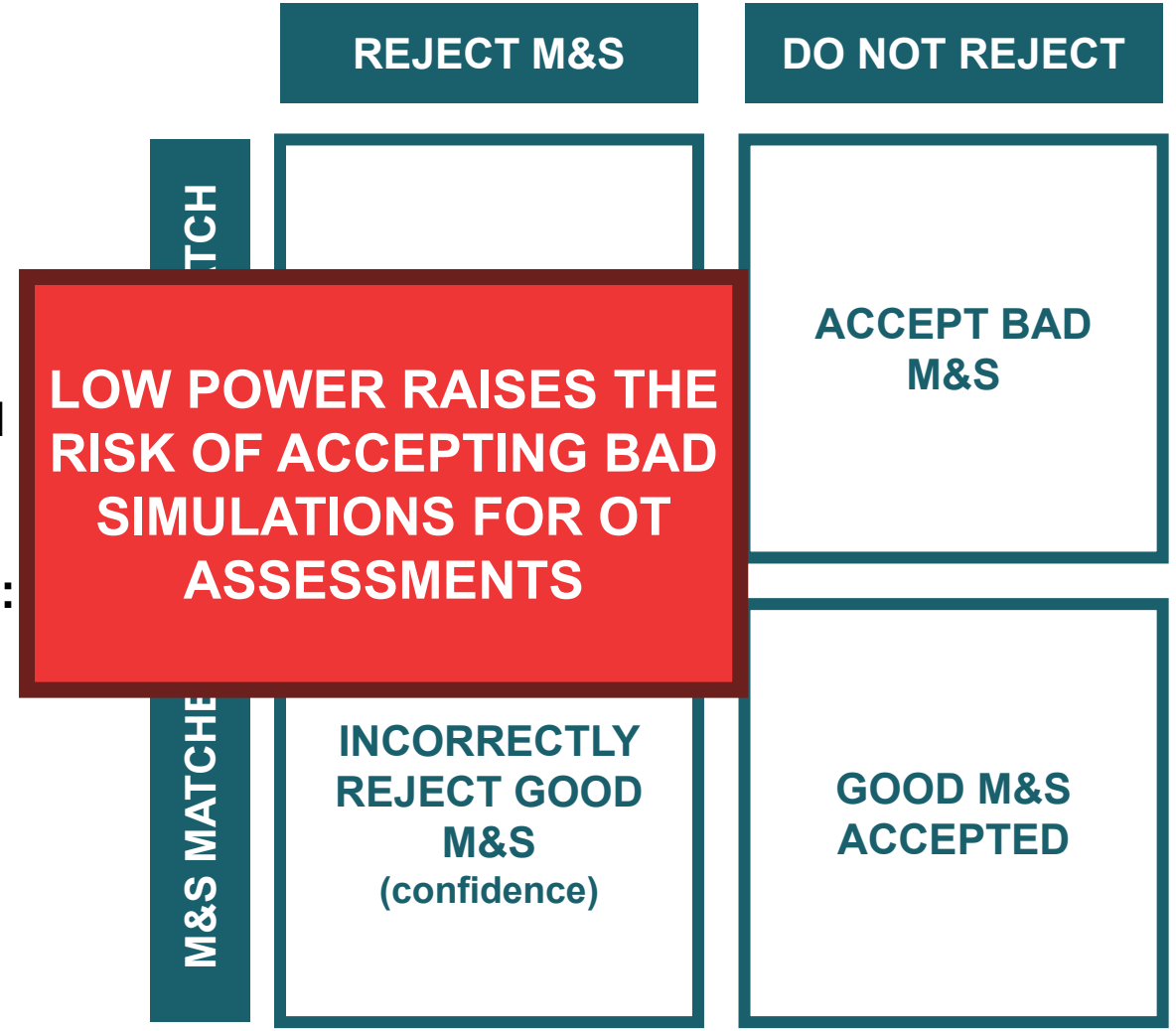


DOT&E: Director, Operational Test and Evaluation; M&S: Modeling and Simulation; VV&A: Verification, Validation, and Accreditation

The M&S community needs to think about statistical risk for VV&A differently than in live testing

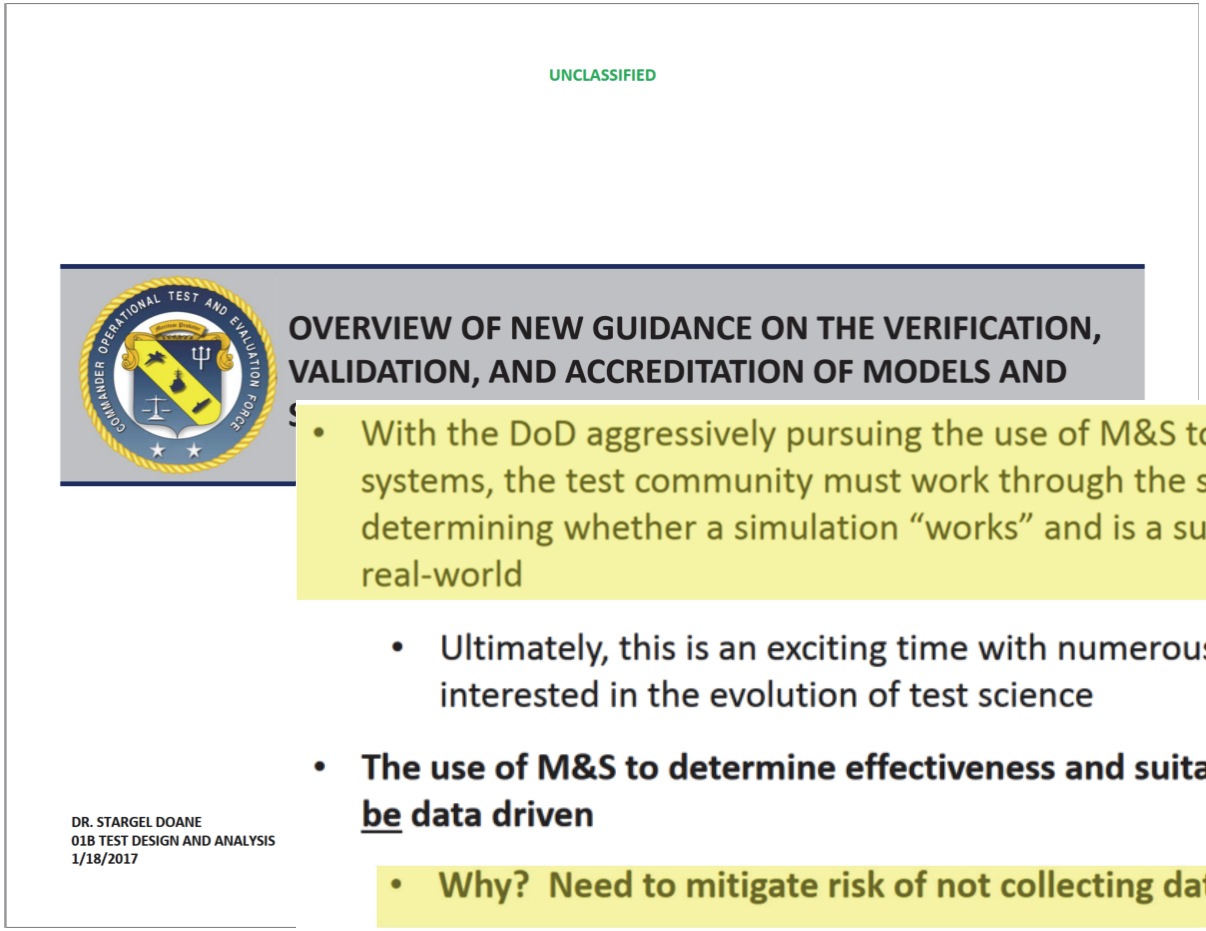
Null Hypothesis: M&S outputs match real-world outcomes

Alternative Hypothesis: M&S outputs typically differ from real-world outcomes



DOT&E: Director, Operational Test and Evaluation; M&S: Modeling and Simulation; VV&A: Verification, Validation, and Accreditation

OPTEVFOR correctly identified the risk associated with using an unsuitable M&S system



UNCLASSIFIED

OVERVIEW OF NEW GUIDANCE ON THE VERIFICATION, VALIDATION, AND ACCREDITATION OF MODELS AND

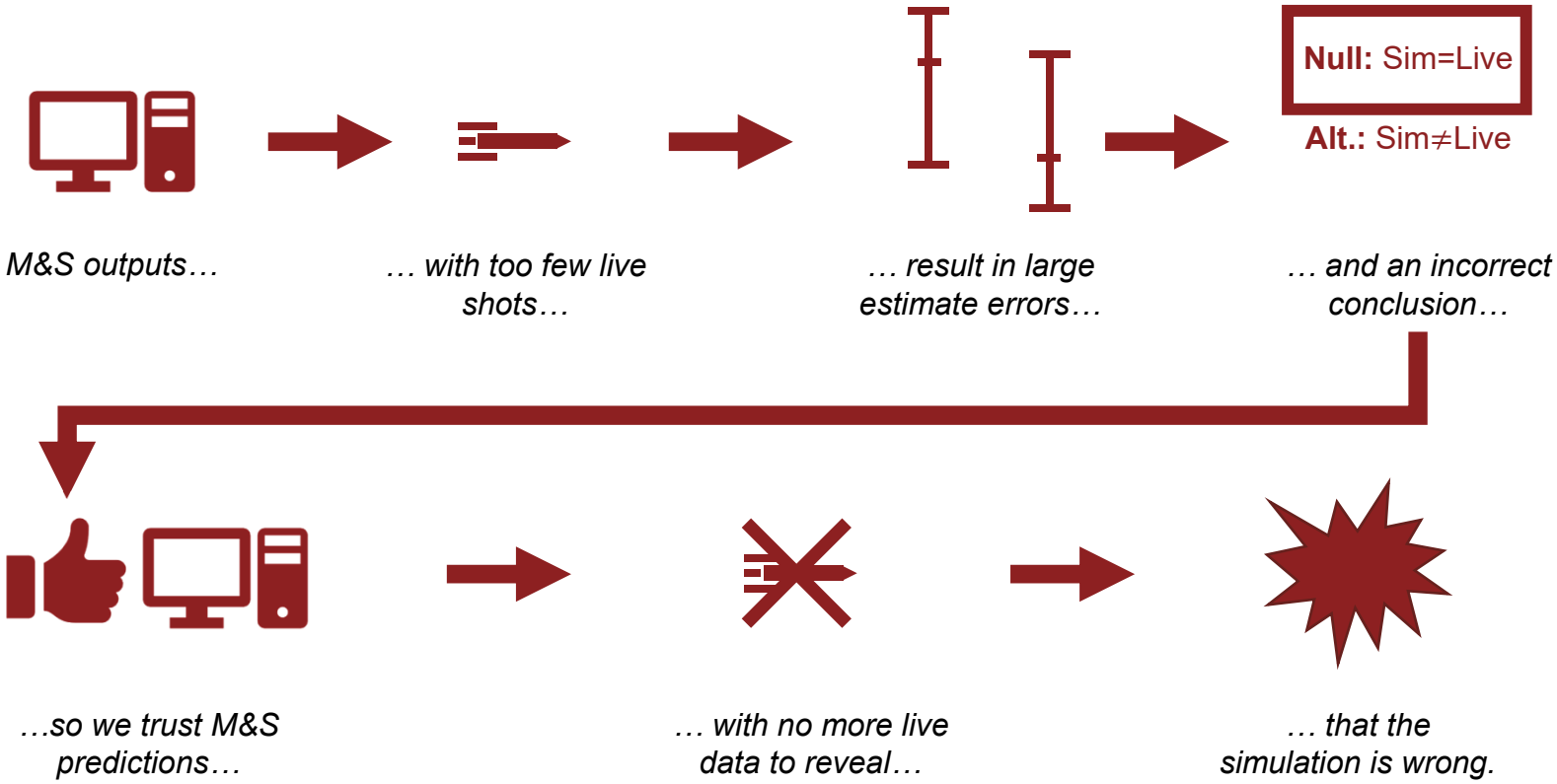
- With the DoD aggressively pursuing the use of M&S to design and evaluate complex systems, the test community must work through the significant challenge of determining whether a simulation "works" and is a sufficient representation of the real-world
  - Ultimately, this is an exciting time with numerous opportunities for those interested in the evolution of test science
- **The use of M&S to determine effectiveness and suitability in Operational Test must be data driven**
  - **Why? Need to mitigate risk of not collecting data from experiment**

DR. STARGEL DOANE
01B TEST DESIGN AND ANALYSIS
1/18/2017

M&S: Modeling and Simulation; DoD: Department of Defense

https://osd.deps.mil/org/dote-extranet/Guidance/Modeling%20and%20Simulation%20Guidance/Overview_of_New_Guidance_on_ModSim_VVA_1-18-2017.pdf

Low power can result in unsuitable M&S systems being used for OT assessments

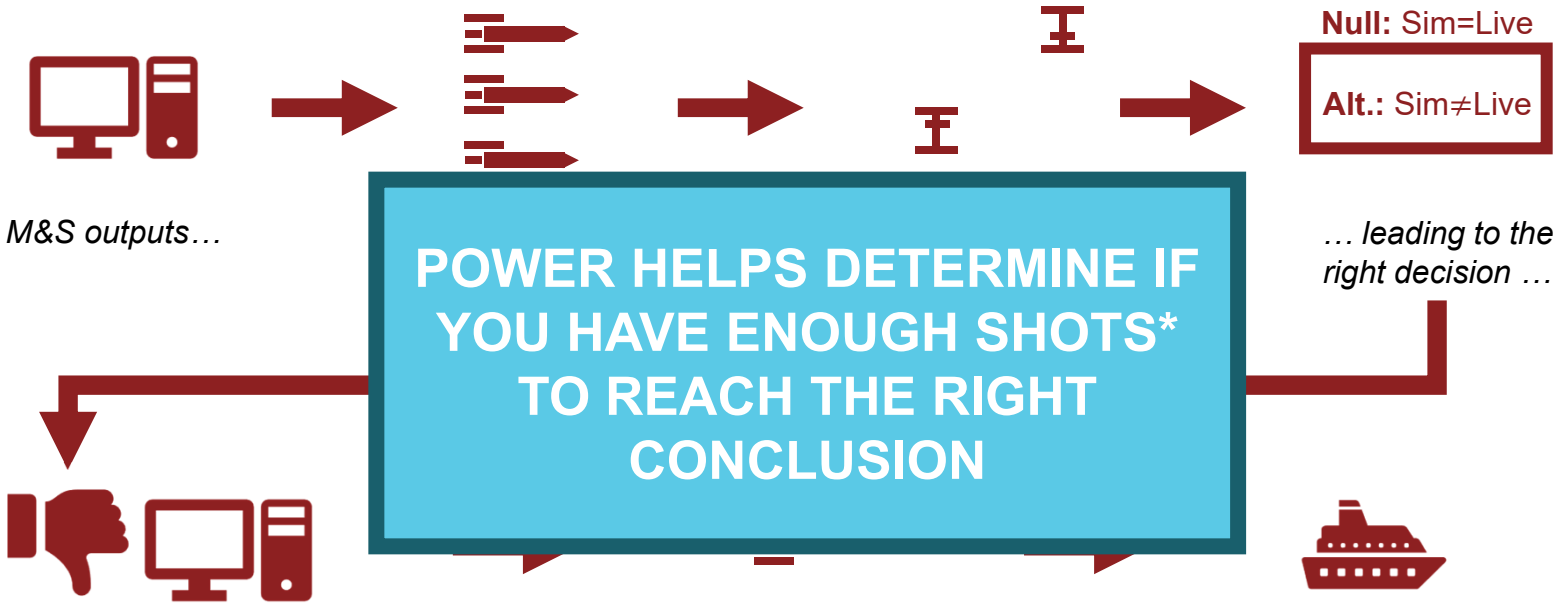


Alt.: Alternative; M&S: Modeling and Simulation; OT: Operational Test

High power reduces the risk of using unsuitable M&S systems for OT assessments

Alt.: Alternative; M&S: Modeling and Simulation; OT: Operational Test

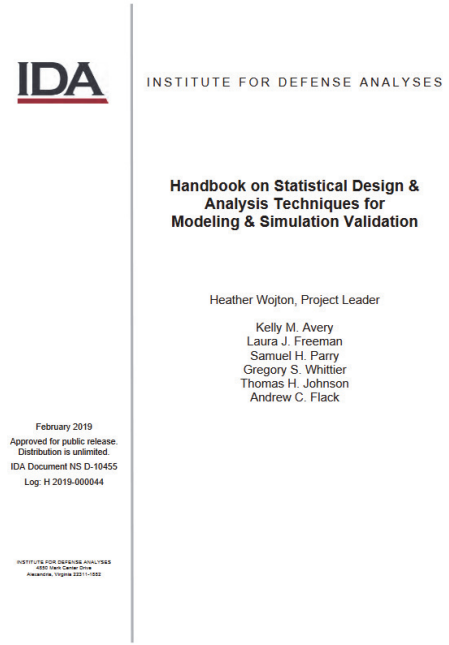
High power reduces the risk of using unsuitable M&S systems for OT assessments



* Among other things; power is not just the number of shots. Also, power is not all we should care about.

Alt.: Alternative; M&S: Modeling and Simulation; OT: Operational Test

The IDA M&S handbook recommends a number of procedures for comparing M&S outputs to live data


IDA

INSTITUTE FOR DEFENSE ANALYSES

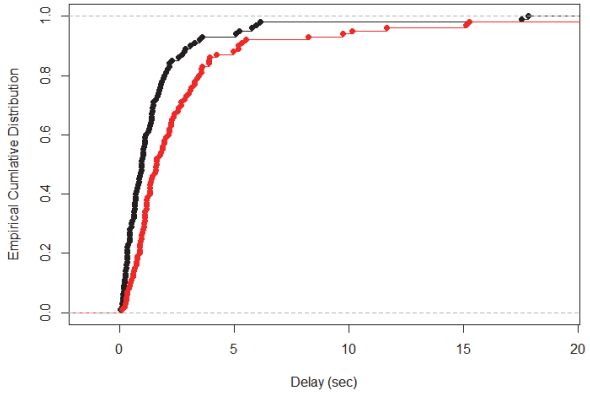
Handbook on Statistical Design & Analysis Techniques for Modeling & Simulation Validation

Heather Wojton, Project Leader

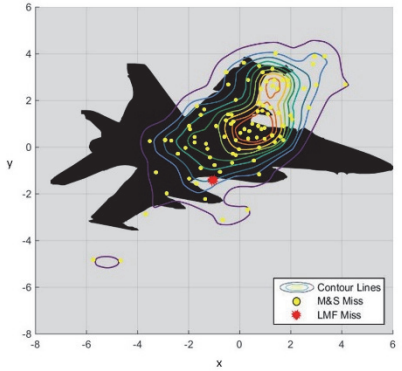
Kelly M. Avery
Laura J. Freeman
Samuel H. Parry
Gregory S. Whittier
Thomas H. Johnson
Andrew C. Flack

February 2019
Approved for public release. Distribution is unlimited.
IDA Document NS D-10455
Log: H 2019-000044

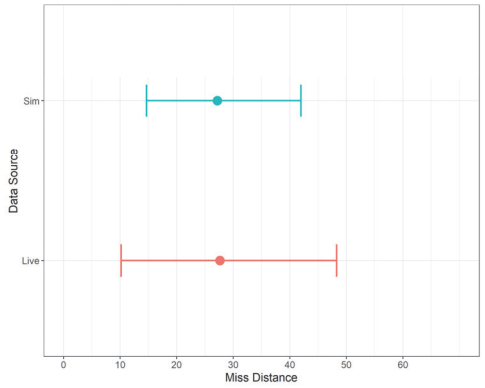
INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882



K-S Test



FCPT



Two-Sample t-Test

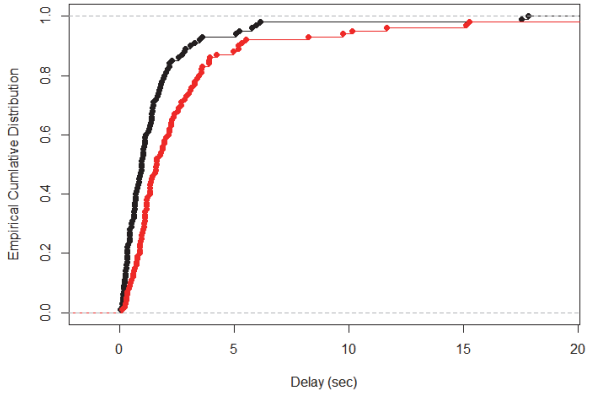
$$Detection\ Range = \beta_0 + \beta_1 TestType + \beta_2 Threat + \beta_3 (TestType * Threat) + \epsilon$$

Data Source Factor

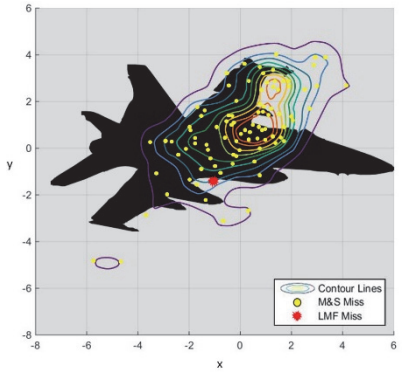
IDA: Institute for Defense Analyses; FCPT: Fisher Combined Probability Test; K-S: Kolmogorov-Smirnov; M&S: Modeling and Simulation

https://testscience.org/wp-content/uploads/formidable/20/Handbook-on-Statistical-Design-Analysis-Techniques-for-Modeling-Simulation-Validation-Reduced.pdf

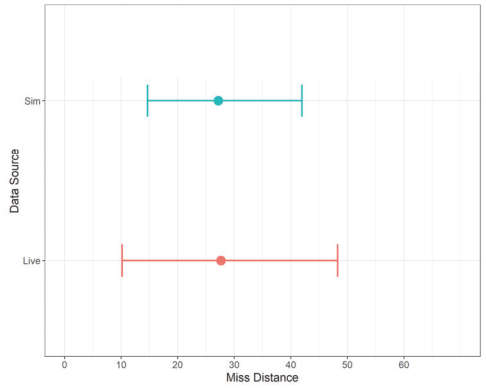
Statistical procedures often use a null hypothesis assuming the M&S is fine



K-S Test



FCPT



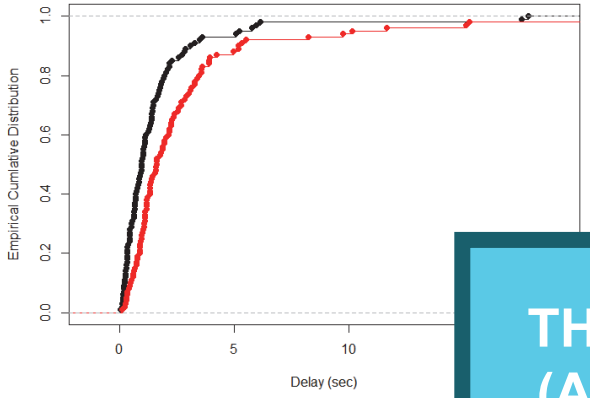
Two-Sample t-Test

$$\textit{Detection Range} = \beta_0 + \beta_1 TestType + \beta_2 Threat + \beta_3 (TestType * Threat) + \epsilon$$

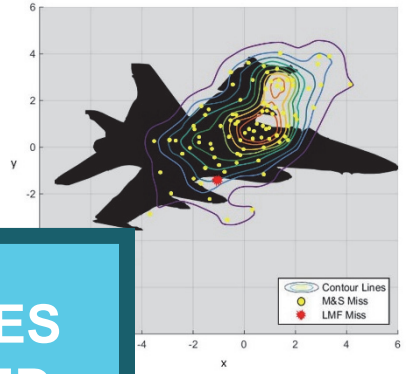
Data Source Factor

FCPT: Fisher Combined Probability Test; K-S: Kolmogorov-Smirnov; M&S: Modeling and Simulation

Statistical procedures often use a null hypothesis assuming the M&S is fine

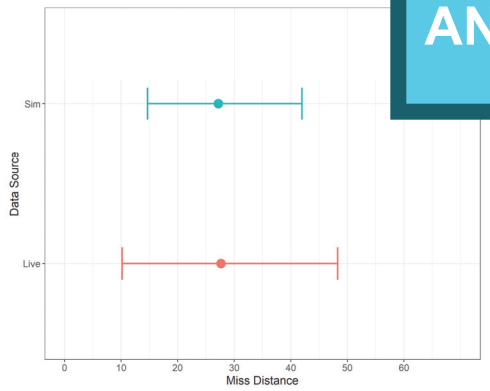


K-S Test



FCPT

THESE PROCEDURES (AND OTHERS) NEED ACCOMPANYING POWER ANALYSES IN PLANNING



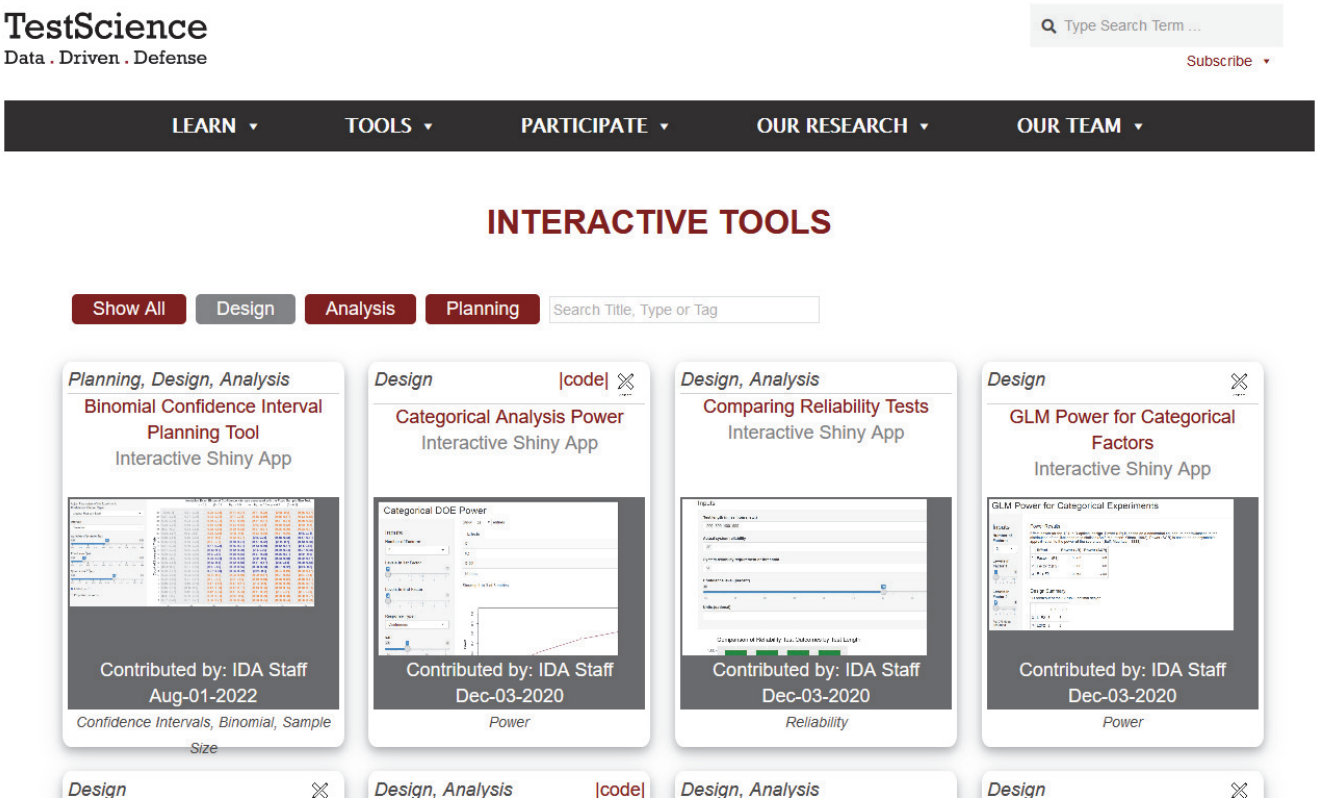
Two-Sample t-Test

$$Detection\ Range = \beta_0 + \beta_1 TestType + \beta_2 Threat + \beta_3 (TestType * Threat) + \epsilon$$

Data Source Factor

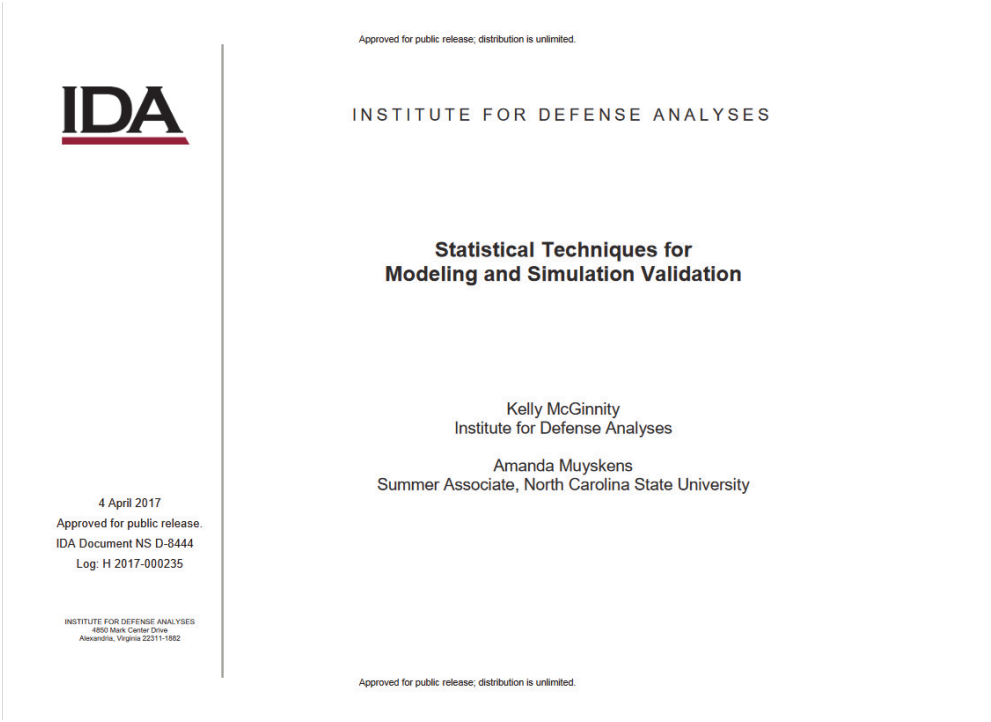
FCPT: Fisher Combined Probability Test; K-S: Kolmogorov-Smirnov; M&S: Modeling and Simulation

The Test Science website has tools for examining power in some contexts



https://testscience.org/interactive-tools/

Use custom Monte Carlo power analysis tailored to the data and statistical procedures planned



Approved for public release; distribution is unlimited.

IDA

INSTITUTE FOR DEFENSE ANALYSES

Statistical Techniques for Modeling and Simulation Validation

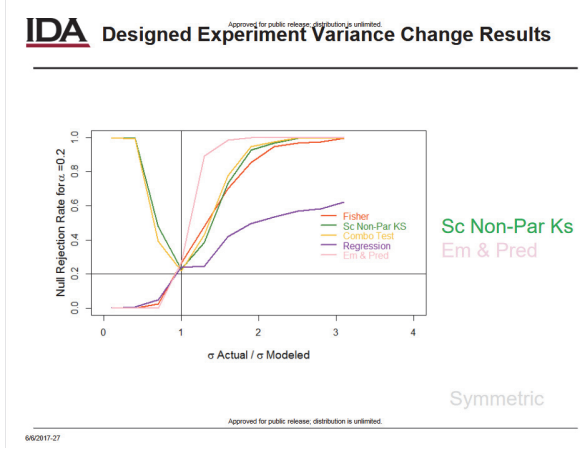
Kelly McGinnity
Institute for Defense Analyses

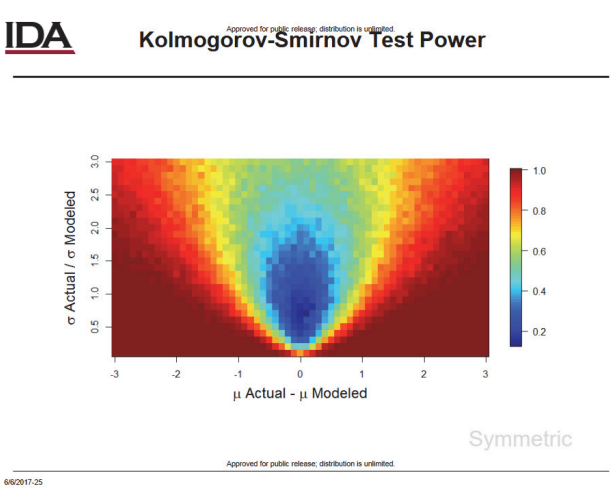
Amanda Muyskens
Summer Associate, North Carolina State University

4 April 2017
Approved for public release.
IDA Document NS D-8444
Log: H 2017-000235

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

Approved for public release; distribution is unlimited.

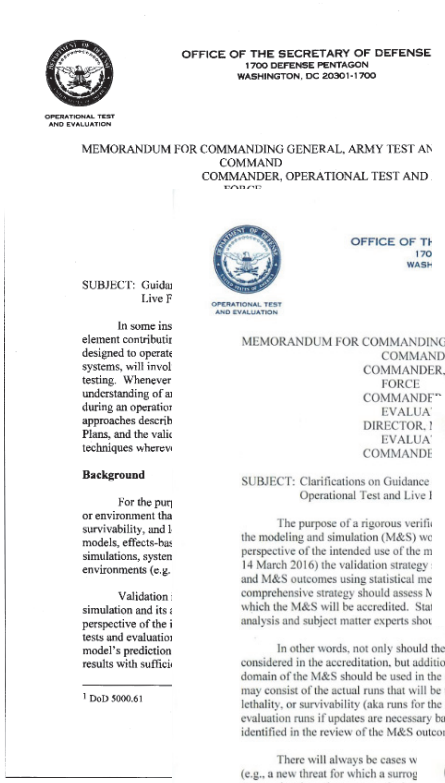




https://www.ida.org/-/media/feature/publications/s/st/statistical-techniques-for-modeling-and-simulation-validation/d-8694.ashx

Will the data we collect reveal important or unanticipated trends in M&S outputs?

VV&A requires comparing live data and M&S outputs



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

OPERATIONAL TEST AND EVALUATION

MEMORANDUM FOR COMMANDING GENERAL, ARMY TEST AN
COMMAND
COMMANDER, OPERATIONAL TEST AND

SUBJECT: Guidu
Live F

In some ins
element contributir
designed to operate
systems, will invol
testing. Whenever
understanding of a
during an operatior
approaches describ
Plans, and the valic
techniques wherev

Background

For the pur
or environment tha
survivability, and l
models, effects-bas
simulations, systen
environments (e.g.

Validation
simulation and its
perspective of the i
tests and evaluation
model's prediction
results with sufficie

[1] DoD 5000.61

OFFICE OF TH
170
WASH

OPERATIONAL TEST AND EVALUATION

MEMORANDUM FOR COMMANDING GENERAL, ARMY TEST AND EVALUATION
COMMAND
COMMANDER, OPERATIONAL TEST AND EVALUATION
FORCE
COMMANDE
EVALUA
DIRECTOR,
EVALUA
COMMANDE

SUBJECT: Clarifications on Guidance
Operational Test and Live

The purpose of a rigorous verifi
the modeling and simulation (M&S) wo
perspective of the intended use of the m
14 March 2016) the validation strategy
and M&S outcomes using statistical me
comprehensive strategy should assess M
which the M&S will be accredited. Stat
analysis and subject matter experts shou

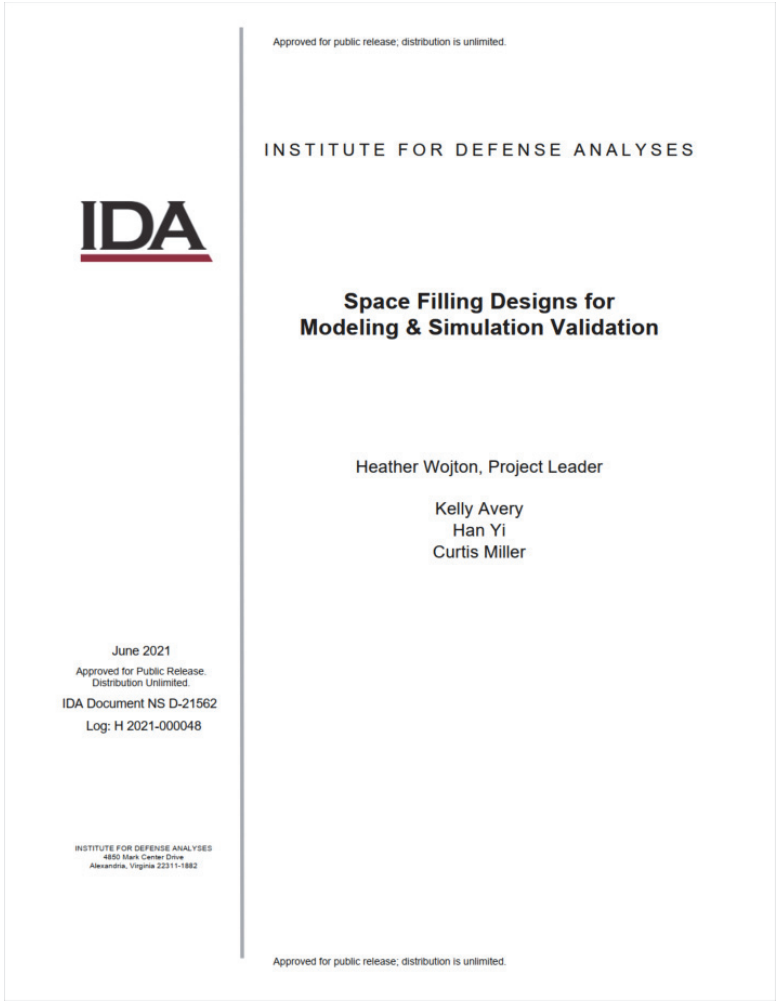
In other words, not only should the simulation runs that match live test conditions be considered in the accreditation, but additionally, M&S runs that span the entire operational domain of the M&S should be used in the accreditation decision. This larger set of M&S runs may consist of the actual runs that will be used for the evaluation of effectiveness, suitability, lethality, or survivability (aka runs for the record). However, they may also differ from the final evaluation runs if updates are necessary based on the comparison to live data or other problems identified in the review of the M&S outcomes.

There will always be cases w
(e.g., a new threat for which a surrog
reasoned and cautious approach shou
gleaned from M&S. In some instan
the M&S for use in the operational s
performance in one area of the opera
no live data are available. In the latt
understood and the uncertainty in th
(a.k.a., emulators or meta-models) sh
operational space and assist in the ur

- The plan for collecting the necessary live and simulation data for M&S validation. The plan should articulate a method for strategically varying the factors that affect system performance with respect to the response variables of interest. It also should describe whether live data will cover the entire operational envelope to be explored with M&S or only a portion of the envelope. If only a portion of the envelope is covered, the plan should clearly describe which portion.
- A strategically selected set of test points designed to compare the M&S and live data. These points should be selected using the principles of experimental design to span as much of the operational space as possible within the constraints of what is feasible to conduct in live testing. See my guidance on design of experiments (DOE) and M&S for more information on selecting a set of comparison data.
- A robust design for the M&S that systematically covers the range of operationally realistic inputs over which the model will be accredited. Space-filling design methodologies are preferred because they not only maximize opportunities for problem detection, but also support the development of statistical emulators that can be compared to live data and assist in quantifying uncertainty in the M&S.

M&S: Modeling and Simulation; VV&A: Verification, Validation, and Accreditation

IDA publications introduce and recommend M&S DOE best practices



Approved for public release; distribution is unlimited.

INSTITUTE FOR DEFENSE ANALYSES

IDA

Space Filling Designs for Modeling & Simulation Validation

Heather Wojton, Project Leader

Kelly Avery
Han Yi
Curtis Miller

June 2021

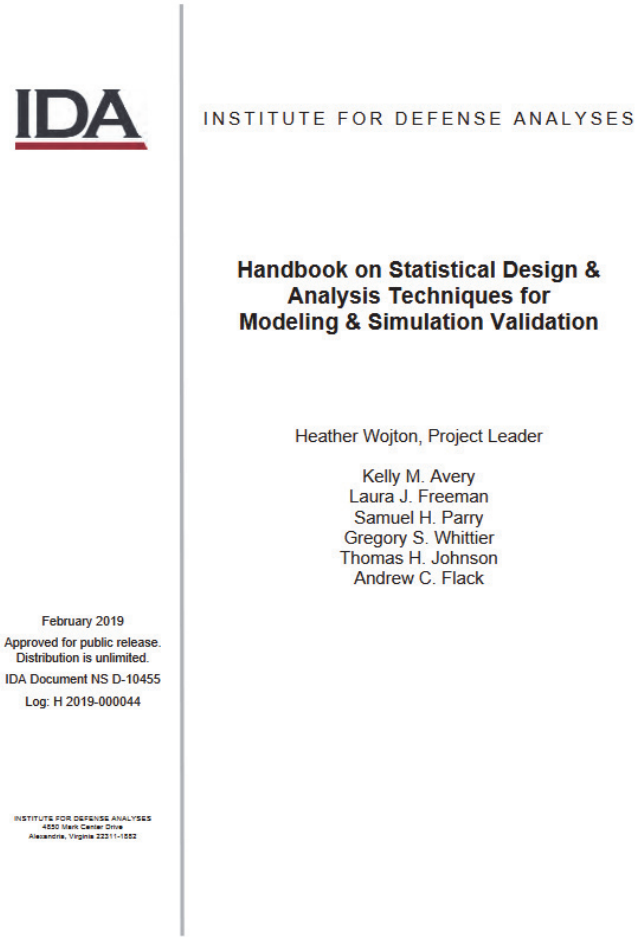
Approved for Public Release. Distribution Unlimited.

IDA Document NS D-21562

Log: H 2021-000048

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

Approved for public release; distribution is unlimited.



IDA

INSTITUTE FOR DEFENSE ANALYSES

Handbook on Statistical Design & Analysis Techniques for Modeling & Simulation Validation

Heather Wojton, Project Leader

Kelly M. Avery
Laura J. Freeman
Samuel H. Parry
Gregory S. Whittier
Thomas H. Johnson
Andrew C. Flack

February 2019

Approved for public release. Distribution is unlimited.

IDA Document NS D-10455

Log: H 2019-000044

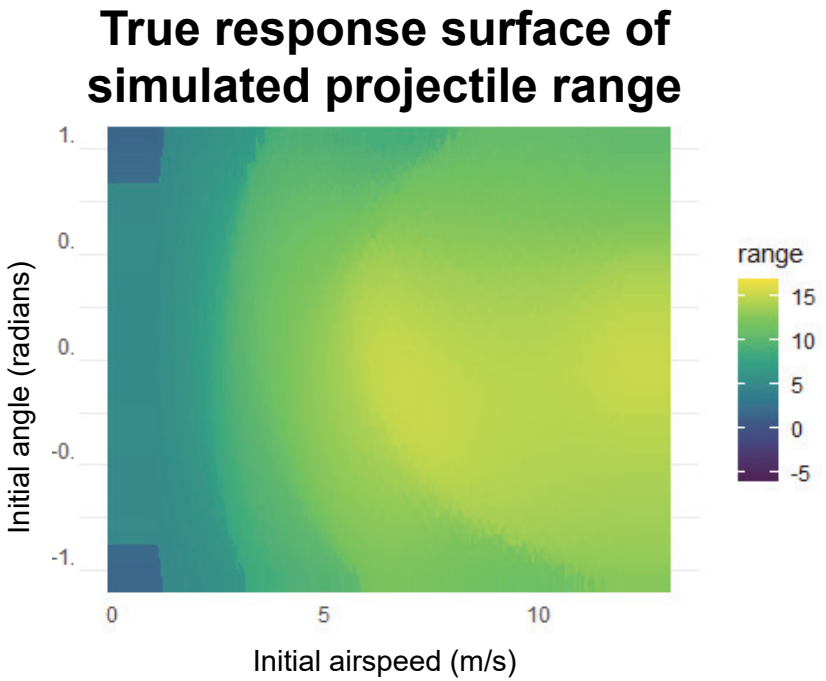
INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

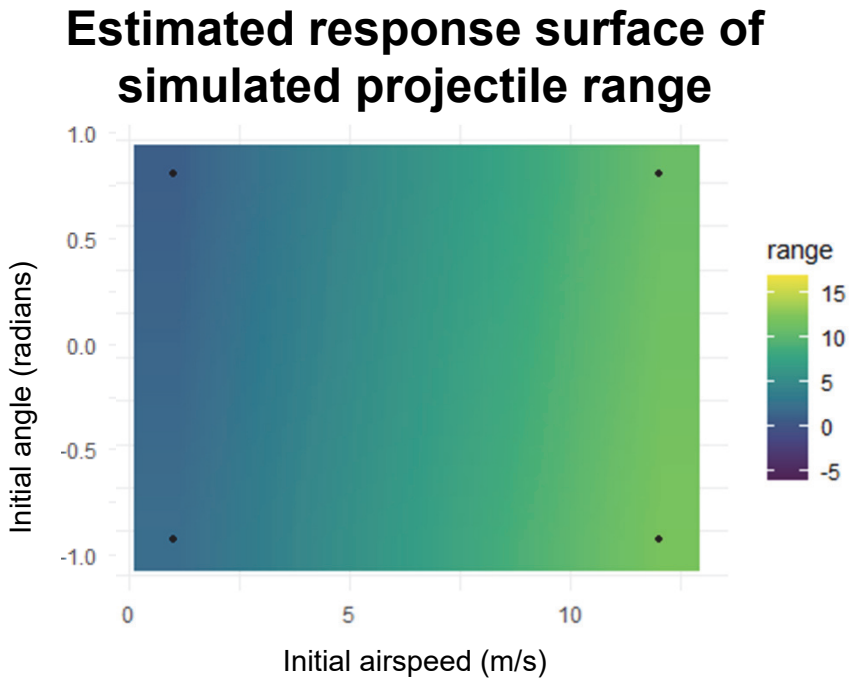
IDA: Institute for Defense Analyses; SFD: Space-Filling Design; DOE: Design of experiments

https://testscience.org/wp-content/uploads/formidable/20/SFD_Literature_Review_Final.html

Space-filling design of experiments helps recover more trends in simulation outputs

A factorial design with a simple linear model fitted will not accurately describe the M&S system's behavior.

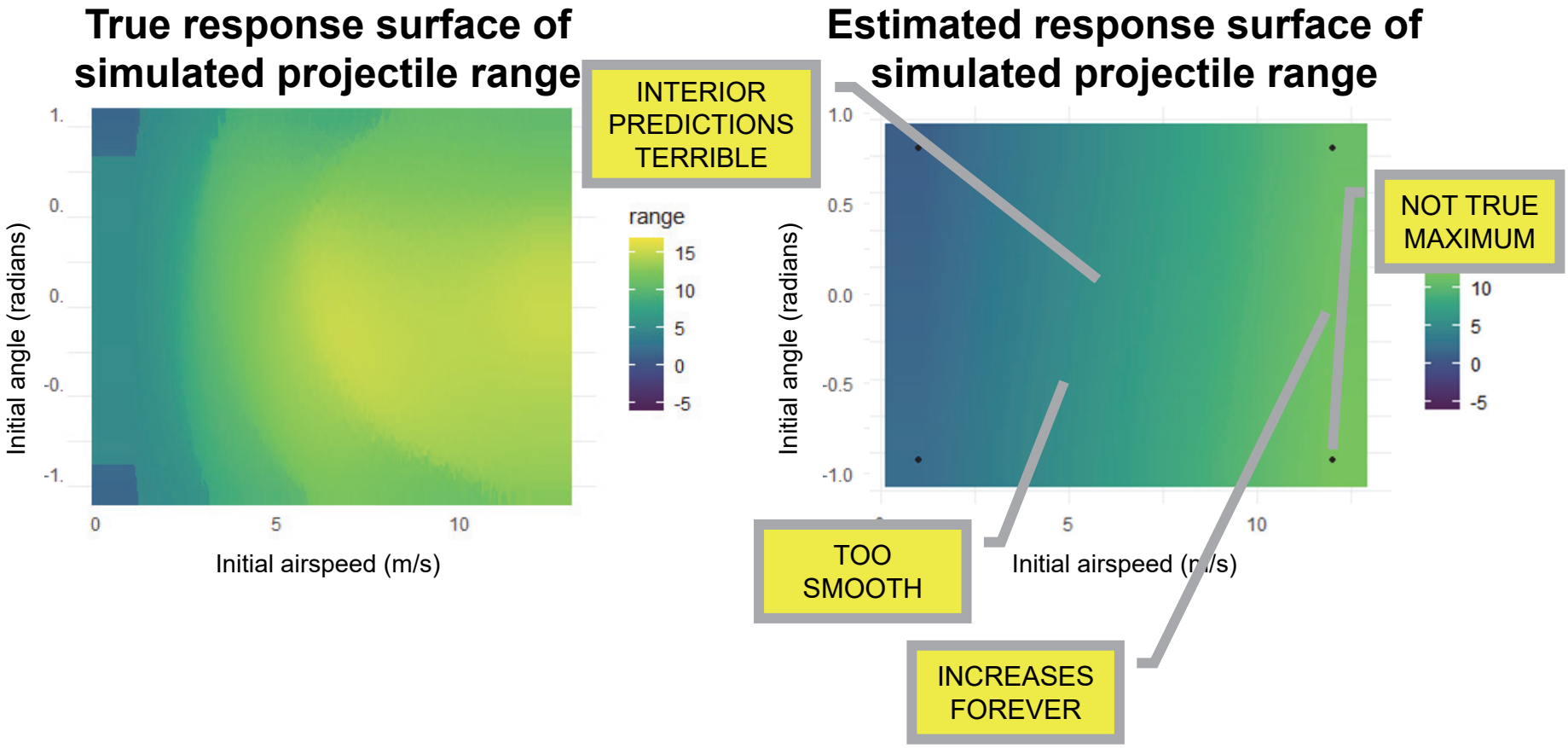




M&S: Modeling and Simulation; SFD: Space-Filling Design; WRPA: World Record Paper Airplane

Space-filling design of experiments helps recover more trends in simulation outputs

A factorial design with a simple linear model fitted will not accurately describe the M&S system's behavior.

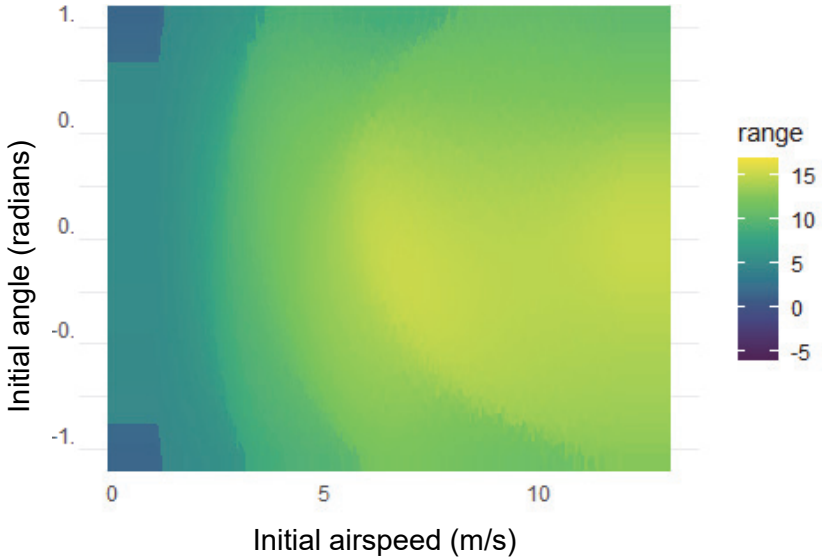


M&S: Modeling and Simulation; SFD: Space-Filling Design; WRPA: World Record Paper Airplane

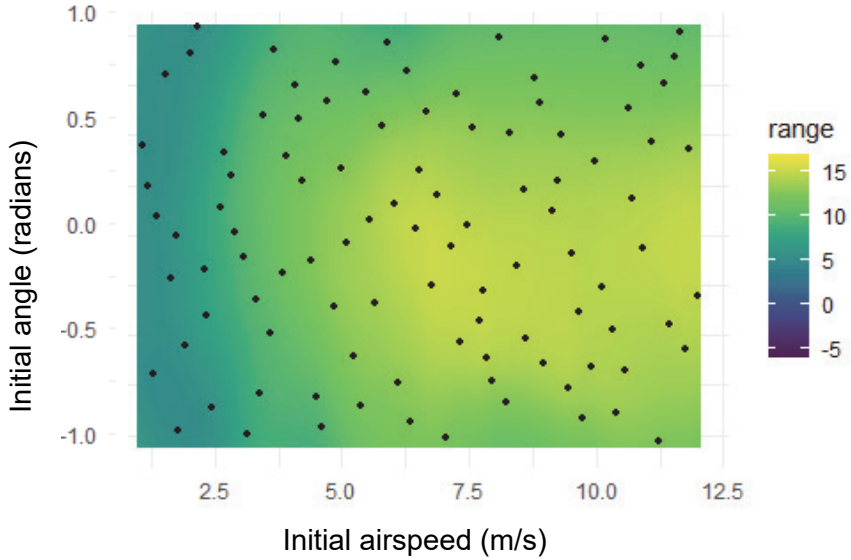
Space-filling design of experiments helps recover more trends in simulation outputs

Analyzing the flights with a Gaussian Process model via a Space-Filling Design yields a good approximation to simulation output.

True response surface of simulated projectile range



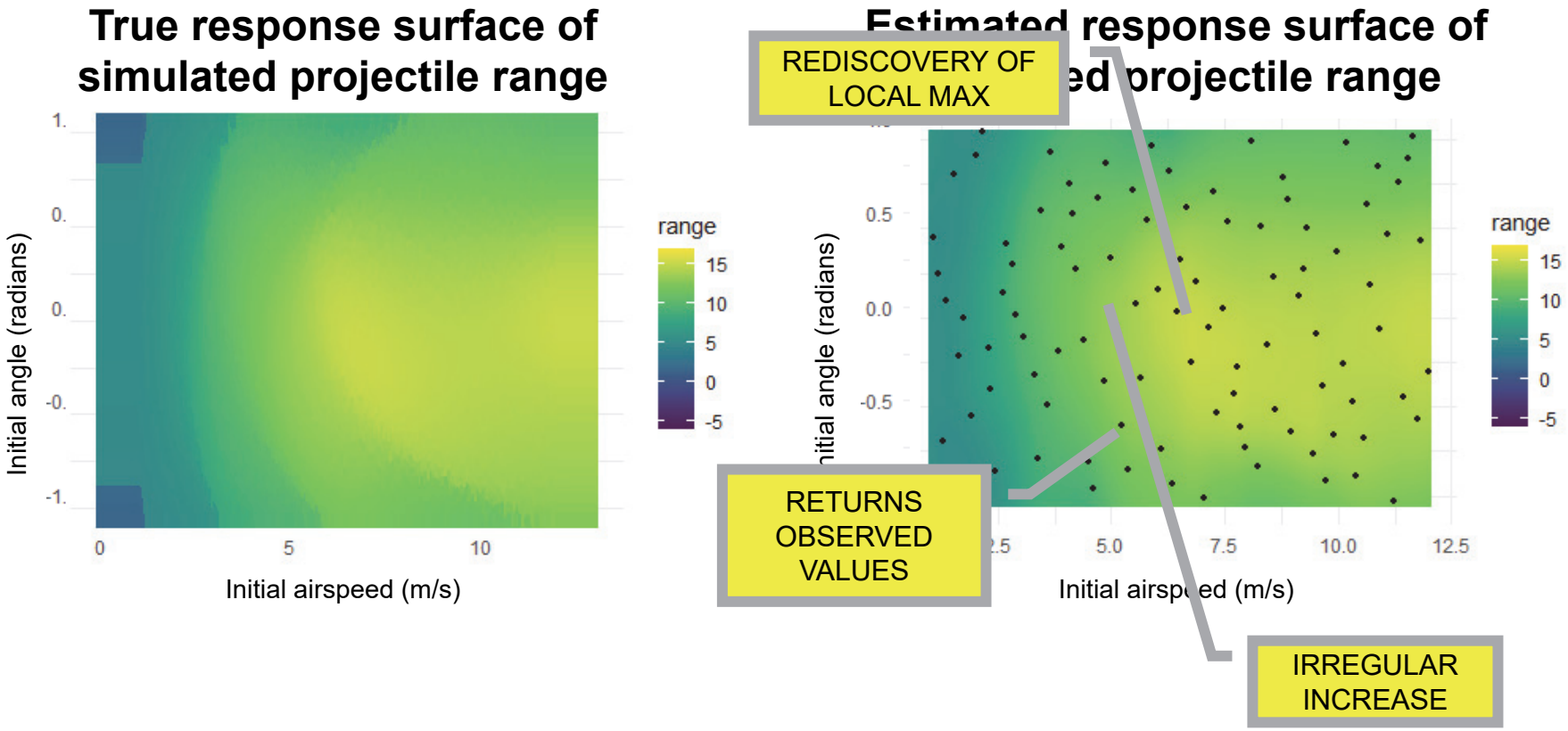
Estimated response surface of simulated projectile range



GP: Gaussian Process; SFD: Space-Filling Design; WRPA: World Record Paper Airplane

Space-Filling design of experiments helps recover more trends in simulation outputs

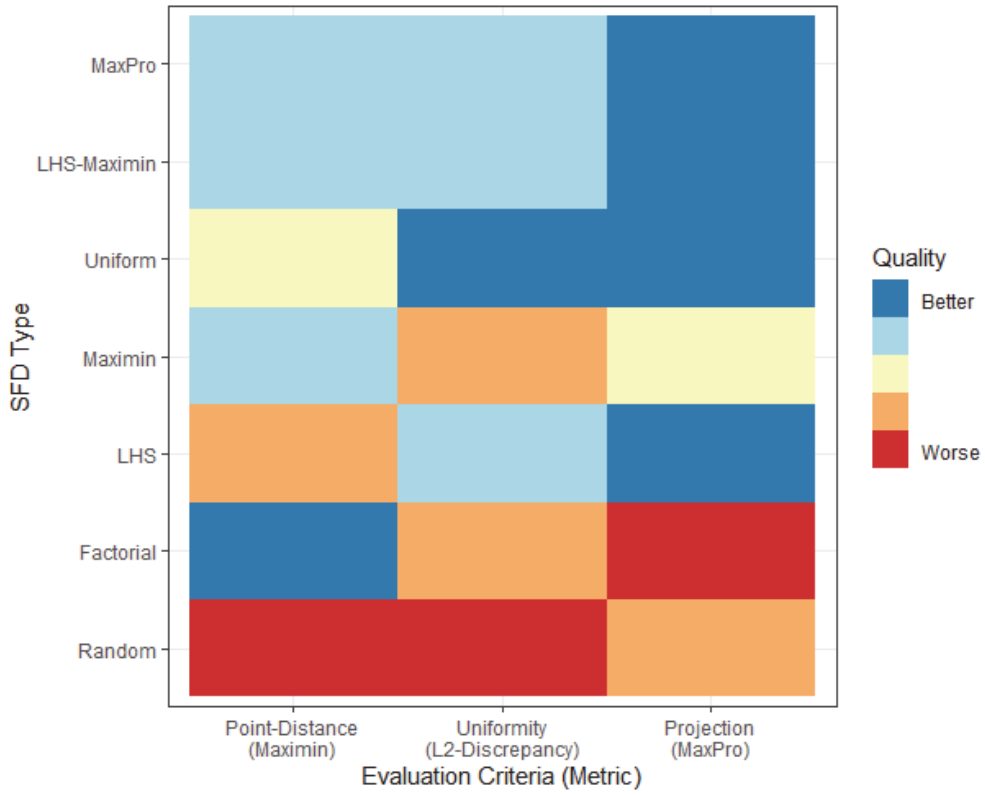
Analyzing the flights with a Gaussian Process model via a Space-Filling Design yields a good approximation to simulation output.

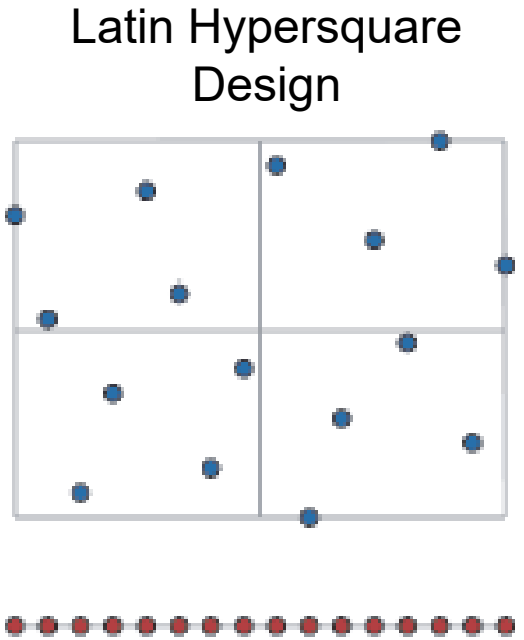


GP: Gaussian Process; SFD: Space-Filling Design; WRPA: World Record Paper Airplane

I prefer maximin sliced Latin hypersquare designs (maximin SLHD) and MaxPro SFDs

General recommendations:
Maximin (Sliced) LHD or MaxPro





SFD: Space-Filling Design

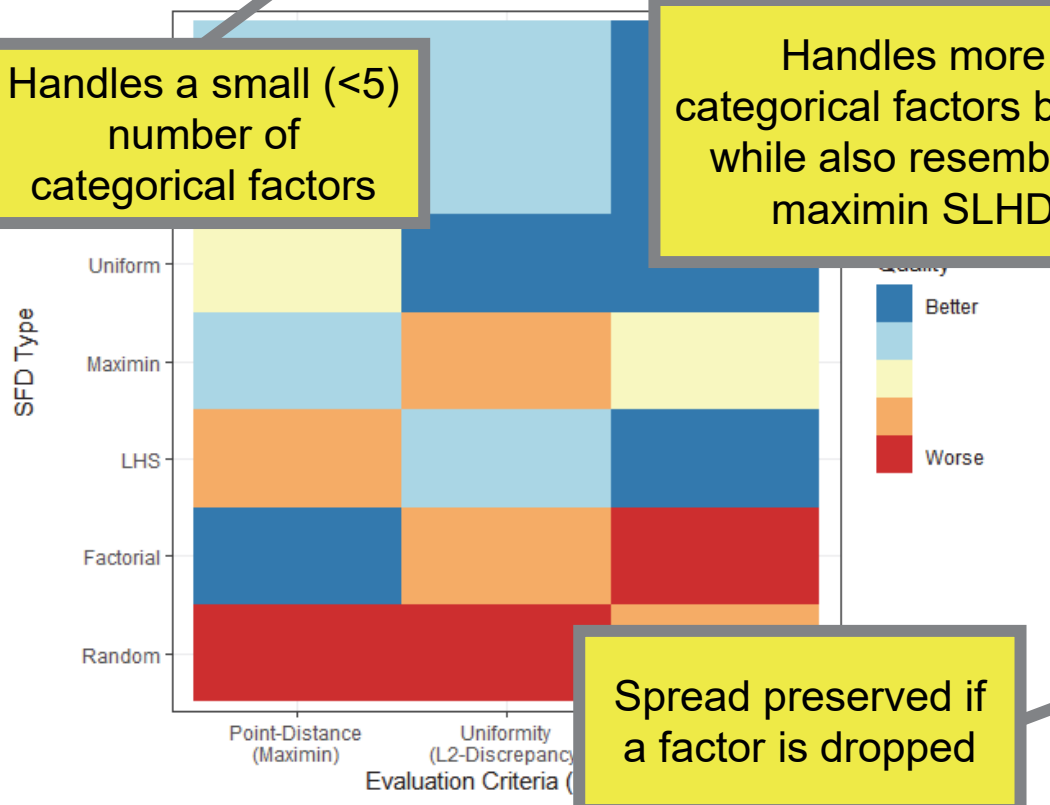
I prefer maximin sliced Latin hypersquare designs (maximin SLHD) and MaxPro SFDs

General recommendations:
Maximin (Sliced) LHD or MaxPro

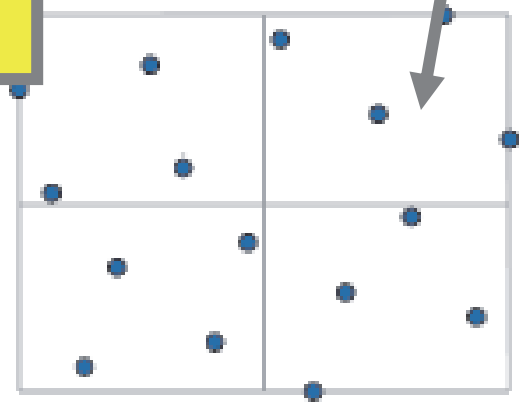
Handles a small (<5) number of categorical factors

Handles more categorical factors better, while also resembling maximin SLHD

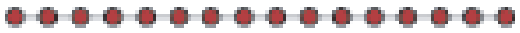
Good spread in the design space



Latin Hypersquare Design



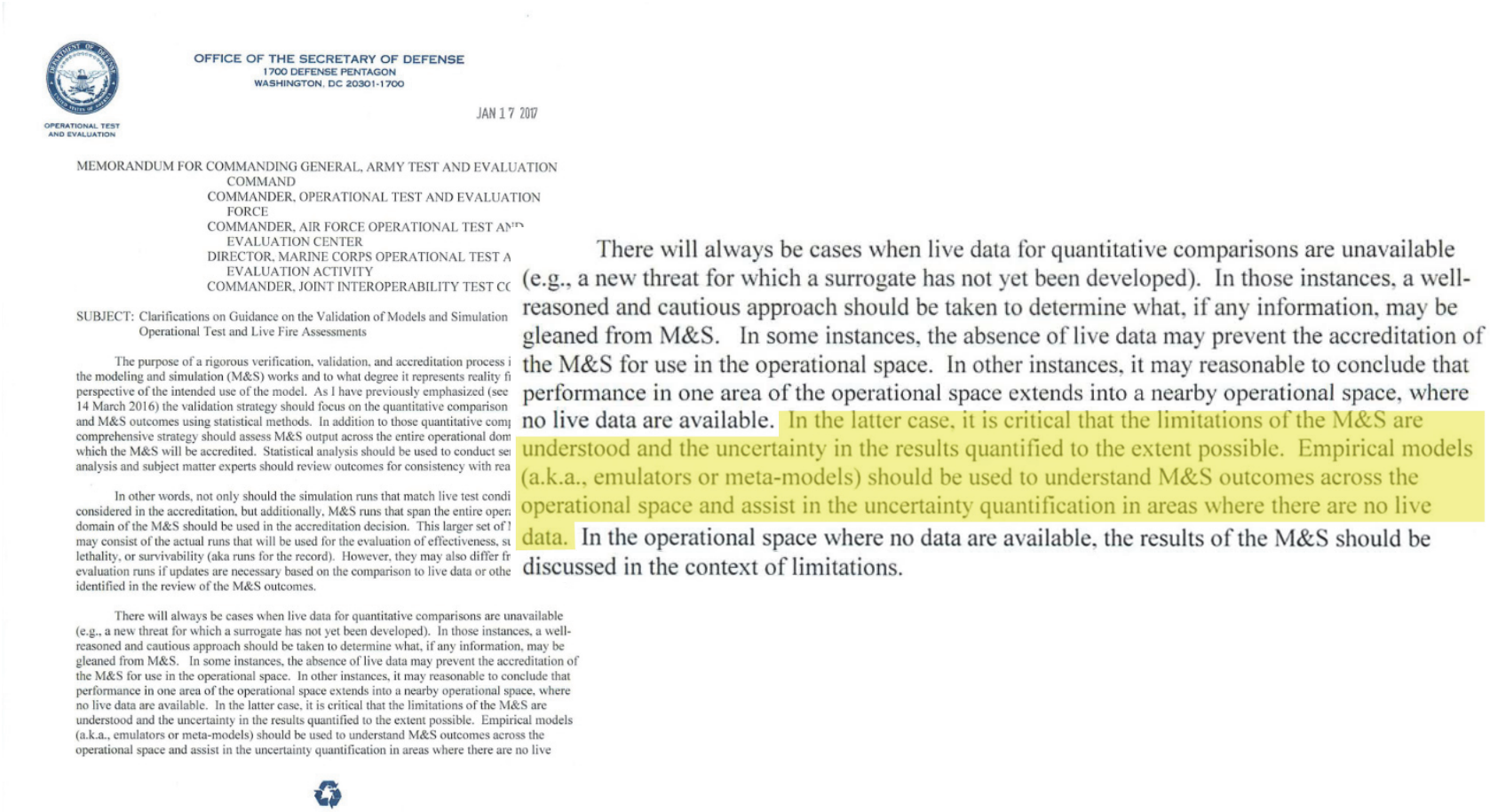
Spread preserved if a factor is dropped



SFD: Space-Filling Design

Are we accurately characterizing M&S predictions?

DOT&E wants to see high-level characterization of M&S systems' outputs



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

OPERATIONAL TEST AND EVALUATION

JAN 17 2017

MEMORANDUM FOR COMMANDING GENERAL, ARMY TEST AND EVALUATION COMMAND
COMMANDER, OPERATIONAL TEST AND EVALUATION FORCE
COMMANDER, AIR FORCE OPERATIONAL TEST AND EVALUATION CENTER
DIRECTOR, MARINE CORPS OPERATIONAL TEST A EVALUATION ACTIVITY
COMMANDER, JOINT INTEROPERABILITY TEST C

SUBJECT: Clarifications on Guidance on the Validation of Models and Simulation Operational Test and Live Fire Assessments

The purpose of a rigorous verification, validation, and accreditation process i the modeling and simulation (M&S) works and to what degree it represents reality fi perspective of the intended use of the model. As I have previously emphasized (see 14 March 2016) the validation strategy should focus on the quantitative comparison and M&S outcomes using statistical methods. In addition to those quantitative comp comprehensive strategy should assess M&S output across the entire operational dom which the M&S will be accredited. Statistical analysis should be used to conduct se analysis and subject matter experts should review outcomes for consistency with rea

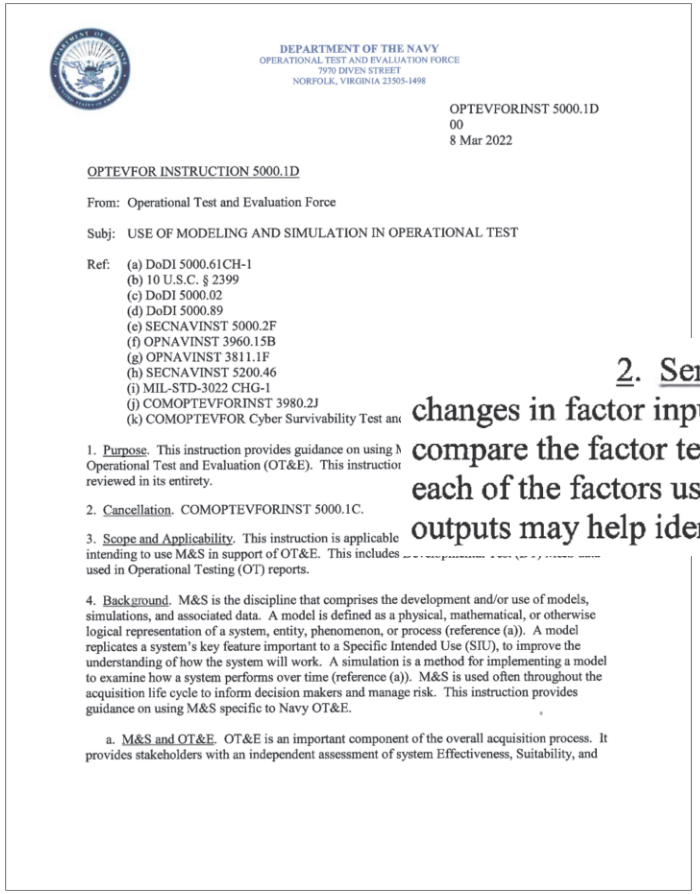
In other words, not only should the simulation runs that match live test condi considered in the accreditation, but additionally, M&S runs that span the entire oper domain of the M&S should be used in the accreditation decision. This larger set of may consist of the actual runs that will be used for the evaluation of effectiveness, s lethality, or survivability (aka runs for the record). However, they may also differ fr evaluation runs if updates are necessary based on the comparison to live data or othe identified in the review of the M&S outcomes.

There will always be cases when live data for quantitative comparisons are unavailable (e.g., a new threat for which a surrogate has not yet been developed). In those instances, a well-reasoned and cautious approach should be taken to determine what, if any information, may be gleaned from M&S. In some instances, the absence of live data may prevent the accreditation of the M&S for use in the operational space. In other instances, it may reasonable to conclude that performance in one area of the operational space extends into a nearby operational space, where no live data are available. In the latter case, it is critical that the limitations of the M&S are understood and the uncertainty in the results quantified to the extent possible. Empirical models (a.k.a., emulators or meta-models) should be used to understand M&S outcomes across the operational space and assist in the uncertainty quantification in areas where there are no live

There will always be cases when live data for quantitative comparisons are unavailable (e.g., a new threat for which a surrogate has not yet been developed). In those instances, a well-reasoned and cautious approach should be taken to determine what, if any information, may be gleaned from M&S. In some instances, the absence of live data may prevent the accreditation of the M&S for use in the operational space. In other instances, it may reasonable to conclude that performance in one area of the operational space extends into a nearby operational space, where no live data are available. In the latter case, it is critical that the limitations of the M&S are understood and the uncertainty in the results quantified to the extent possible. Empirical models (a.k.a., emulators or meta-models) should be used to understand M&S outcomes across the operational space and assist in the uncertainty quantification in areas where there are no live data. In the operational space where no data are available, the results of the M&S should be discussed in the context of limitations.

DOT&E: Director, Operational Test and Evaluation; M&S: Modeling and Simulation; VV&A: Verification, Validation, and Accreditation

OPTEVFOR calls for meta-model use in OPTEVFORINST 5000.1D



DEPARTMENT OF THE NAVY
OPERATIONAL TEST AND EVALUATION FORCE
7970 DIVEN STREET
NORFOLK, VIRGINIA 23505-1498

OPTEVFORINST 5000.1D
00
8 Mar 2022

OPTEVFOR INSTRUCTION 5000.1D

From: Operational Test and Evaluation Force

Subj: USE OF MODELING AND SIMULATION IN OPERATIONAL TEST

Ref: (a) DoDI 5000.61CH-1
(b) 10 U.S.C. § 2399
(c) DoDI 5000.02
(d) DoDI 5000.89
(e) SECNAVINST 5000.2F
(f) OPNAVINST 3960.15B
(g) OPNAVINST 3811.1F
(h) SECNAVINST 5200.46
(i) MIL-STD-3022 CHG-1
(j) COMOPTEVFORINST 3980.2J
(k) COMOPTEVFOR Cyber Survivability Test an

1. Purpose. This instruction provides guidance on using M
Operational Test and Evaluation (OT&E). This instructio
reviewed in its entirety.

2. Cancellation. COMOPTEVFORINST 5000.1C.

3. Scope and Applicability. This instruction is applicable
intending to use M&S in support of OT&E. This includes
used in Operational Testing (OT) reports.

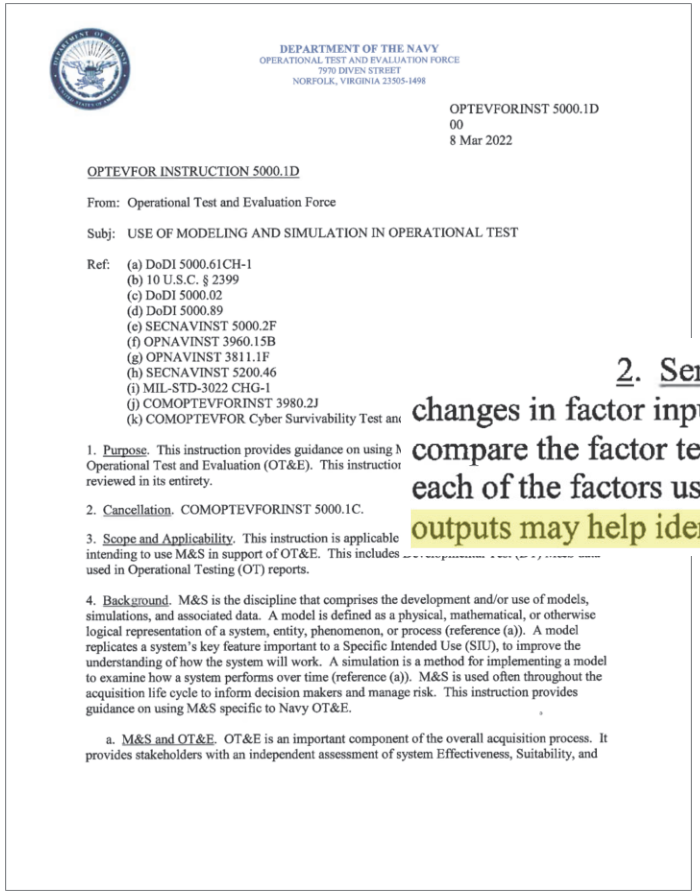
4. Background. M&S is the discipline that comprises the development and/or use of models, simulations, and associated data. A model is defined as a physical, mathematical, or otherwise logical representation of a system, entity, phenomenon, or process (reference (a)). A model replicates a system's key feature important to a Specific Intended Use (SIU), to improve the understanding of how the system will work. A simulation is a method for implementing a model to examine how a system performs over time (reference (a)). M&S is used often throughout the acquisition life cycle to inform decision makers and manage risk. This instruction provides guidance on using M&S specific to Navy OT&E.

a. M&S and OT&E. OT&E is an important component of the overall acquisition process. It provides stakeholders with an independent assessment of system Effectiveness, Suitability, and

2. Sensitivity Analysis (Factor Excursions): Evaluate the relative effects of changes in factor inputs to the outputs of the model. If an empirical meta-model can be created, compare the factor terms to observed changes in the model outputs. Evaluate the effect size of each of the factors using SME expertise. Small input changes that have significant effects on outputs may help identify sources of error or uncertainty in results.

OPTEVFOR: Operational Test and Evaluation Force

OPTEVFOR calls for meta-model use in OPTEVFORINST 5000.1D



DEPARTMENT OF THE NAVY
OPERATIONAL TEST AND EVALUATION FORCE
7970 DIVEN STREET
NORFOLK, VIRGINIA 23505-1498

OPTEVFORINST 5000.1D
00
8 Mar 2022

OPTEVFOR INSTRUCTION 5000.1D

From: Operational Test and Evaluation Force

Subj: USE OF MODELING AND SIMULATION IN OPERATIONAL TEST

Ref: (a) DoDI 5000.61CH-1
(b) 10 U.S.C. § 2399
(c) DoDI 5000.02
(d) DoDI 5000.89
(e) SECNAVINST 5000.2F
(f) OPNAVINST 3960.15B
(g) OPNAVINST 3811.1F
(h) SECNAVINST 5200.46
(i) MIL-STD-3022 CHG-1
(j) COMOPTEVFORINST 3980.2J
(k) COMOPTEVFOR Cyber Survivability Test an

1. Purpose. This instruction provides guidance on using M
Operational Test and Evaluation (OT&E). This instructior
reviewed in its entirety.

2. Cancellation. COMOPTEVFORINST 5000.1C.

3. Scope and Applicability. This instruction is applicable
intending to use M&S in support of OT&E. This includes
used in Operational Testing (OT) reports.

4. Background. M&S is the discipline that comprises the development and/or use of models, simulations, and associated data. A model is defined as a physical, mathematical, or otherwise logical representation of a system, entity, phenomenon, or process (reference (a)). A model replicates a system's key feature important to a Specific Intended Use (SIU), to improve the understanding of how the system will work. A simulation is a method for implementing a model to examine how a system performs over time (reference (a)). M&S is used often throughout the acquisition life cycle to inform decision makers and manage risk. This instruction provides guidance on using M&S specific to Navy OT&E.

a. M&S and OT&E. OT&E is an important component of the overall acquisition process. It provides stakeholders with an independent assessment of system Effectiveness, Suitability, and

2. Sensitivity Analysis (Factor Excursions): Evaluate the relative effects of changes in factor inputs to the outputs of the model. If an empirical meta-model can be created, compare the factor terms to observed changes in the model outputs. Evaluate the effect size of each of the factors using SME expertise. Small input changes that have significant effects on outputs may help identify sources of error or uncertainty in results.

Statistical analysis methods and experimental designs should not inhibit discovery of input sensitivities in the operational space

OPTEVFOR: Operational Test and Evaluation Force

IDA publications and presentations introduce and recommend metamodeling best practices

Space Filling Designs and Metamodeling for Understanding Modeling & Simulation Behavior

Curtis Miller

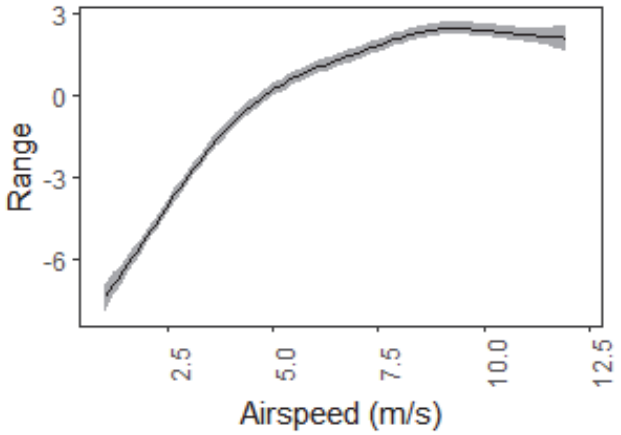
April 26, 2022

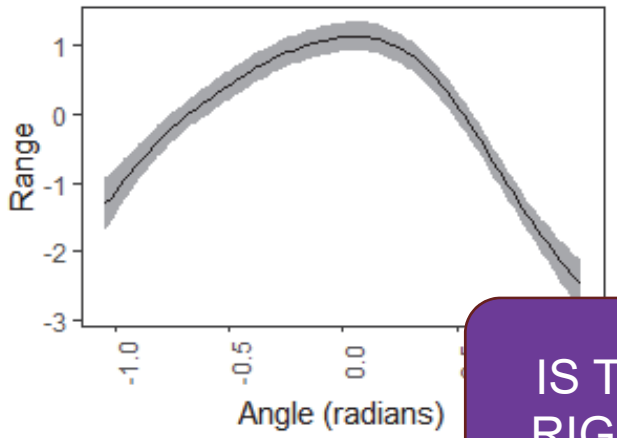
Institute for Defense Analyses
730 East Glebe Road ● Alexandria, Virginia 22305

https://dataworks.testscience.org/

Metamodels allow for qualitative assessment of simulation system performance

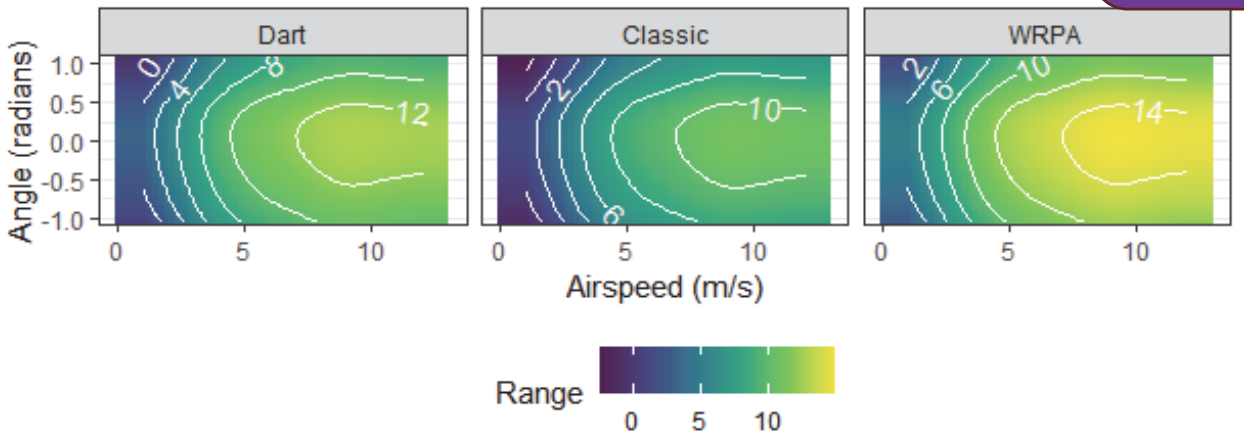
$$\text{range}_i = \beta_0 + \beta_{\text{cl}}\text{classic}_i + \beta_{\text{wr}}\text{wrpa}_i + f_{\text{as}}(\text{airspeed}_i) + f_{\text{an}}(\text{angle}_i) + \epsilon_i$$





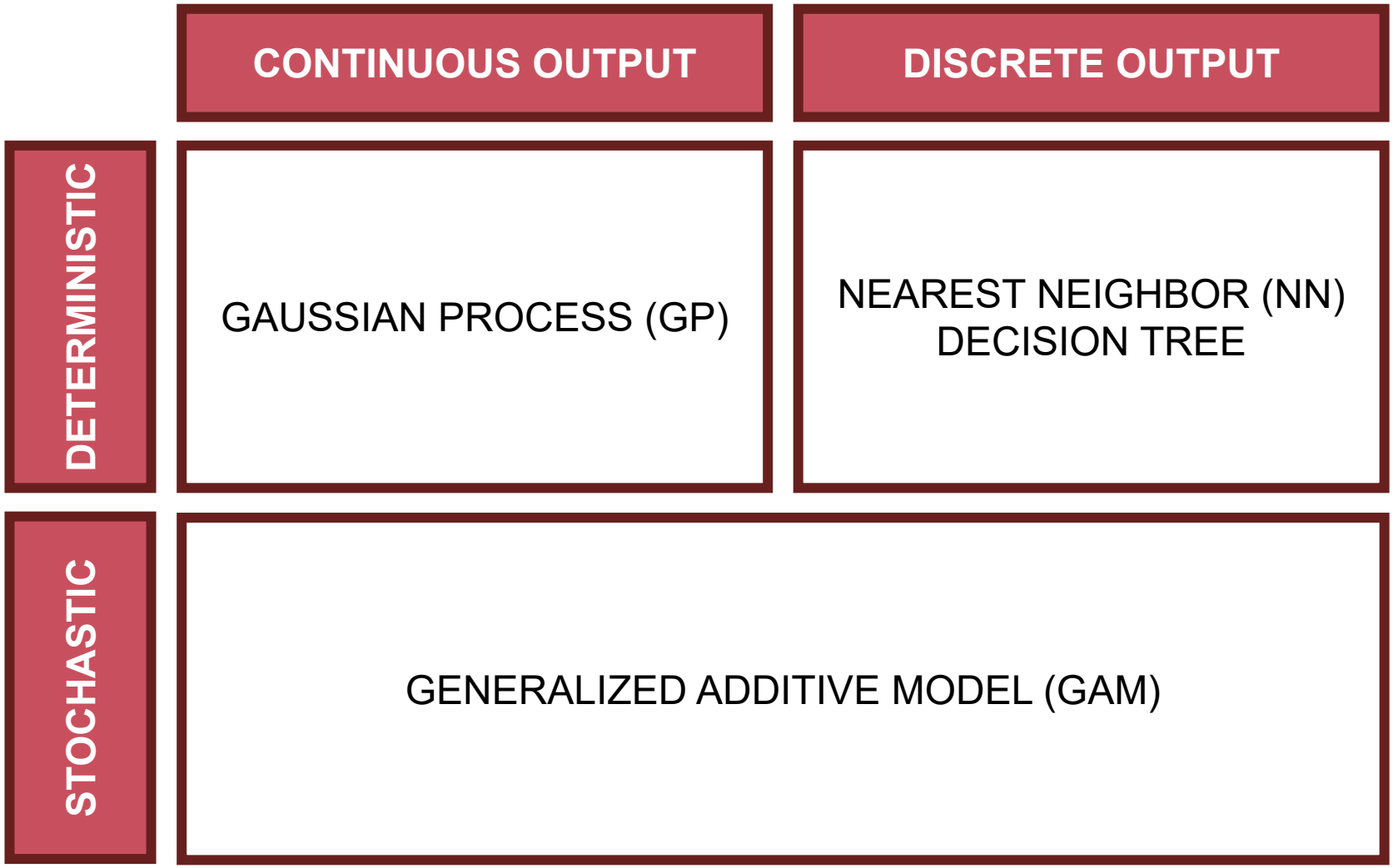
Average Range:
Dart: 9.3 m
Classic: 7.4 m
WRPA: 11.3 m

IS THIS RIGHT?



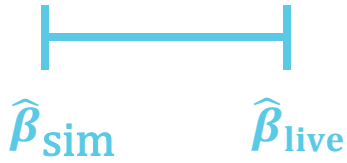
M&S: Modeling and Simulation; WRPA: World Record Paper Airplane

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

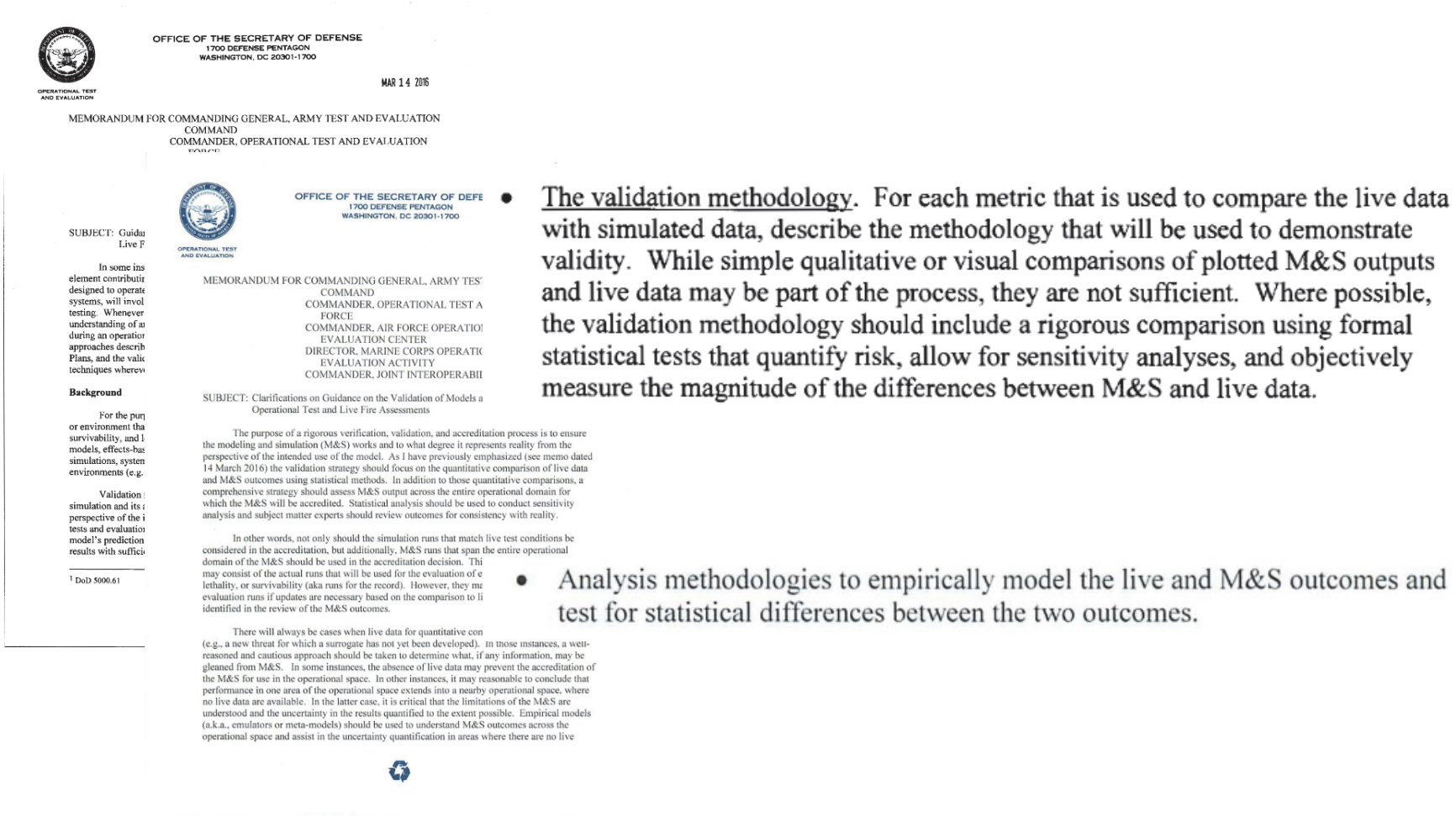


M&S: Modeling and Simulation

Do M&S predictions align with real world outcomes?



VV&A requires statistical tests for comparing live data and M&S outputs



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

OPERATIONAL TEST AND EVALUATION

MAR 14 2016

MEMORANDUM FOR COMMANDING GENERAL, ARMY TEST AND EVALUATION COMMAND
COMMANDER, OPERATIONAL TEST AND EVALUATION FORCE
COMMANDER, AIR FORCE OPERATIONAL EVALUATION CENTER
DIRECTOR, MARINE CORPS OPERATIONAL EVALUATION ACTIVITY
COMMANDER, JOINT INTEROPERABILITY COMMAND

SUBJECT: Clarifications on Guidance on the Validation of Models and Operational Test and Live Fire Assessments

The purpose of a rigorous verification, validation, and accreditation process is to ensure the modeling and simulation (M&S) works and to what degree it represents reality from the perspective of the intended use of the model. As I have previously emphasized (see memo dated 14 March 2016) the validation strategy should focus on the quantitative comparison of live data and M&S outcomes using statistical methods. In addition to those quantitative comparisons, a comprehensive strategy should assess M&S output across the entire operational domain for which the M&S will be accredited. Statistical analysis should be used to conduct sensitivity analysis and subject matter experts should review outcomes for consistency with reality.

In other words, not only should the simulation runs that match live test conditions be considered in the accreditation, but additionally, M&S runs that span the entire operational domain of the M&S should be used in the accreditation decision. This may consist of the actual runs that will be used for the evaluation of lethality, or survivability (aka runs for the record). However, they may evaluation runs if updates are necessary based on the comparison to live identified in the review of the M&S outcomes.

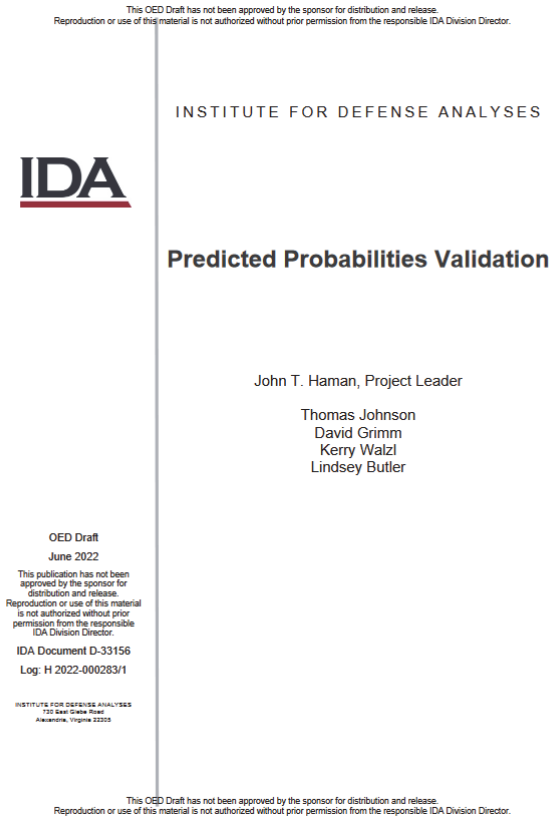
There will always be cases when live data for quantitative con (e.g., a new threat for which a surrogate has not yet been developed). In those instances, a well-reasoned and cautious approach should be taken to determine what, if any information, may be gleaned from M&S. In some instances, the absence of live data may prevent the accreditation of the M&S for use in the operational space. In other instances, it may reasonable to conclude that performance in one area of the operational space extends into a nearby operational space, where no live data are available. In the latter case, it is critical that the limitations of the M&S are understood and the uncertainty in the results quantified to the extent possible. Empirical models (a.k.a., emulators or meta-models) should be used to understand M&S outcomes across the operational space and assist in the uncertainty quantification in areas where there are no live

- The validation methodology. For each metric that is used to compare the live data with simulated data, describe the methodology that will be used to demonstrate validity. While simple qualitative or visual comparisons of plotted M&S outputs and live data may be part of the process, they are not sufficient. Where possible, the validation methodology should include a rigorous comparison using formal statistical tests that quantify risk, allow for sensitivity analyses, and objectively measure the magnitude of the differences between M&S and live data.

- Analysis methodologies to empirically model the live and M&S outcomes and test for statistical differences between the two outcomes.

M&S: Modeling and Simulation; VV&A: Verification, Validation, and Accreditation

IDA has studied additional statistical methods for problems not described in the handbook



This OED Draft has not been approved by the sponsor for distribution and release.
Reproduction or use of this material is not authorized without prior permission from the responsible IDA Division Director.

INSTITUTE FOR DEFENSE ANALYSES

IDA

Predicted Probabilities Validation

John T. Haman, Project Leader

Thomas Johnson
David Grimm
Kerry Walzl
Lindsey Butler

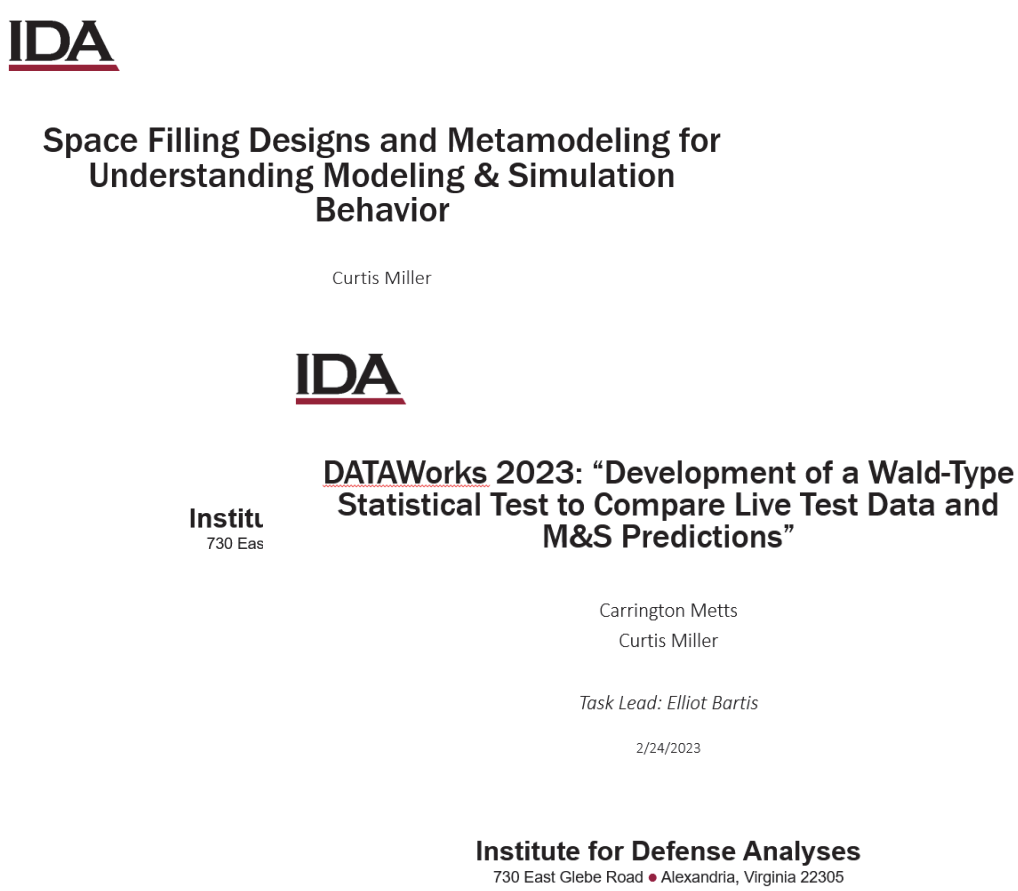
OED Draft
June 2022

This publication has not been approved by the sponsor for distribution and release. Reproduction or use of this material is not authorized without prior permission from the responsible IDA Division Director.

IDA Document D-33156
Log: H 2022-000283/1

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

This OED Draft has not been approved by the sponsor for distribution and release.
Reproduction or use of this material is not authorized without prior permission from the responsible IDA Division Director.



IDA

Space Filling Designs and Metamodeling for Understanding Modeling & Simulation Behavior

Curtis Miller

IDA

DATAWorks 2023: "Development of a Wald-Type Statistical Test to Compare Live Test Data and M&S Predictions"

Carrington Metts
Curtis Miller

Task Lead: Elliot Bartis

2/24/2023

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

IDA: Institute for Defense Analyses

What purpose does getting M&S OT VV&A right serve?

Know the enemy and know yourself; in a hundred battles you will never be in peril. When you are ignorant of the enemy but know yourself, your chances of winning or losing are equal. If ignorant both of your enemy and of yourself, you are certain in every battle to be in peril.

—Sun Tzu

Know the enemy and know yourself; in a hundred battles you will never be in peril. When you are ignorant of the enemy but know yourself, your chances of winning or losing are equal. If ignorant both of your enemy and of yourself, you are certain in every battle to be in peril.

—Sun Tzu

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
XX-05-2023	Final	START DATE	END DATE May 2023

4. TITLE AND SUBTITLE
Statistical Methods Development Work for M&S Validation

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
HQ0034-19-D-0001		
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**
BD-09-2299(90)	229990	

6. AUTHOR(S)
Miller, Curtis, G.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
Institute for Defense Analyses 730 East Glebe Road Alexandria, Virginia 22305	NS D-33460 H 2023-000116

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER
Director, Operational Test and Evaluation 1700 Defense Pentagon Washington, DC 20301		

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
We discuss four areas in which statistically rigorous methods contribute to modeling and simulation validation studies. These areas are statistical risk analysis, space-filling experimental designs, metamodel construction, and statistical validation. Taken together, these areas implement DOT&E guidance on model validation. In each area, IDA has contributed either research methods, user-friendly tools, or both. We point to our tools on testscience.org, and survey the research methods that we've contributed to the M&S validation literature.

This presentation was given at the PEO IWS M&S in T&E Workshop on May 4, 2023.

15. SUBJECT TERMS
Modeling and Simulation (M&S) Validation; statistical methods; Design of Experiments (DOE); Operational Testing; Director of Operational Test and Evaluation (DOT&E)

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	c. THIS PAGE		
Unclassified	Unclassified	Unclassified	SAR	

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER
John Haman	703-845-2132

INSTRUCTIONS FOR COMPLETING SF 298

1. REPORT DATE.
Full publication date, including day, month, if available. Must cite at least the year and be Year 2000 compliant, e.g. 30-06-1998; xx-06-1998; xx-xx-1998.

2. REPORT TYPE.
State the type of report, such as final, technical, interim, memorandum, master's thesis, progress, quarterly, research, special, group study, etc.

3. DATES COVERED.
Indicate the time during which the work was performed and the report was written.

4. TITLE.
Enter title and subtitle with volume number and part number, if applicable. On classified documents, enter the title classification in parentheses.

5a. CONTRACT NUMBER.
Enter all contract numbers as they appear in the report, e.g. F33615-86-C-5169.

5b. GRANT NUMBER.
Enter all grant numbers as they appear in the report, e.g. AFOSR-82-1234.

5c. PROGRAM ELEMENT NUMBER.
Enter all program element numbers as they appear in the report, e.g. 61101A

5d. PROJECT NUMBER.
Enter all project numbers as they appear in the report, e.g. 1F665702D1257; ILIR.

5e. TASK NUMBER.
Enter all task numbers as they appear in the report, e.g. 05; RF0330201; T4112.

5f. WORK UNIT NUMBER.
Enter all work unit numbers as they appear in the report, e.g. 001; AFAPL30480105.

6. AUTHOR(S).
Enter name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. The form of entry is the last name, first name, middle initial, and additional qualifiers separated by commas, e.g. Smith, Richard, J, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES). Self-explanatory.

8. PERFORMING ORGANIZATION REPORT NUMBER. Enter all unique alphanumeric report numbers assigned by the performing organization, e.g. BRL-1234; AFWL-TR-85-4017-Vol-21-PT-2.

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES). Enter the name and address of the organization(s) financially responsible for and monitoring the work.

10. SPONSOR/MONITOR'S ACRONYM(S).
Enter, if available, e.g. BRL, ARDEC, NADC.

11. SPONSOR/MONITOR'S REPORT NUMBER(S).
Enter report number as assigned by the sponsoring/monitoring agency, if available, e.g. BRL-TR-829; -215.

12. DISTRIBUTION/AVAILABILITY STATEMENT.
Use agency-mandated availability statements to indicate the public availability or distribution limitations of the report. If additional limitations/ restrictions or special markings are indicated, follow agency authorization procedures, e.g. RD/FRD, PROPIN, ITAR, etc. Include copyright information.

13. SUPPLEMENTARY NOTES. Enter information not included elsewhere such as: prepared in cooperation with; translation of; report supersedes; old edition number, etc.

14. ABSTRACT. A brief (approximately 200 words) factual summary of the most significant information.

15. SUBJECT TERMS. Key words or phrases identifying major concepts in the report.

16. SECURITY CLASSIFICATION. Enter security classification in accordance with security classification regulations, e.g. U, C, S, etc. If this form contains classified information, stamp classification level on the top and bottom of this page.

17. LIMITATION OF ABSTRACT. This block must be completed to assign a distribution limitation to the abstract. Enter UU (Unclassified Unlimited) or SAR (Same as Report). An entry in this block is necessary if the abstract is to be limited.