INSTITUTE FOR DEFENSE ANALYSES

Statistical Techniques for Modeling and Simulation Validation

Laura J. Freeman, *Project Leader*
Kelly M. Avery

September 2017

IDA NS D-8694

Log: H 2017-000493

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

About This Publication

In the world of DoD test and evaluation, collecting sufficient data to evaluate system performance against operationally realistic threats is often not possible due to cost and resource restrictions, safety concerns, or lack of adequate or representative threats. Thus, modeling and simulation (M&S) tools are frequently used to augment live testing in order to facilitate a more complete evaluation of performance. When M&S is used as part of an operational evaluation, the M&S capability should first be rigorously validated to ensure it is representing the real world adequately enough for the intended use. Specifically, the usefulness and limitations of the M&S should be well characterized, and uncertainty quantified to the extent possible. Many statistical techniques are available to rigorously compare M&S output with live test data. This document will describe some of these methodologies and present recommendations for a variety of data types and sizes. We will show how design for computer experiments can be used to efficiently cover the simulation domain and inform live testing. Experimental design and corresponding statistical analysis techniques for comparing live and simulated data will be discussed and compared. A simulation study shows that regression analysis is the most powerful comparison when experimental design techniques are used, while more robust non-parametric techniques provide widely applicable solutions for the comparison.

Acknowledgments

Technical review was performed by Dean Thomas and William G. Gardner from the Operational Evaluation Division.

For more information:
Laura J. Freeman, Project Leader
lfreeman@ida.org • (703) 845-2084

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

INSTITUTE FOR DEFENSE ANALYSES

IDA NS D-8694

Statistical Techniques for Modeling and Simulation Validation

Laura J. Freeman, *Project Leader*
Kelly M. Avery

This page intentionally left blank.

Executive Summary

In Defense system test and evaluation, collecting sufficient data to evaluate system performance against operationally realistic threats is often not possible due to cost and resource restrictions, safety concerns, or lack of adequate or representative threats. Thus, modeling and simulation (M&S) tools frequently are used to augment live testing in order to facilitate a more complete evaluation of performance.

When relying on M&S to support an operational evaluation, one should thoroughly understand how well the M&S represents the system of interest across all conditions in which the system will be used. The process of determining the degree to which a model and its associated data provide an accurate representation of the real world from the perspective of the intended uses of the model is known as validation.

Any number of techniques can be used to support the validation of a model, but a portion of any validation should include a quantitative comparison of the simulation output to available live data. Such an analysis provides an objective statistical measure of the differences between the M&S and the live test, quantifies the uncertainty in the model, and can identify areas of high risk that might require additional testing.

Design of experiments techniques can be used to efficiently cover the factor space of interest and gather the appropriate data from both the computer simulation environment and the live environment to support a meaningful comparison.

This briefing provides an overview of some statistical design and analysis methods that can help support the characterization of the accuracy of the model by providing quantitative comparisons of M&S data to reference data gathered from live testing. The briefing also describes a simulation study used to obtain recommendations for which techniques to use in different situations. For example, if a designed experiment was executed, more advanced and sensitive statistical methods such as regression analysis should be considered.

This page intentionally left blank.

Statistical Techniques for Modeling and Simulation Validation

Dr. Kelly Avery

Dr. Laura Freeman

Institute for Defense Analyses

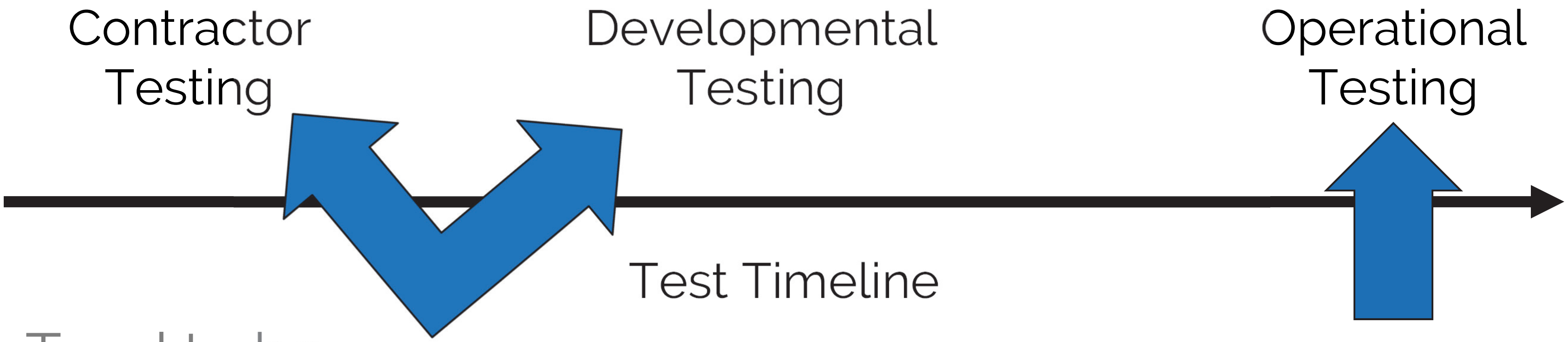
Simulation Innovations Workshop 2017

This page intentionally left blank.

Operational Testing

This page intentionally left blank.

DoD Test Paradigm



Contractor Testing

Developmental Testing

Operational Testing

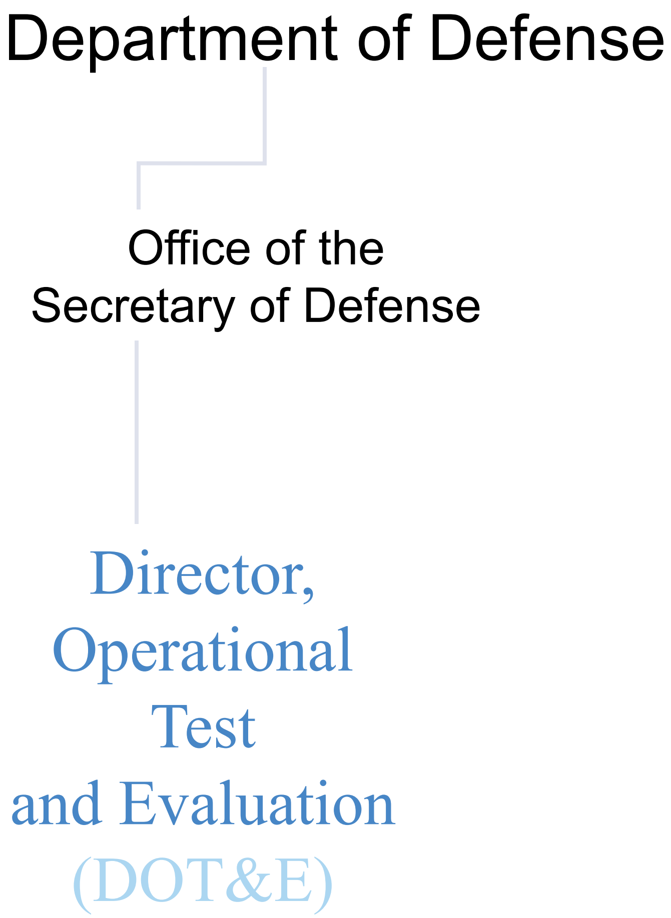
Test Timeline

Tend to be requirements driven

Focuses on mission success

Requirements documents are often missing important mission considerations

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

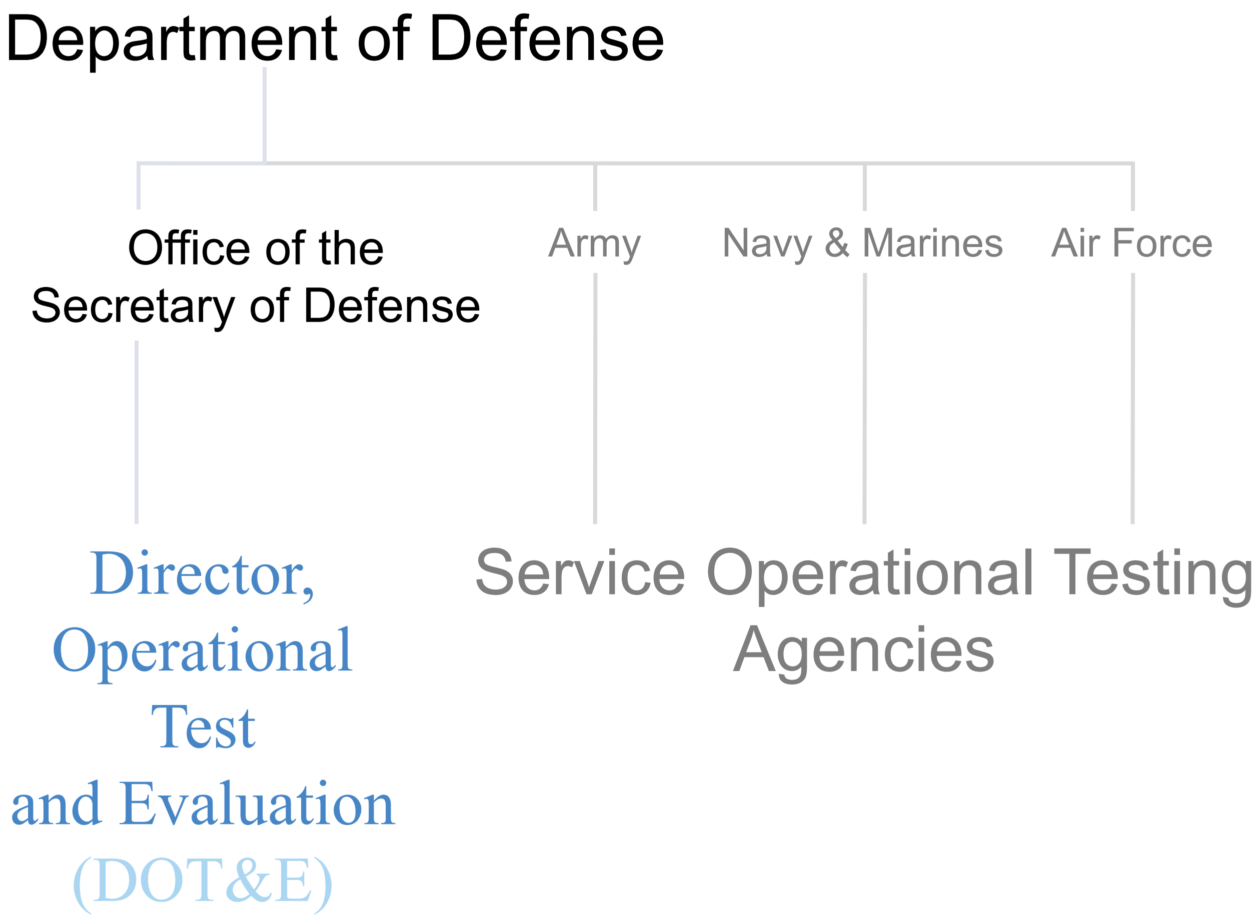


Department of Defense

Office of the
Secretary of Defense

Director,
Operational
Test
and Evaluation
(DOT&E)

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



Department of Defense

Office of the Secretary of Defense

Army

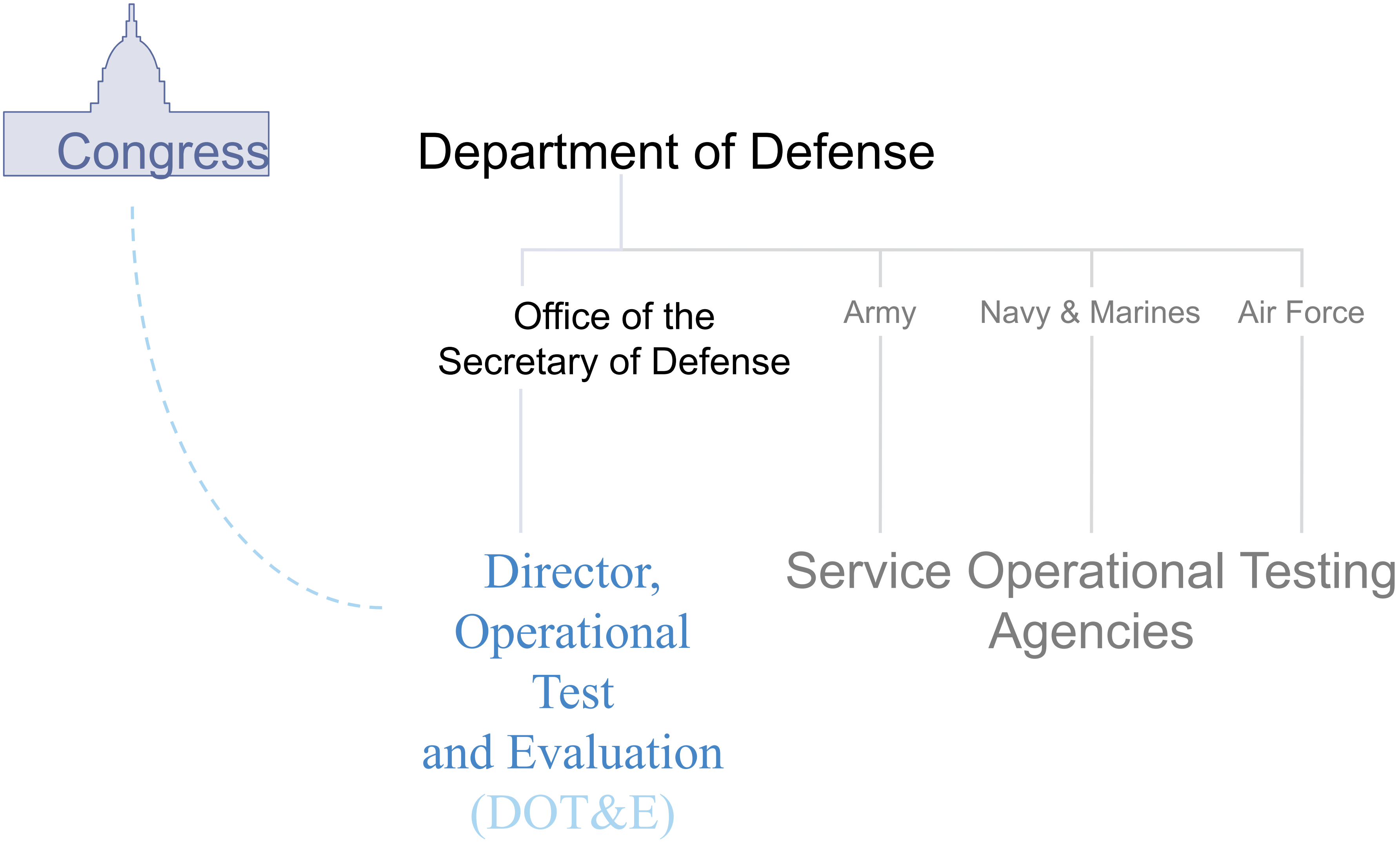
Navy & Marines

Air Force

Director, Operational Test and Evaluation (DOT&E)

Service Operational Testing Agencies

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



Congress

Department of Defense

Office of the Secretary of Defense

Army

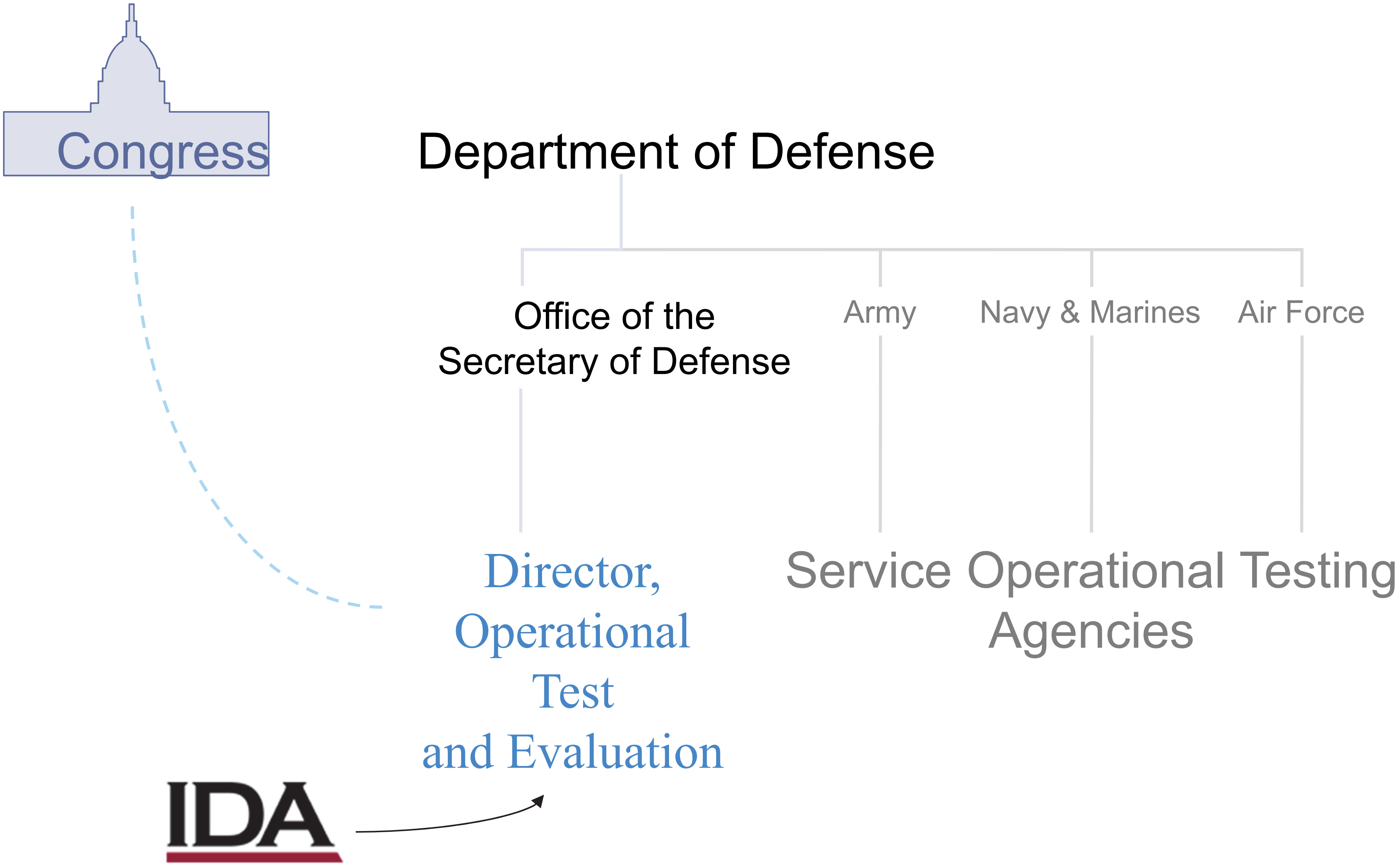
Navy & Marines

Air Force

Director, Operational Test and Evaluation (DOT&E)

Service Operational Testing Agencies

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



Congress

Department of Defense

Office of the Secretary of Defense

Army

Navy & Marines

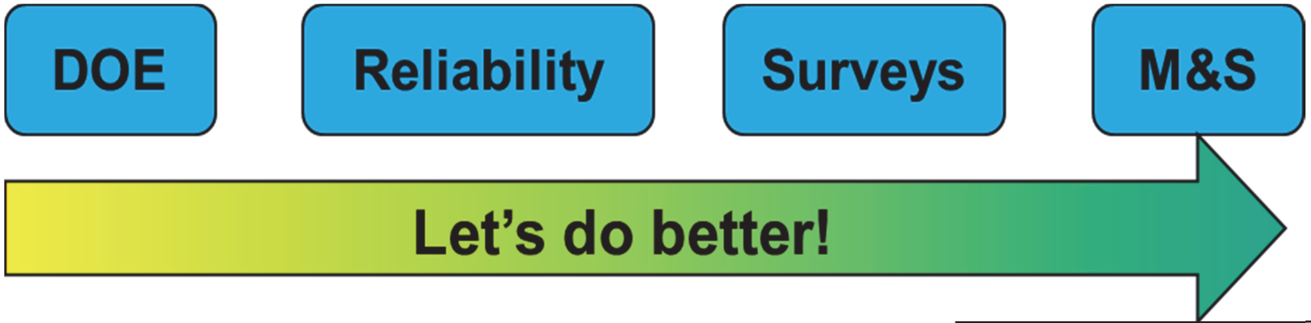
Air Force

Director, Operational Test and Evaluation

Service Operational Testing Agencies

IDA

DOT&E Sets Policy and Guidance for Conducting Operational Testing



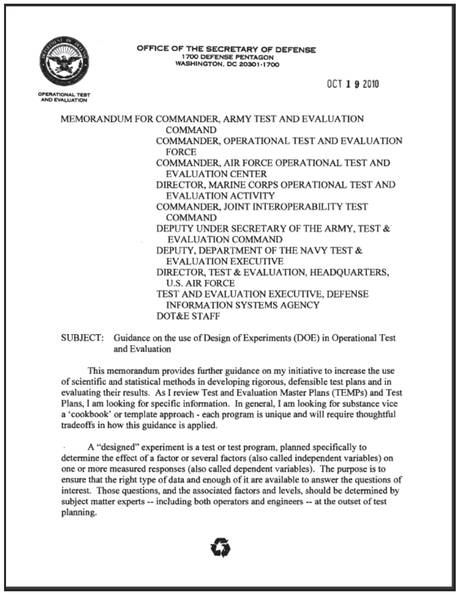
DOE

Reliability

Surveys

M&S

Let's do better!



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

OPERATIONAL TEST
AND EVALUATION

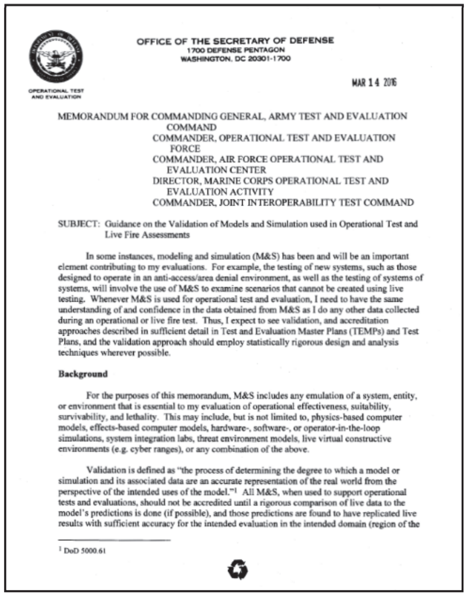
OCT 1 9 2010

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

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

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

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



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

OPERATIONAL TEST
AND EVALUATION

MAR 1 4 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 Models and Simulation used in Operational Test and Live Fire Assessments

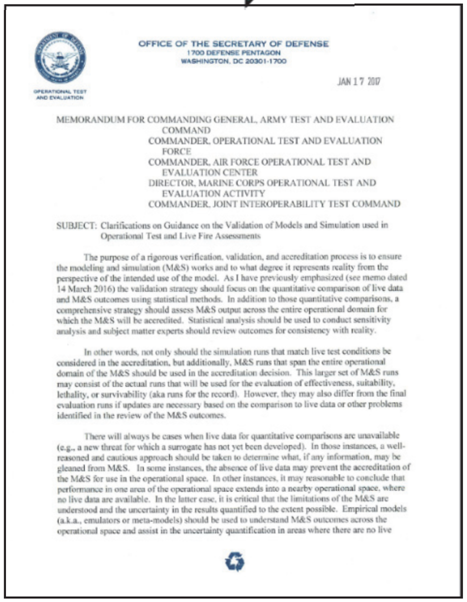
In some instances, modeling and simulation (M&S) has been and will be an important element contributing to my evaluations. For example, the testing of new systems, such as those designed to operate in an anti-access/area denial environment, as well as the testing of systems of systems, will involve the use of M&S to examine scenarios that cannot be created using live testing. Whenever M&S is used for operational test and evaluation, I need to have the same understanding of and confidence in the data obtained from M&S as I do any other data collected during an operational or live fire test. Thus, I expect to see validation, and accreditation approaches described in sufficient detail in Test and Evaluation Master Plans (TEMPs) and 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



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

OPERATIONAL TEST
AND EVALUATION

JAN 1 7 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 AND 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 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.

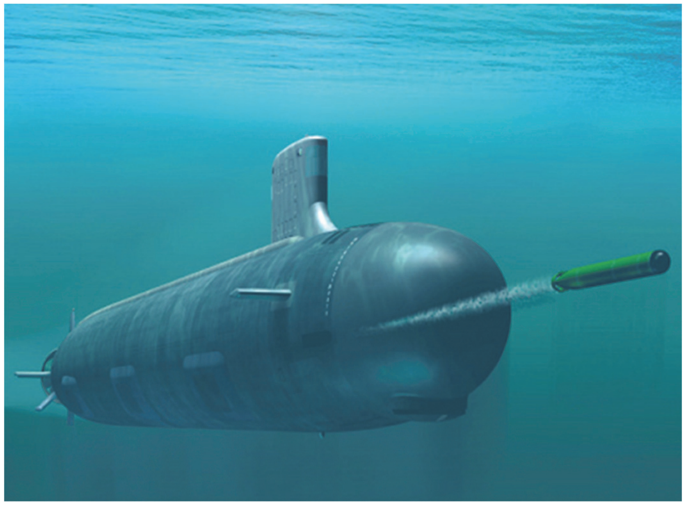
In other words, not only should the simulation runs that match live test conditions be considered in the accreditation, but additionally, M&S runs that span the entire operational domain of the M&S should be used in the accreditation decision. This larger set of M&S runs may consist of the actual runs 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 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

Striving to increase the rigor of DoD Test and Evaluation

Uses for M&S in OT

Supplement or augment live test data when experiments are cost and/or safety prohibitive



Examine threats incapable of being reproduced for testing

Characterize rare events or threats

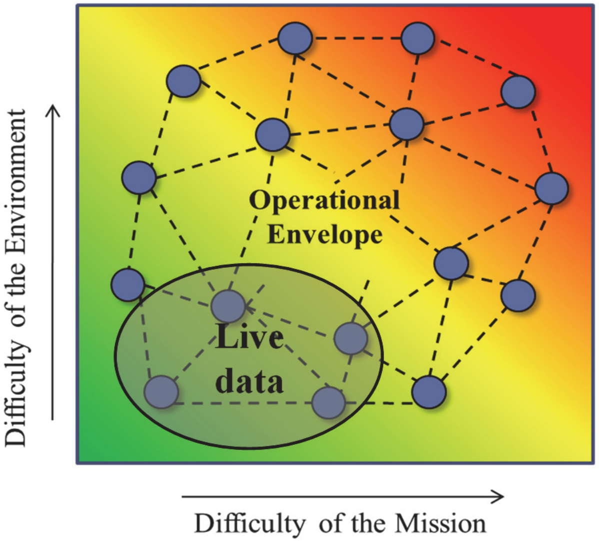
Allow for end-to-end mission evaluation

Inform experimental design decisions

M&S can never fully replace testing in the true operational environment (open air, at sea, etc.)

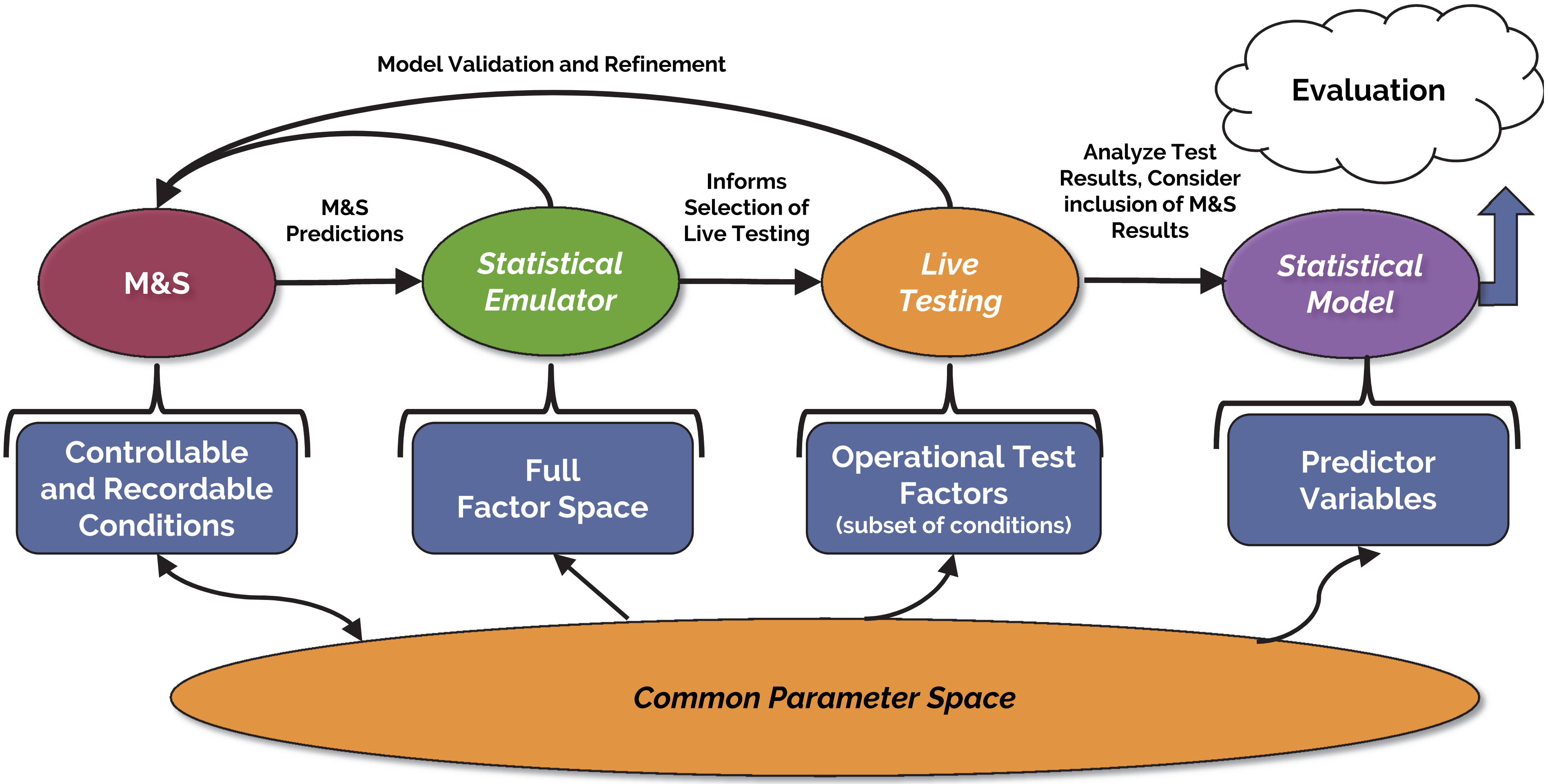
DOT&E has been encouraging statistically-based validation techniques in addition to traditional techniques

Apply design for computer experiments principles to create statistical emulators to assess M&S output across the entire operational domain



Employ DOE and formal statistical analysis techniques to compare live and M&S data

Generic Framework



Identify the common set of variables that spans the operational space

This page intentionally left blank.

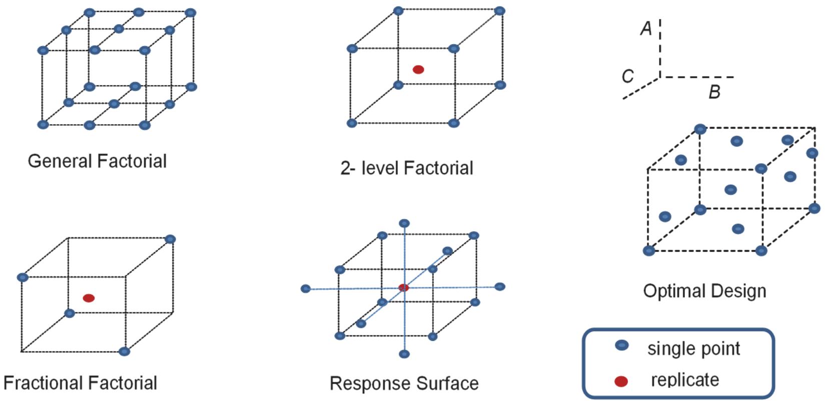
Design of Experiments

This page intentionally left blank.

DOE can be used to efficiently cover large input spaces

Provides a scientific, structured, objective test methodology answering the key questions of test:

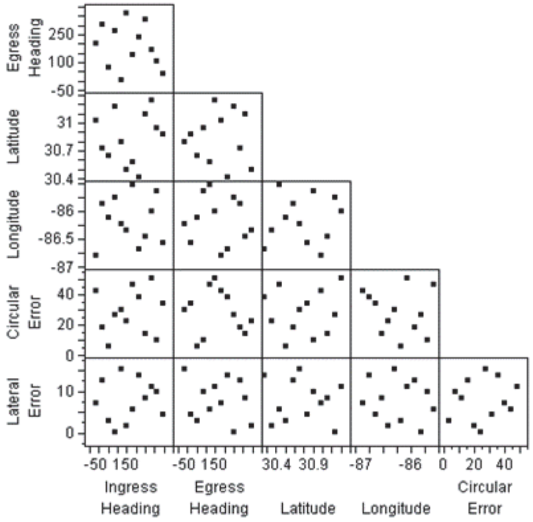
- How many points?
- Which points?
- In what order?
- How to analyze?



Two broad classes:

- **Classical** – geared towards stochastic test outcomes
- **Computer** – assume a deterministic outcome

A	B	C	D	E	F	G	H	I	J
0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	1	1	1	1	1
0	1	1	1	0	0	0	1	1	1
1	0	1	1	0	1	1	0	0	1
1	1	0	1	1	0	1	0	1	0
1	1	1	0	1	1	0	1	0	0



Use DOE to ensure sufficient data for validation

Could require a combination of computer experiments and classical designs, i.e.

- Space filling to efficiently cover the full M&S space
- Classical designs to select limited live tests from possible live shots

Facilitates the creation of statistical emulators to support

- Sensitivity analysis
- Quick model characterization
- Quantifying uncertainty
- Identifying risk areas
- Planning future live testing

Choice of design informs validation techniques

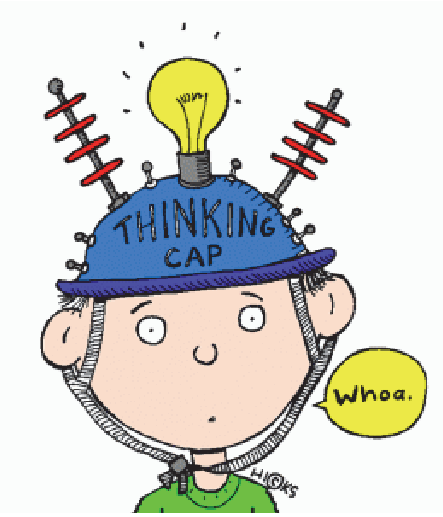
Goal: Maximize the probability that we identify bugs in the M&S

Statistical Techniques for Comparing Live and Simulation Data

This page intentionally left blank.

Goal of the Statistical Comparison (in English)

Do the simulation data and the live experimental data agree?



What uncertainty is there in the simulation data?
If they don't agree, can we identify the specific conditions where they disagree?
Are the differences between M&S and experimental data important?

Goal of the Statistical Comparison (in stats speak)

H_0: Simulation output matches the live data

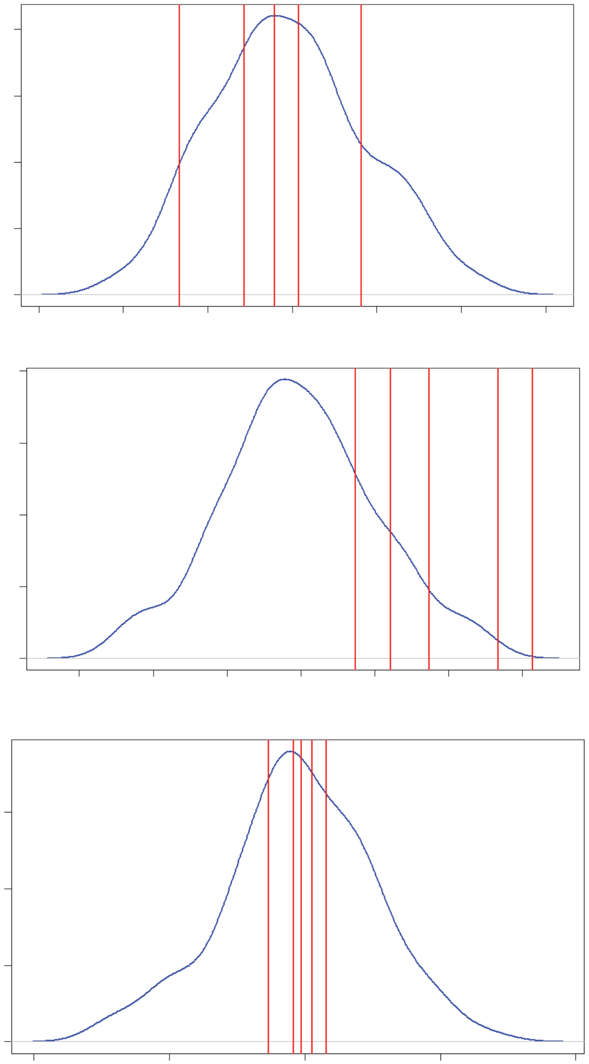
H_1: Simulation output does not match the live data

"Matching" can be in terms of a variety of parameters, including the means, the variances, and the distributions

Goal is to maximize power given a specified confidence level

Higher power and confidence translates into less uncertainty about the difference between live and sim

Can assess the statistical performance of various techniques using Monte Carlo simulation



Statistical Test Options (not exhaustive!)

Parametric Tests

- t-test (or log t-test)
- Kolmogorov-Smirnov Test

Non-parametric Tests

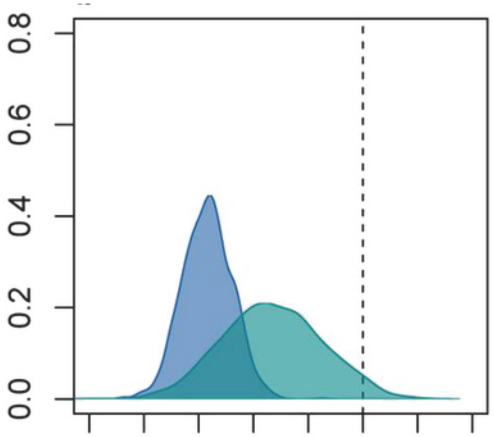
- Kolmogorov-Smirnov Test
- Fisher's Combined Probability Test
- Wilcoxon Rank Sum Test
- Fisher's Exact Test

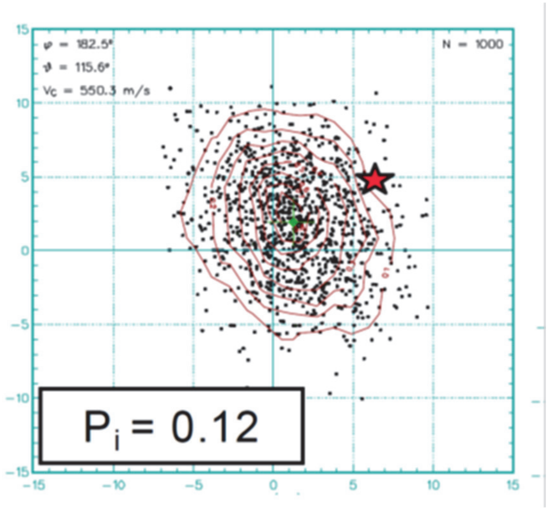
Regression Testing

- Multiple Regression (Linear, lognormal, or logistic)
- Emulation and Prediction

Considerations

- Ignore factors or take factors into account?
- Are a combination of techniques necessary in some cases?





Preview of Recommendations

The properties of your observed data and the structure of your factors dictates which method is best

Multiple statistical techniques can be used to check for various types of differences between live and sim

General classes of comparison methods that tend to work well:

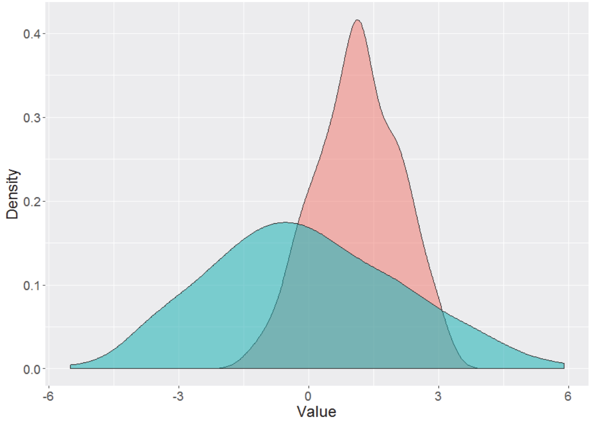
- Non-parametric Kolmogorov-Smirnov test and Fisher's Combined Probability Test
  - Work well for distribution comparisons
- Regression analysis (to include variations like logistic and lognormal) with indicator variable for live/sim
  - Works best for matched designed experiments
- Statistical emulation and prediction
  - Works well for lots of M&S data and limited live data

Recommendations determined via Monte Carlo power simulations

Kolmogorov-Smirnov (K-S) Test

Compare the distribution of live data to the distribution of M&S data

- The K-S test calculates the maximum distance between two CDFs



Parametric: Compare each of the data sets (live and sim) to a *reference distribution* (e.g. normal)

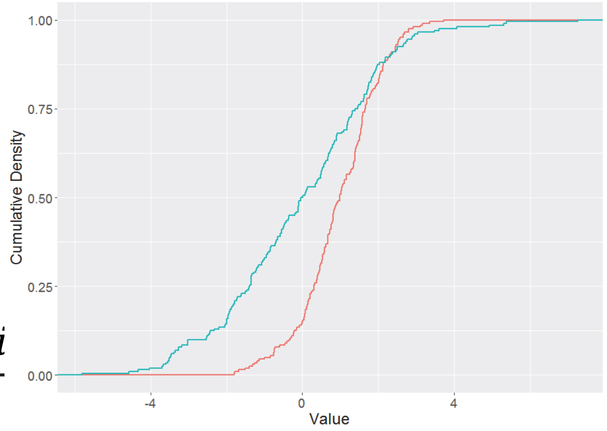
Non-parametric: Compare each of the data sets (live and sim) to *each other*

Works better for our problem

Scaling the data first can account for different conditions

- For each distinct condition:

$$Scaled\ data = \frac{each\ individual\ data\ point\ - mean\ (all\ data\ in\ that\ conditi}{stan\ dev\ (all\ data\ in\ that\ condition)}$$



Note: All data are notional

Fisher’s Combined Probability Test

Compares distributions of continuous data

- Simulation “cloud” vs. 1 or more live shots per condition
- Nonparametric

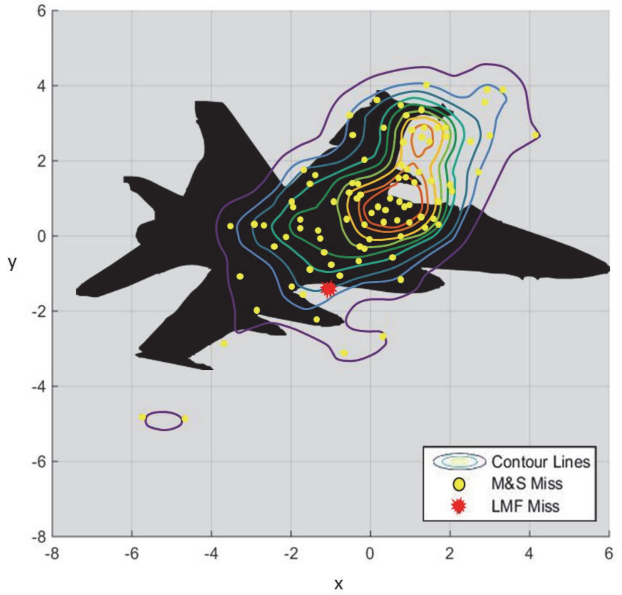
p-values can be calculated in a variety of ways

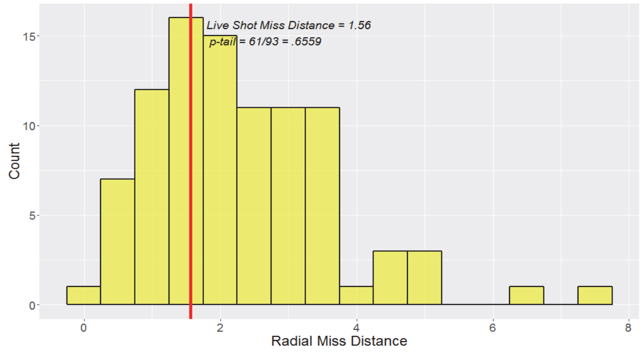
- 2 dimensionally using contours
- 1 dimensionally using miss distance quantiles

Use a goodness-of-fit procedure to check for overall uniformity of the p-values

- Fisher’s Combined probability test: X = -2 Σ ln(p) follows a chi-square distribution with 2N degrees of freedom
  - Sensitive to one failed test condition
- Kolmogorov-Smirnov test: compares observed p-values to a true uniform distribution

No formal test of factor effects





Note: All data are notional

Regression Modeling: Parameterizing Live vs. Sim

Pool live and M&S data and build a statistical model

- Include an indicator term that indicates whether the data point comes from live or M&S (*test type*), as well as interaction terms between *test type* and other factors of interest
- For example,

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

- If the *Test Type* effect is statistically significant, then the M&S runs are not providing data that are consistent with the live runs
- If the interaction term is significant, there many be a problem with the simulation under some conditions but not others

The type of regression depends on the nature of the observed data

- Symmetric – use linear regression
- Skewed – use lognormal regression
- Binary – use logistic regression

Method works best if you used a designed experiment for both live and sim

- Must compute interaction terms to avoid rolling up results
- Strength is detecting differences in means

Monte Carlo Power Simulation

Risk analysis technique

- Model possible outcomes by randomly drawing from a probability distribution and calculating results over and over, each time using a difference set of random values

Result of interest in this case is whether or not a statistical test rejects the null hypothesis, given the alternative is true

- Doing this many many times and calculating the proportion of times the test succeeds is an estimate of statistical power!

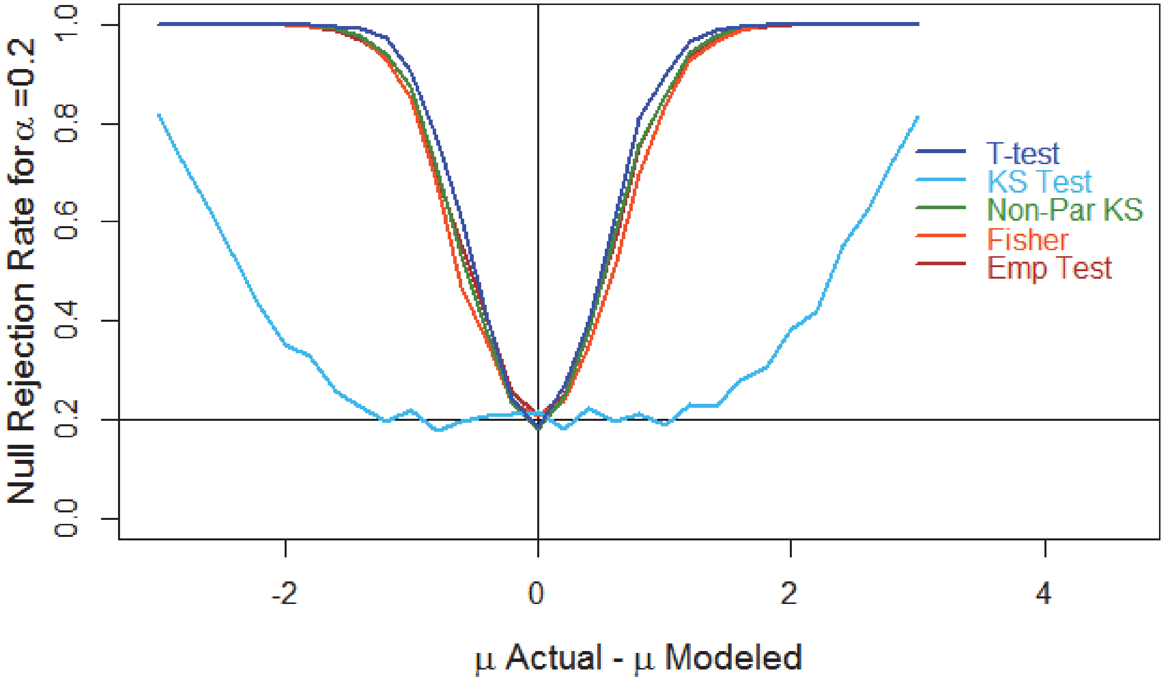
Simulation conditions:

- Distribution of response variable (symmetric, skewed, binary)
- Sample size of live test (small, medium, large*)
- Structure of factors (univariate, distributed level effects, designed experiment)
- Effect size of interest (range of differences in means and variance ratios)

Running the simulation for each statistical test over all conditions of interest and comparing their power allows us to make informed recommendations

* For symmetric and skewed data: Small = 2-5, Moderate = 5-10, Large = 11-20; For binary data: Small = 20, Moderate = 40, Large = 100

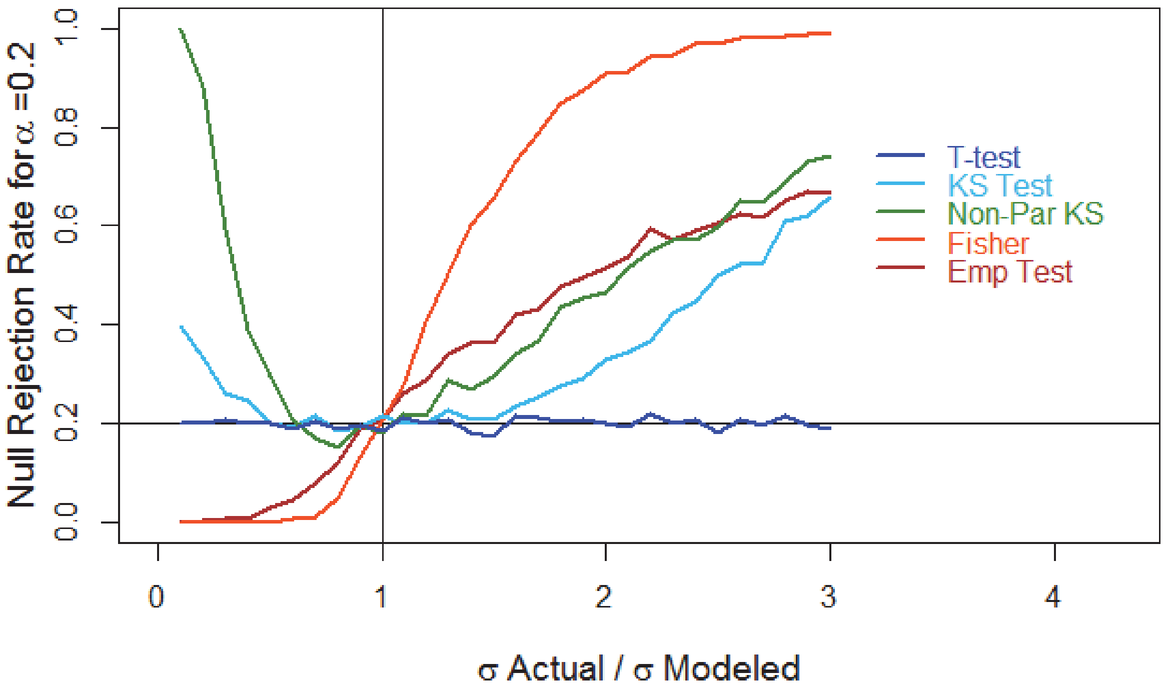
Power Curves for Various Tests:
Detecting mean changes in univariate data



T-test

Symmetric

Power Curves for Various Tests: Detecting variance changes in univariate data

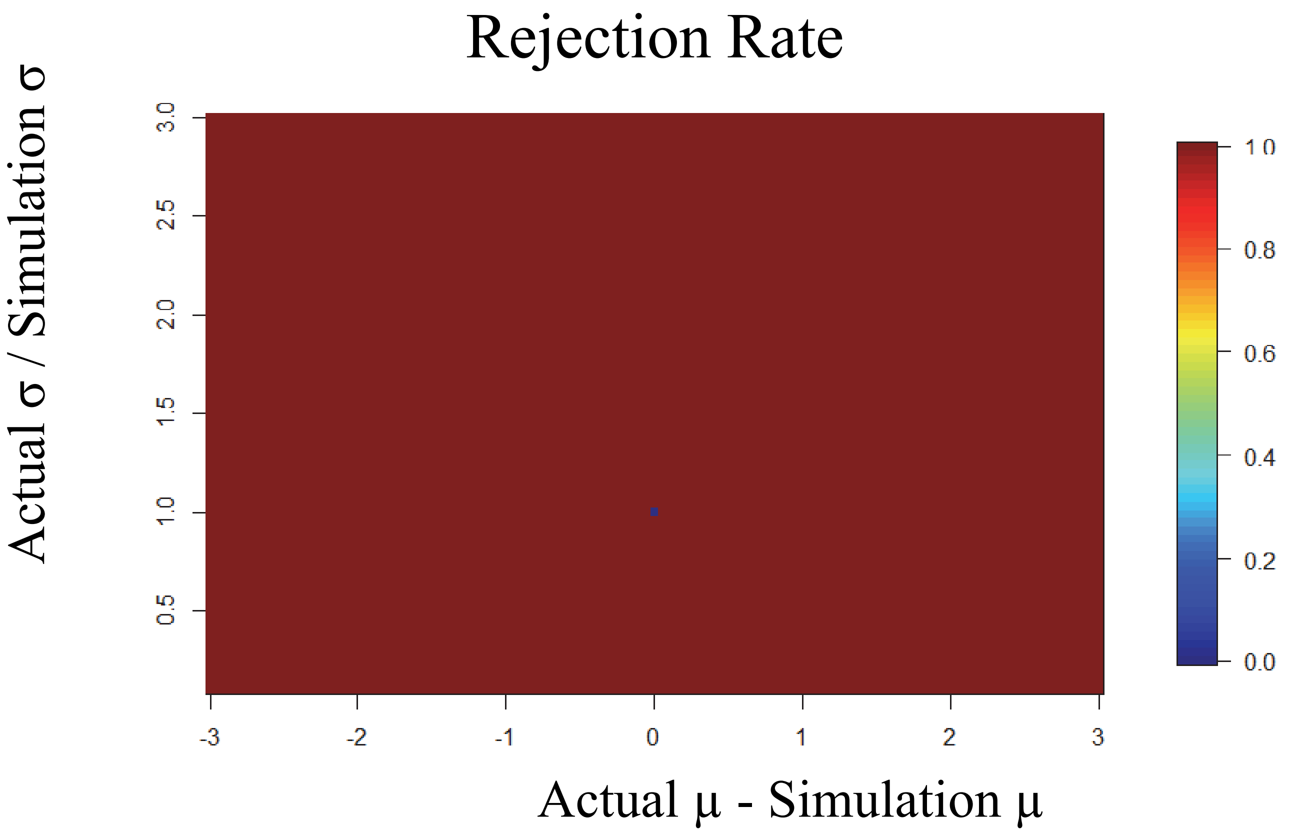


Fisher

Non-Par KS

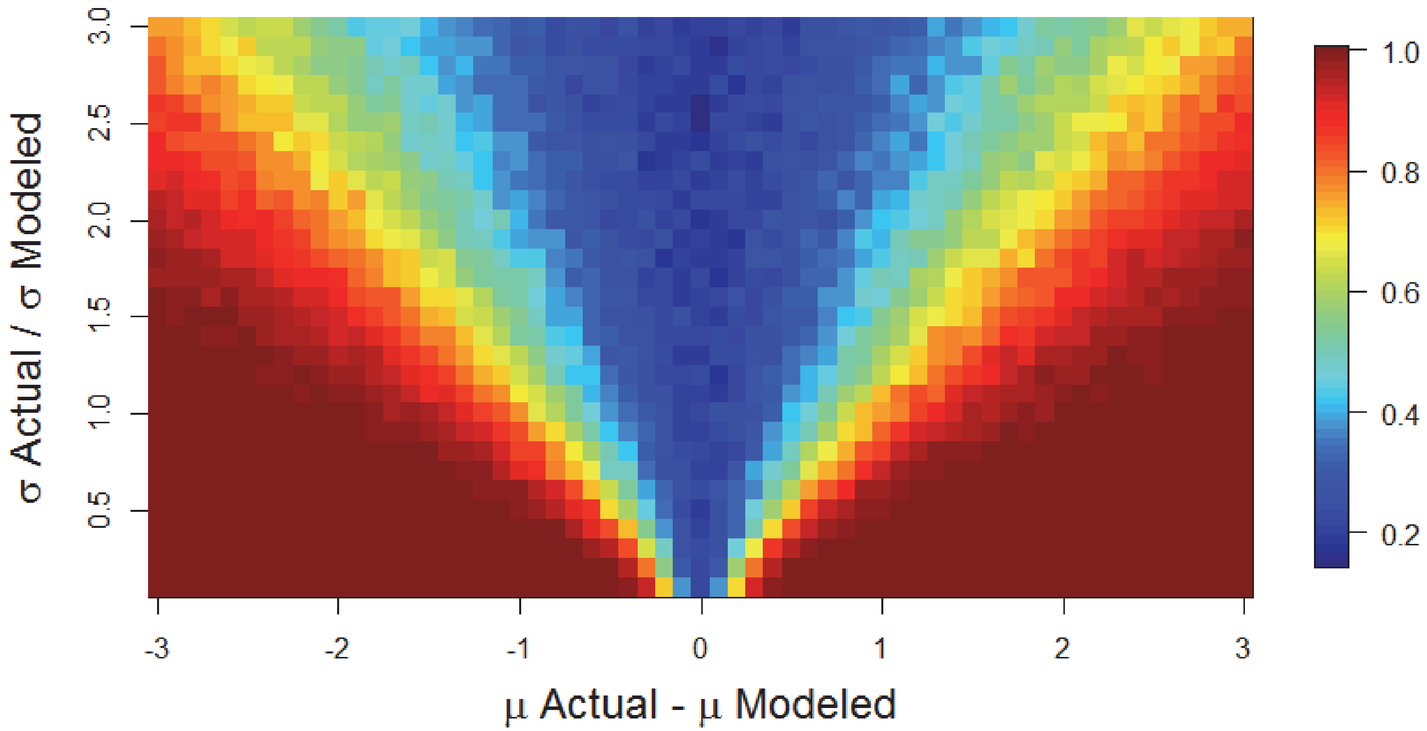
Symmetric

Power Heat Map: Ideal Test



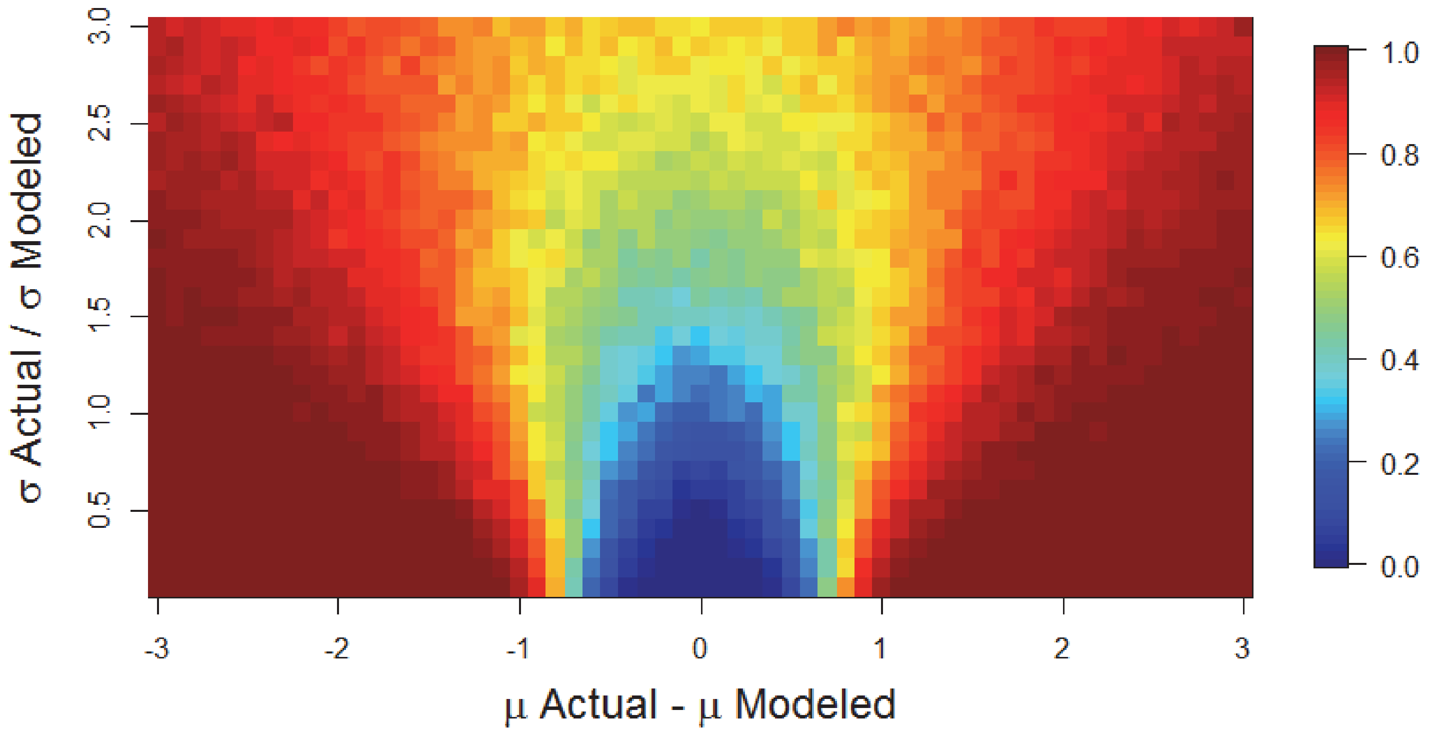
Notice the blue dot at no change in mean and variance!

Power Heat Map: T-test



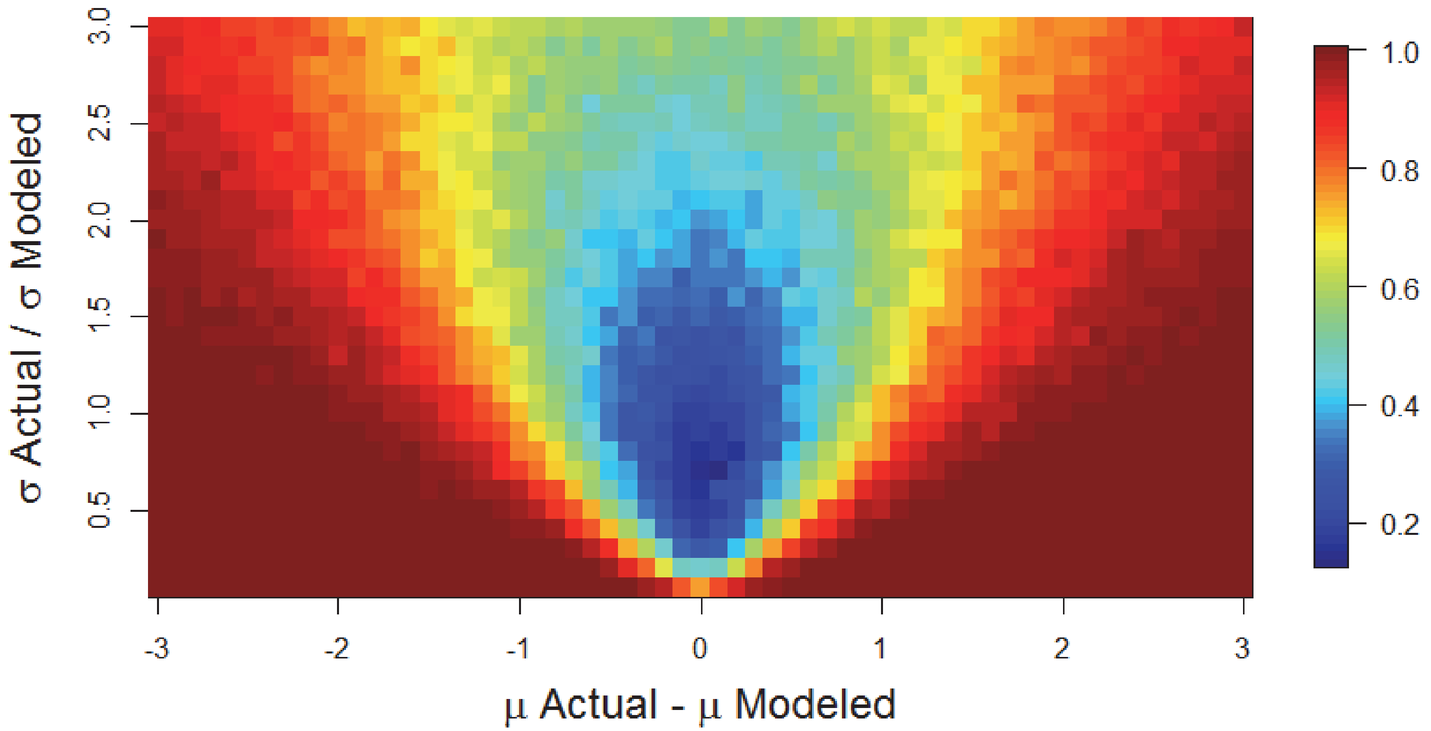
Symmetric

Power Heat Map: Fisher’s Test



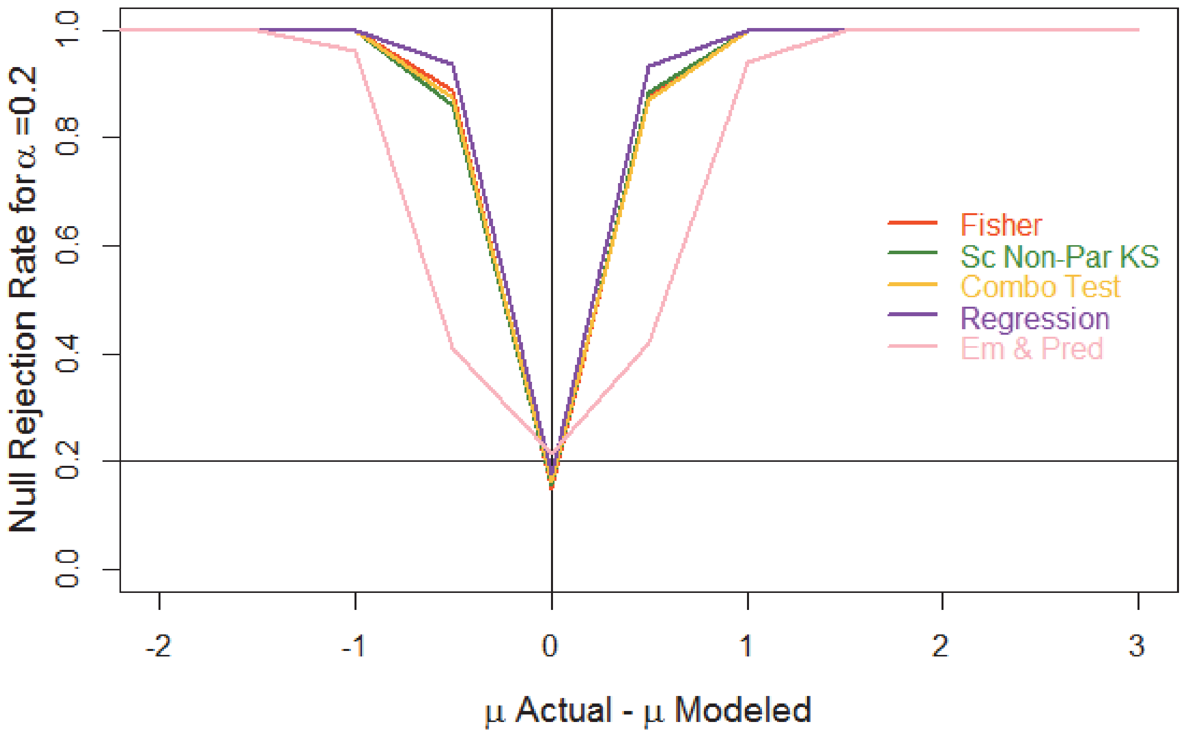
Symmetric

Power Heat Map: Kolmogorov-Smirnov Test



Symmetric

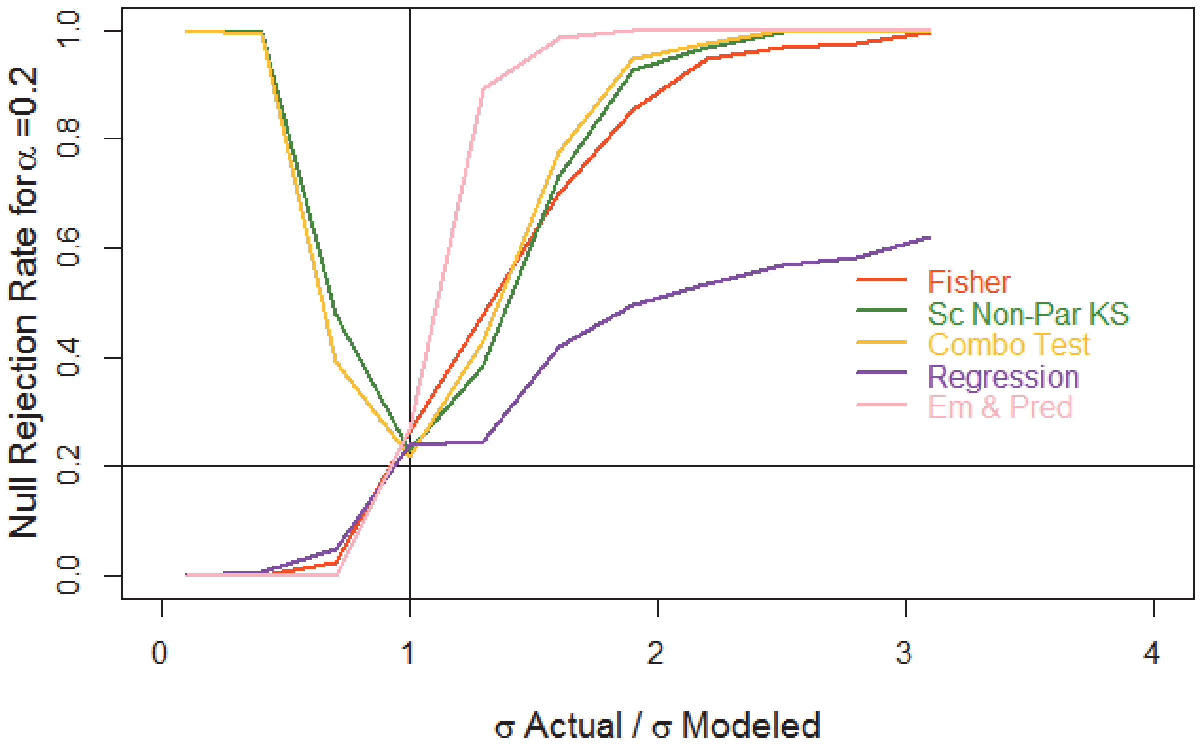
Power Curves for Various Tests:
Detecting mean changes in a designed experiment



Regression

Symmetric

Power Curves for Various Tests:
Detecting variance changes in a designed experiment

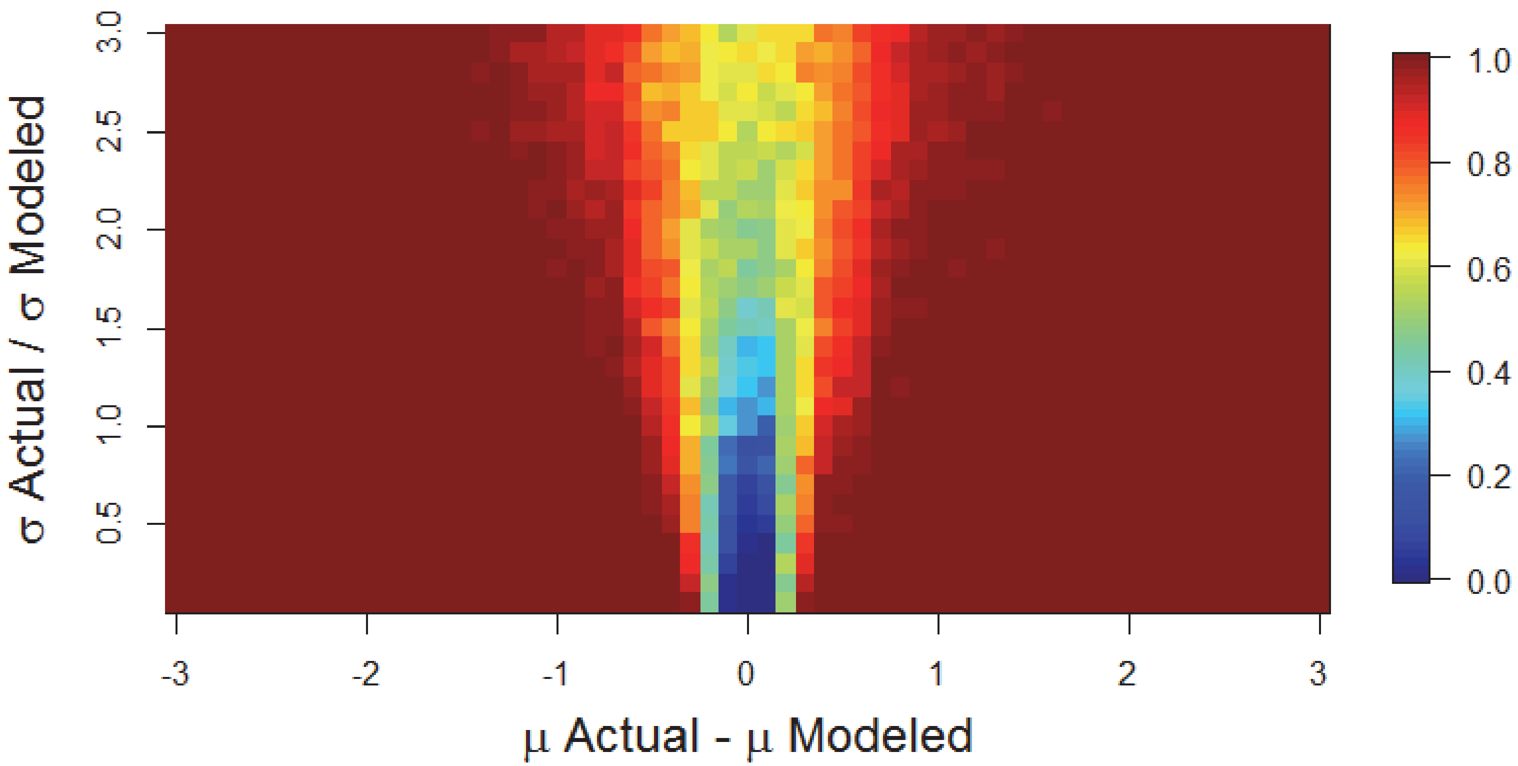


Sc Non-Par Ks

Em & Pred

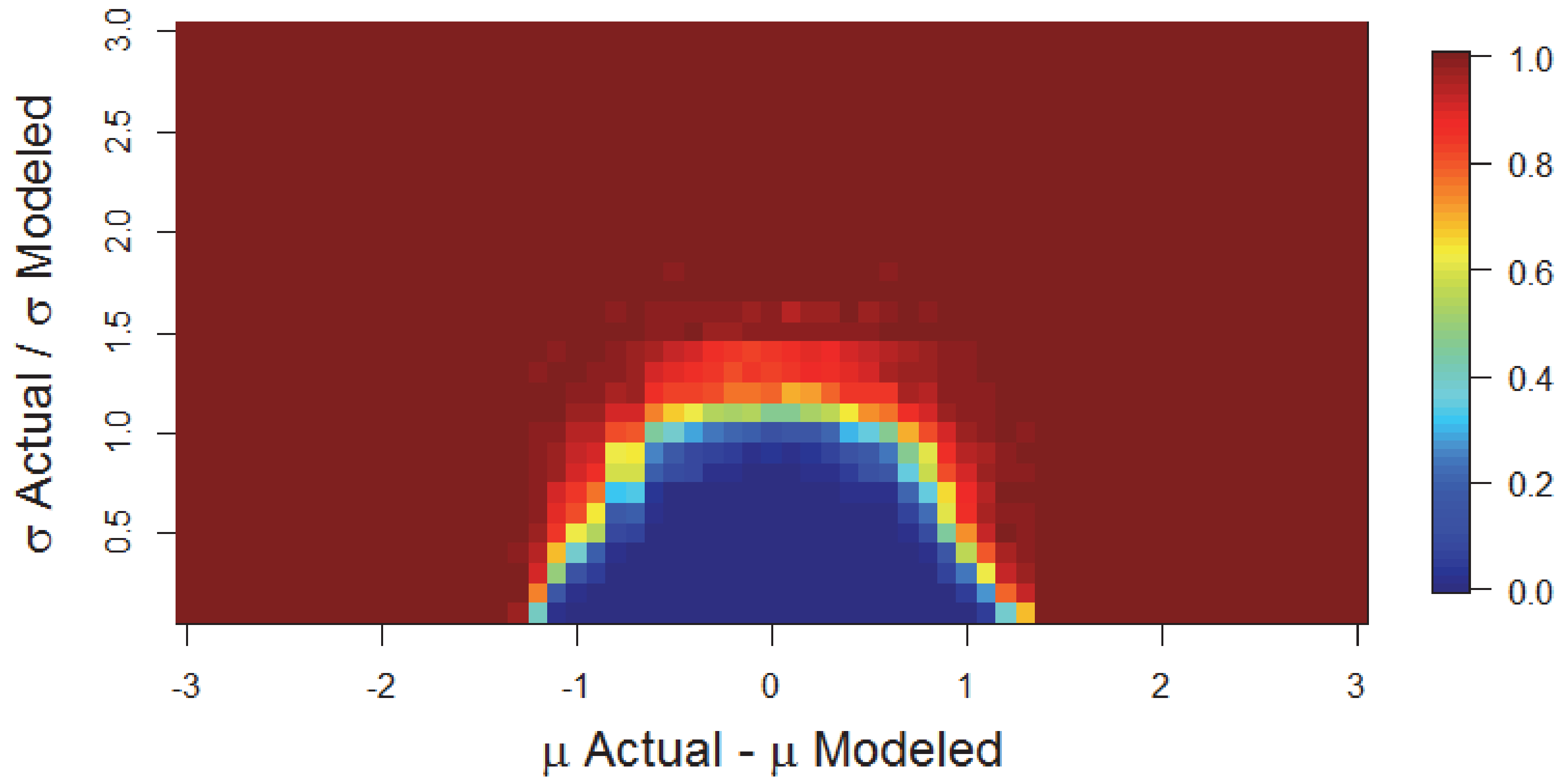
Symmetric

Power Heat Map: Regression



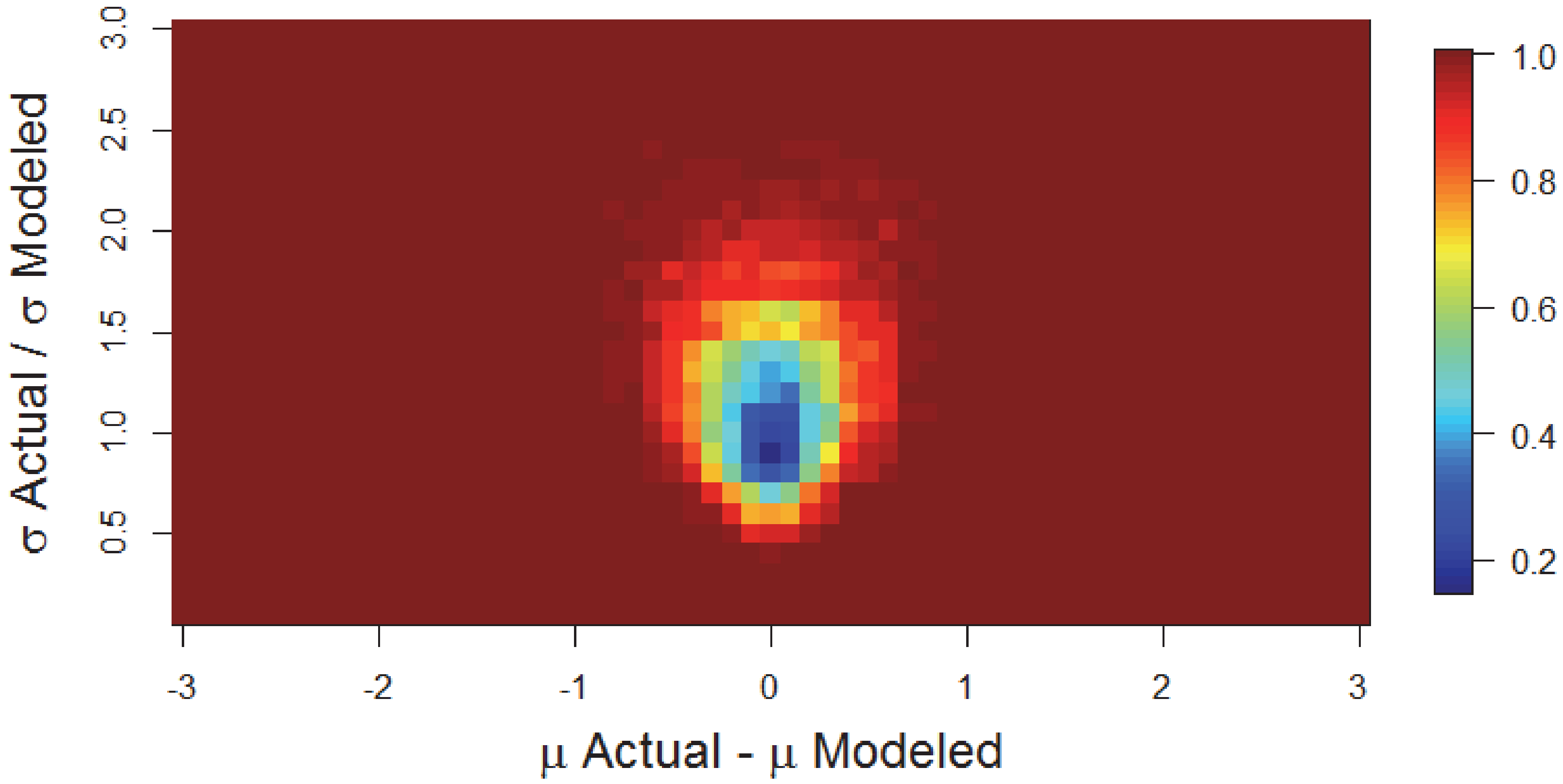
Symmetric

Power Heat Map: Emulation & Prediction



Symmetric

Power Heat Map: Kolmogorov-Smirnov Test



Symmetric

Detailed Recommendations

Distribution	Structure Of Factors	Small Sample Sizes	Moderate Samples Sizes	Large Sample Sizes
Symmetric	Univariate	Fisher's Combined	T-test Fisher's Combined Non-Par KS	T-test Fisher's Combined Non-Par KS
	Distributed Level Effects	Combo Test	Sc Non-Par KS	Sc Non-Par KS
	Designed Experiment	Linear Regression Sc Non-Par KS Emulation & Pred	Linear Regression Sc Non-Par KS Emulation & Pred	Sc Non-Par KS
Skewed	Univariate	Fisher's Combined	Log T-test Fisher's Combined Non-Par KS	Log T-test Fisher's Combined Non-Par KS
	Distributed Level Effects	Combo Test	Sc Non-Par KS	Sc Non-Par KS
	Designed Experiment	Lognormal Regression Sc Non-Par KS Emulation & Pred	Lognormal Regression Sc Non-Par KS Emulation & Pred	Sc Non-Par KS
Binary	Univariate	Fisher's Exact	Fisher's Exact	Fisher's Exact
	Distributed Level Effects	Logistic Regression	Logistic Regression	Logistic Regression
	Designed Experiment	Logistic Regression	Logistic Regression	Logistic Regression

Notes on sample sizes:
Simulation sample size = 100 in all cases; Live sample size (symmetric and skewed): Small = 2-5, Moderate = 5-10, Large = 11-20; Live sample size (binary): Small = 20, Moderate = 40, Large = 100

Conclusions

Design of experiments techniques should be used to efficiently cover the simulation domain and inform live testing

Regression analysis is the most powerful comparison when matched experimental designs are used

More robust nonparametric techniques, such as the Kolmogorov-Smirnov test, provide widely applicable solutions for the comparison

Future work:

- Bayesian statistical methods
- Calibration

This page intentionally left blank.

Statistical Techniques for Modeling and Simulation Validation

Dr. Kelly McGinnity Avery
Dr. Laura J. Freeman
Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, VA 22311
703-845-2265, 703-845-2084
kavery@ida.org, lfreeman@ida.org

Keywords: Operational test, Validation, Statistics, Modeling and simulation

ABSTRACT: *In Defense system test and evaluation, collecting sufficient data to evaluate system performance against operationally realistic threats is often not possible due to cost and resource restrictions, safety concerns, or lack of adequate or representative threats. Thus, modeling and simulation (M&S) tools frequently are used to augment live testing in order to facilitate a more complete evaluation of performance.*

When M&S is used as part of an operational evaluation, the M&S capability first should be rigorously validated to ensure it is representing the real world adequately for the intended use. Specifically, the usefulness and limitations of the M&S should be well characterized, and uncertainty should be quantified to the extent possible. Many statistical techniques are available to rigorously compare M&S output with live test data. This paper will describe some of these methodologies and present recommendations for a variety of data types and sizes.

We will show how design for computer experiments can be used to efficiently cover the simulation domain and inform live testing. Experimental design and corresponding statistical analysis techniques for comparing live and simulated data will be discussed and compared. A simulation study shows that regression analysis is the most powerful comparison when experimental design techniques are used, while more robust non-parametric techniques provide widely applicable solutions for the comparison.

1. Introduction

The Department of Defense (DoD) acquires some of the world's most complex systems. These systems push the limits of existing technology and span a wide range of domains. Before the DoD acquires any high risk, high cost, or large scale weapon system or military equipment, that system must first undergo a series of tests. The final test prior to full rate production and fielding is operational testing (OT). OT evaluates operational effectiveness, suitability, and survivability. This evaluation relies on characterizing supporting performance metrics related to these overarching constructs when the system is employed by representative, trained users in an environment as close as possible to the operational setting in which the system will be used in combat.

Evaluations of operational effectiveness, suitability, and survivability increasingly rely on M&S to supplement live testing. For example, in the case of a single use air-to-air weapon, limited numbers of weapons and limited aerial targets result in a high cost per shot, prohibiting a full characterization of the weapon across many different conditions. Testers also cannot put users in harm's way, so in the case of testing a helicopter's self-defense system, an inert missile could be fired at a manned aircraft, a live warhead could be fired at an unmanned aircraft, but the full end-to-end mission never could be evaluated in a live test environment. In other cases, a system's ability to defeat a certain foreign threat might not be testable simply because we do not have access to that threat system.

When relying on M&S to support an OT evaluation for reasons such as those described above, one should thoroughly understand how well the M&S represents the system of interest across all conditions in which the system will be used. The process of determining the degree to which a model and its associated data provide an accurate representation of the real world from the perspective of the intended uses of the model is known as validation [1]. Any number of techniques can be used to support the validation of a model [2], but a portion of any validation should include a quantitative comparison of the simulation output to available live data. In many end-to-end evaluations, live data is not available. In those cases M&S should be compared to the most realistic data available. Such an analysis provides an objective statistical measure of the differences between the M&S and the live test, quantifies the uncertainty in the model, and can identify areas of high risk that might require additional testing.

Design of experiments (DOE) techniques, discussed in Chapter 2, can be used to efficiently cover the factor space of interest and gather the appropriate data from both the computer simulation environment and the live environment to support a meaningful comparison. Chapter 3 provides an overview of some statistical methods that can help support the characterization of the accuracy of the model by providing quantitative comparisons of M&S data to reference data gathered from live testing. Chapter 4 describes a simulation study used to obtain recommendations for which techniques to use in different situations. For example, if a designed experiment was executed, more advanced and sensitive statistical methods such as regression analysis should be considered. Finally, Chapter 5 outlines some conclusions and areas for future study.

2. Design of Experiments

Design of experiments (DOE) is a rich scientific methodology, containing many design types [3]. Two important subsets for this paper are classical and computer experiments. Figure 2.1 shows common types of classical designs. Classical designs include factorial, fractional factorial, response surface, and, more recently, optimal designs. The classical designs seek to partition variability in test outcomes into the variability that can be explained and variability that cannot be explained. The explainable variability is captured by the input variables (termed factors) and low order polynomial functions of input variables (e.g., quadratic terms, interactions). The remaining variability is unexplained. Replicates can be used to partition the unexplained variability into pure error (i.e. the inherent variability under a fixed set of input conditions) and lack of fit variability (variability potentially explained by lurking variables and higher order terms not include in the model). The statistical significance of a variable or group of variables is determined by comparing the variability explained by those variables to the unexplained variability. Due to this definition of significance, these designs are specifically geared towards stochastic test outcomes.

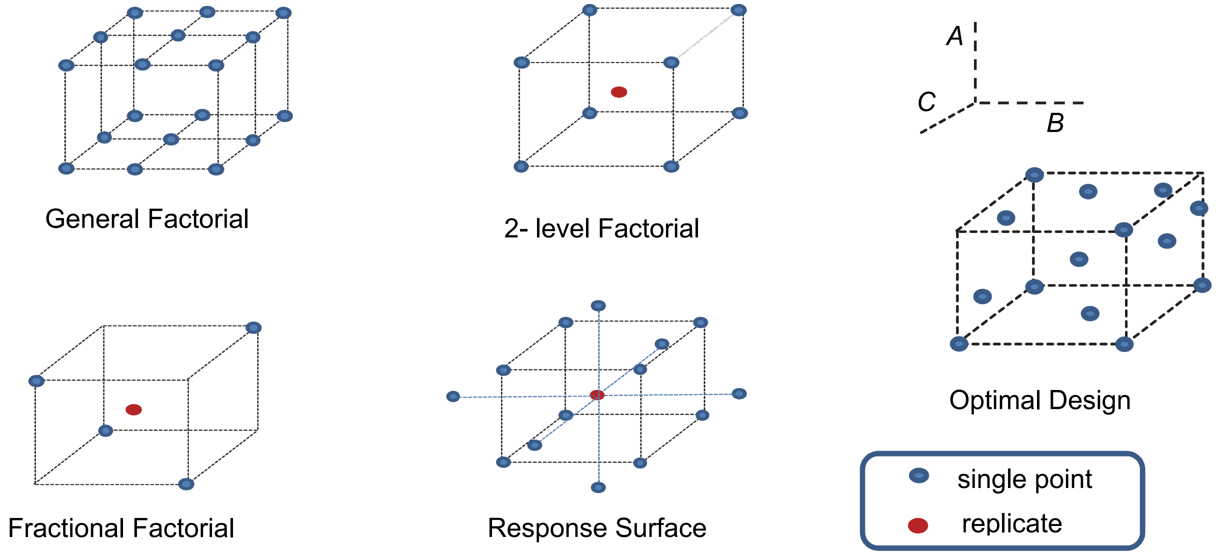
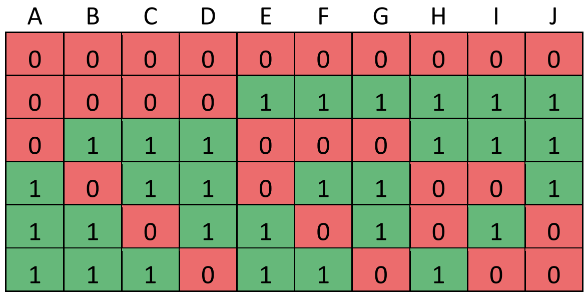


Figure 2.1. Classical design approaches for a three factor (A,B,C) experiment.

The other important subset of DOE is Design for Computer Experiments. Computer experiments assume that the outcome of the test is deterministic (i.e., for the same set of input conditions, multiple replicates of a test will result in the exact same output). Replication is not useful in a computer experiment, because there is no pure error. Rather, computer experiments seek to maximize the probability of finding defects in software that are based on combinations of input conditions. There are two primary classes of computer experiments: factor covering (categorical inputs) and space filling (continuous inputs). Figure 2.2 shows a factor covering design for a computer simulation with 10 input variables (A-J) that have two settings (on=1/off=0). Notice that 10 variables, each with two levels, would require 1024 runs for a full factorial. The factor covering design shown in figure 2 covers all possible two-way combinations in only six runs. However, if a computer fault (e.g. crash or error) occurs, additional information will be required to determine the root cause because several two-way pairs are run simultaneously.



A	B	C	D	E	F	G	H	I	J
0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	1	1	1	1	1
0	1	1	1	0	0	0	1	1	1
1	0	1	1	0	1	1	0	0	1
1	1	0	1	1	0	1	0	1	0
1	1	1	0	1	1	0	1	0	0

Figure 2.2 Factor covering design for 10 factors with two levels (0/1).

Figure 2.3, on the other hand, shows a 12 run space filling Latin hypercube design [4] across five continuous variables, each with an infinite number of possible inputs. The space filling designs attempts to distribute a specified number of points across the space to maximize fault detection.

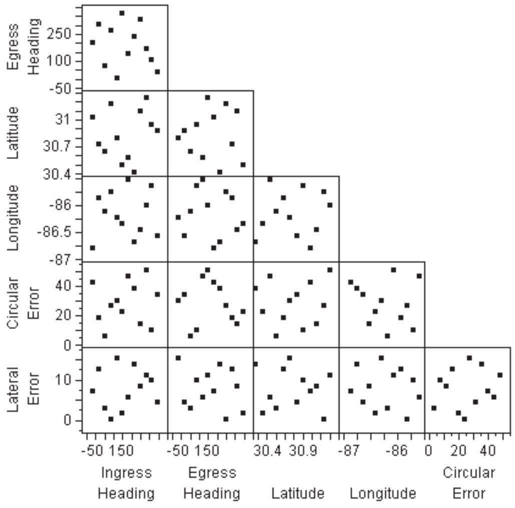


Figure 2.3. Example space filling design with 12 data points spanning 5 factors.

All three design strategies are important to consider in the context of designing tests to cover the M&S space, the live test space, and match test points between live tests and M&S. The test design strategy selected will inform what validation analyses are possible and/or optimum.

Figure 2.4 shows conceptually how one would select points from M&S and live testing. A key element of this selection is what points to conduct in both live and M&S testing. Direct matching of points provides the most powerful validation strategy. Additionally, if the subset of matched points are based on a classical designed experiment, the validation strategy can include regression analysis, which, as discussed in the next section, is the most powerful validation approach. If the matched data do not come from a designed experiment, statistical comparison is still possible but will require aggregation of the comparison across conditions using a less powerful approach. If data cannot be matched directly, emulation and prediction is the best approach for validation.

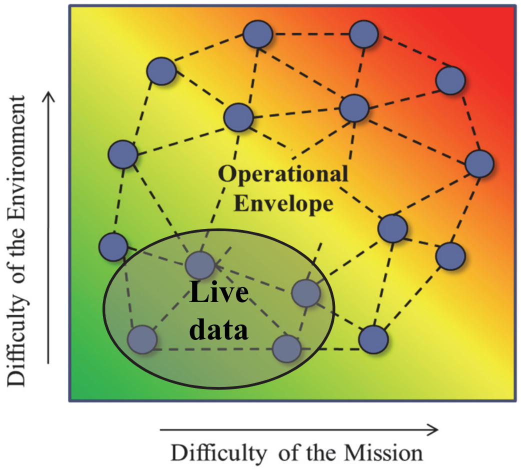


Figure 2.4. Notional picture of data selection across M&S and live testing.

By using a combination of computer experiments and classical designs we can ensure sufficient data for validation and maximize the probability that we identify problems using the M&S. A common strategy used in the air-to-air missile test paradigm is to use space filling designs to efficiently cover the full M&S space and use classical designs to select the limited live tests from possible live shots within range and safety restrictions.

3. Statistical Techniques for Comparing Live and Simulation Data

The primary purpose of statistical validation is to detect and quantify differences between the live test data and simulation output. Understanding these differences and the uncertainty associated with them will inform if the simulation is adequate for the intended use. It is important to note that a simulation need not match live data exactly in order to be useful.

Differences between live (referent) data and simulation output could be in terms of mean (or any measure of center), variance (or measure of spread), or distribution more generally. It is typically assumed that at least as much simulation data can be obtained as live data.

Figure 3.1 compares the distribution of several dozen simulation data points (blue curve) to a handful of live data points (red lines) under the same condition. In the top panel, the two data sets appear to match quite well in terms of center and spread. In the center panel, the mean of the live data is noticeably higher than that of the simulation data. In the bottom panel, the spread of the live data is much smaller than that of the simulation data. While these particular differences are easy to see by eye, statistical techniques can be crucial to detecting such differences in the presence of multiple factors and interactions or noisy data.

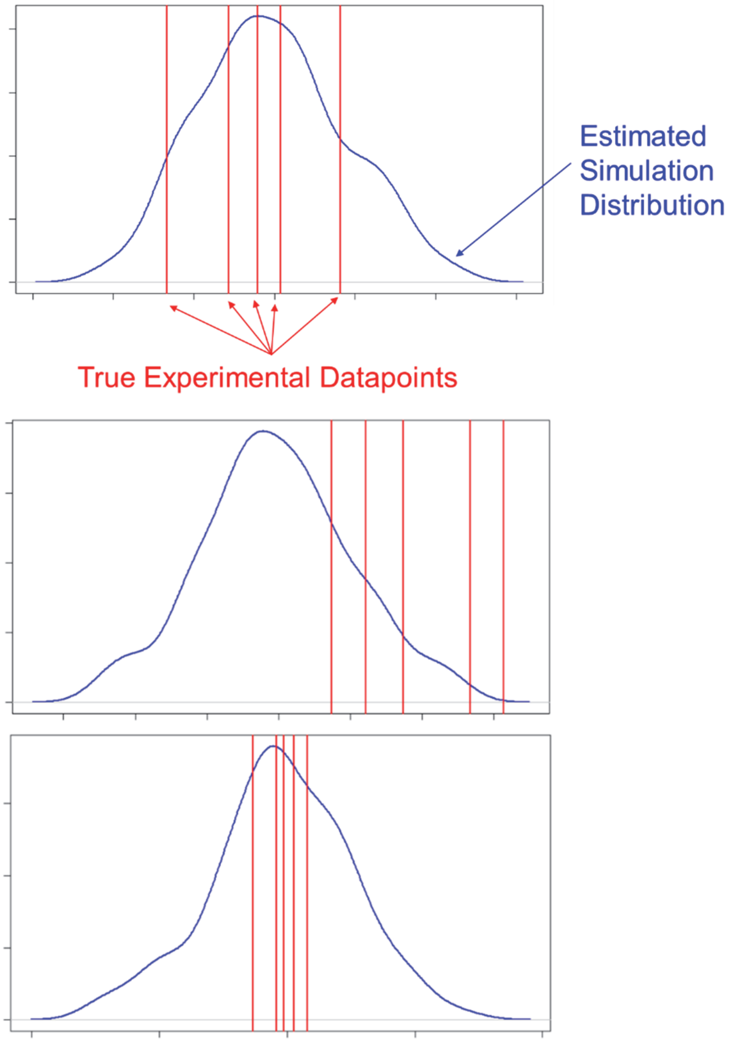


Figure 3.1. Comparison of live data to simulation output where data sets match (top panel), are different in center (middle panel), and are different in spread (bottom panel).

The most appropriate method for statistically comparing live data and simulated output will depend on a variety of circumstances. There is no one-size-fits-all solution. In some cases, it may be useful or necessary to apply multiple techniques in order to fully understand and describe the strengths and weaknesses of the M&S. The distribution of the response variable, or metric of interest, will inherently drive the techniques that can be used. For example, binary (pass/fail) data should not be analyzed in the same way as continuous data, such as miss distance or detection time. The presence or absence of factors (variable operating conditions) will also influence which methods are preferred. Applying a simple hypothesis test on the means, for example, to a data set obtained across multiple diverse conditions should be avoided as it will cover up any potential differences in how the M&S performs across the operational space. Finally, the amount and structure of live data collected limits the application of certain statistical techniques. In cases where only a handful of live data points are available, approaches that involve building a statistical model with the live data may not be feasible. If a designed experiment is used to select the data points for comparison, then all of the following techniques are available. If there is no direct data matching between live tests and M&S, validation approaches are limited. However, even when there is no direct match between live tests and M&S, the simulation data still can be modeled to determine how well it can predict live outcomes, and quantitative comparisons still can be made.

The following subsections outline different categories of techniques that can be used to support model validation in certain cases. The list of methods is not meant to be exhaustive or prescriptive, nor are any of the techniques new or innovative. The value here is the application of these statistical techniques within an M&S context. Chapter 4 outlines a simulation-based method for determining which techniques are optimal under specific conditions.

3.1 Univariate Techniques

One overarching class of statistical methods are hypothesis tests on a single parameter. Such tests compare a certain metric of interest, such as the mean or the variance, of one data set with that of another data set. Examples of methods in this class include the t-test, the F-test for equality of variances, empirical median or quantile tests, and several nonparametric procedures, such as the Wilcoxon signed rank and rank sum tests [5]. These tests are aggregate tests in that they are not designed to account for factor differences.

Probably the most common of these tests, and the one that performed best in our simulation study for univariate data, is the t-test. The t-test is a parametric test used to compare the means of two data sets. It is only valid if the data is approximately normal[1] and observations are independent of one another. The t-test is a powerful tool for detecting differences in means, provided the parametric assumptions are met. However, the test is limited in the sense that it only determines changes in means. The t-test is not designed to check for differences in variance, nor does it account for any possible factor effects.

Overall, these techniques can be useful for quickly invalidating an M&S, but since they only capture one dimension of the data, they typically are not sufficient for a comprehensive statistical comparison between M&S and live tests.

3.2 Distribution Techniques

A second class of methods compare the entire distribution, or shape, of a data set to that of another data set, rather than focusing on a particular summary statistic such as the mean or variance. Examples of statistical tests that employ distributions include the Kolmogorov Smirnov (KS) Test, Anderson Darling [6], and Fisher's Combined Probability Test [7].

Methods that worked well for many settings in our simulation study were the non-parametric KS test, and Fisher's Combined Probability Test. The nonparametric KS test uses an empirical distance metric to determine whether two samples are drawn from populations with the same distribution. The KS test can be adapted to account for data collected under different conditions by scaling[2] the data first. One of the biggest strengths of the KS test is that it does not depend on the underlying distribution being tested. It is computationally simple and provides a good starting point

[1] If the data is lognormally distributed, a t-test can be performed on the log-transformed data using the same procedure.

[2] For each individual data point, subtract the mean of the data in that condition and divide by the standard deviation of the data in that condition.

to direct further analysis, as it considers the overall shape of a distribution rather than focusing specifically on central tendency or dispersion. Also, it can be modified easily to account for factors (i.e., data collected under different conditions). The main limitation of the KS test in the context of validating M&S is that multiple live data points are required in order to form a distribution.

Fisher's Combined Probability Test is a method for combining the results from several independent tests, each testing common hypotheses of interest. In the M&S context, "tests" might be environmental condition of interest across which you want to validate the model. While there are multiple ways to calculate the relative significance of each individual test, generally speaking, the goal is to figure out where in the distribution of M&S outcomes a single live data point lies. If enough of the live data points fall in the extremes of the model distribution, that is an indicator that the model is not representing reality. Fisher's Combined Probability Test is a powerful technique for detecting differences in variance between the live data and simulation outcomes. However, one should be cautious that a single extreme result is not driving the overall conclusion. In this case, the assessor should explore and discuss why and how the extreme result occurred and whether or not that alone should invalidate the M&S. In addition, Fisher's test does not account for factor effects the way a statistical model would. If a structure designed experiment was performed in both the live and simulated environments, this is not the best technique to apply.

3.3 Regression based techniques

Statistical regression methods are used to model and analyze several variables, or factors, at a time. If data is collected across a variety of conditions using a classical design approach, regression techniques can easily separate the effects of each individual factor on the response, as well as detect any interaction, or synergistic relationship, between factors. [3]

In the M&S context, one way to leverage the benefits of such techniques is to combine the data from the live test with that of the M&S and fit a single regression model that includes data source (live or M&S) as a model term, in addition to any other factors of interest, and interactions between data source and the other factors of interest. In this scenario, if the source term is statistically significant, that indicates that the M&S and the live data are different from each other. If an interaction term between data source and another factor is significant, the M&S may perform well under some conditions, but not others. An important consideration when using this single regression model technique is to ensure the factor spaces in the live and simulation environments are the same. In addition, the sample sizes between the live and simulation data should be relatively even, otherwise correlation problems could occur. Therefore, only M&S data matched to the live data space should be included in the regression model. Given sufficient statistical power, this technique is a good way to detect differences in means when a matched designed experiment was executed.

Another regression-based technique to consider, especially if live data is limited or direct matching between M&S and live is not possible, is emulation and prediction. In this case, outputs from the M&S only are used to build an empirical emulator, or statistical model, characterizing the response as a function of each factor. As live data points become available, they are compared to the prediction interval generated from the emulator under the same conditions. If a live point falls within the prediction interval, there is evidence that the simulation is performing well under those conditions. Conversely, if a live point falls outside of the prediction interval, one should investigate why the emulator is failing under certain conditions and test for any systematic patterns. This method is most effective when used in conjunction with a designed experiment, and when multiple simulation runs can be performed for each condition. Emulation and prediction is a powerful technique for detecting when the variance of the live data is larger than that of the simulation data.

Basic linear regression requires the response variable to be approximately normally distributed. However, there are analogous methods for other response distributions. If the data is right skewed, lognormal regression tends to work well, while if the data is binomial, logistic regression is the correct approach.

3.4 Binary and categorical techniques

Most of the statistical techniques for validation discussed thus far are only appropriate for continuous responses. If the outcome of interest is binary (pass/fail) or categorical in nature, a different class of approaches is needed. Some examples of such techniques include Fisher's Exact Test, chi-squared tests, and permutation tests [8].

Fisher's Exact Test is a powerful procedure for assessing differences in a binary response variable between live and simulation when there are no factors under consideration. If a designed experiment with multiple factors was performed, logistic regression techniques (introduced in the previous section) can be employed. However, for the small sample sizes typically available in DoD testing, statistical validation of a binary outcome typically is not possible as the uncertainty associated with the results will be large.

4. Simulation Study

Monte Carlo power simulations were performed under various settings to determine the most appropriate technique(s). Monte Carlo simulation uses iterations of random draws from a probability distribution to approximate a solution to a problem [9]. In this context, we are interested in how reliably a particular statistical technique signals that the M&S differs from live data, given that they really are different. If we simulate thousands of live data sets and thousands of M&S data sets that are different by some user specified-amount, and calculate the percentage of times the test is "correct," this is an estimate of the statistical power of that test.

The specific conditions studied via Monte Carlo simulation[3] were:

- Distribution of response (symmetric, skewed, binary)
- Live sample size (100 a small sample size is 2 to 4 samples for continuous data or 20 for binary data, medium size is 6 to 10 samples for continuous data or 50 for binary data, large size is 11-20 samples for continuous data or 100 for binary data)
- Structure of factors (univariate, distributed level effects[4], designed experiment).

The sample size for the simulation was assumed to be 100 in all cases.

Table 4.1 below shows the recommended techniques in each category. In the cases where there are multiple methods per cell, more than one test is required to have high power to test differences in both mean and variance. In addition, some tests are more sensitive to cases where the live test data is more variable than the simulation data, while others perform better in the reverse case, where the simulation data is more variable than the live data. Thus, it may be necessary to use up to three techniques depending on the goals of the validation study.

Generally speaking, robust non-parametric techniques, such as the Kolmogorov-Smirnov test and Fisher's Combined Probability Test, are widely applicable across the majority of conditions with continuous data. Whenever a designed experiment was performed, regression modeling techniques generally performed well since they are designed to optimally leverage all information in the presence of factors.

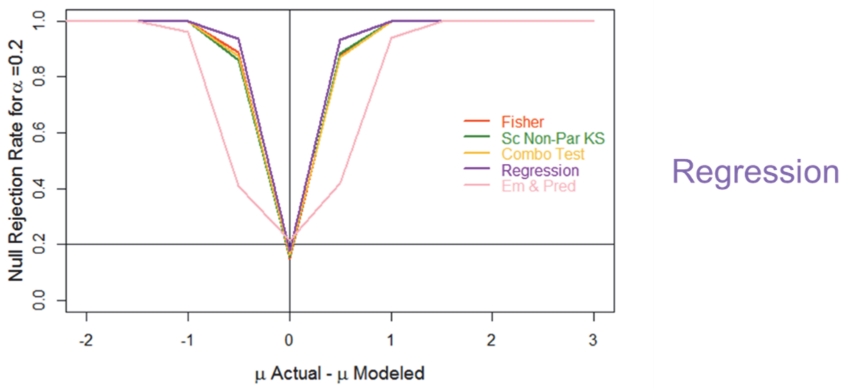
[3] In all cases, Type I error was fixed at 0.2. In cases where the actual Type I error was not as initially specified, an empirical correction was applied.

[4] The intent of the "distributed level effects" category is to capture the case where multiple conditions are tested, but the factors are not varied in any systematic way, as would be the case in a designed experiment. To simulate this, the differences in response variable across the set of conditions followed a lognormal distribution.

		Sample Size		
Distribution	**Structure Of Factors**	**Small**	**Moderate**	**Large**
Skewed	Univariate	Fisher's Combined	Log T-test Fisher's Combined Non-Par KS	Log T-test Fisher's Combined Non-Par KS
	Distributed Level Effects	Scaled Non-Par KS	Scaled Non-Par KS	Scaled Non-Par KS
	Designed Experiment	Regression Scaled Non-Par KS Emulation & Pred	Regression Scaled Non-Par KS Emulation & Pred	Scaled Non-Par KS
Symmetric	Univariate	Fisher's Combined	T-test Fisher's Combined Non-Par KS	T-test Fisher's Combined Non-Par KS
	Distributed Level Effects	Scaled Non-Par KS	Scaled Non-Par KS	Scaled Non-Par KS
	Designed Experiment	Regression Scaled Non-Par KS Emulation & Pred	Regression Scaled Non-Par KS Emulation & Pred	Scaled Non-Par KS
Binary	Univariate	Fisher's Exact	Fisher's Exact	Fisher's Exact
	Distributed Level Effects	Logistic Regression	Logistic Regression	Logistic Regression
	Designed Experiment	Logistic Regression	Logistic Regression	Logistic Regression

Table 4.1. Statistical Methods Recommendations

Figures 4.1 and 4.2 provide examples of the power curves and heat maps used to decide the “winning” method(s) in each category above. For brevity, only the results for a designed experiment with symmetric data with a moderate sample size are included. Figure 4.1 shows power curves for mean and variance for all applicable methods. The ideal curve will hit .2 at no change in mean/variance and then increase to 1 as quickly as possible. In this case regression performs best for a change in the mean, while the non-parametric KS and emulation and prediction perform best for a change in the variance. Notice that non-parametric KS does better if the live test variance is larger than the M&S variance, and the emulation and prediction method does better if the live test variance is larger than the M&S variance.



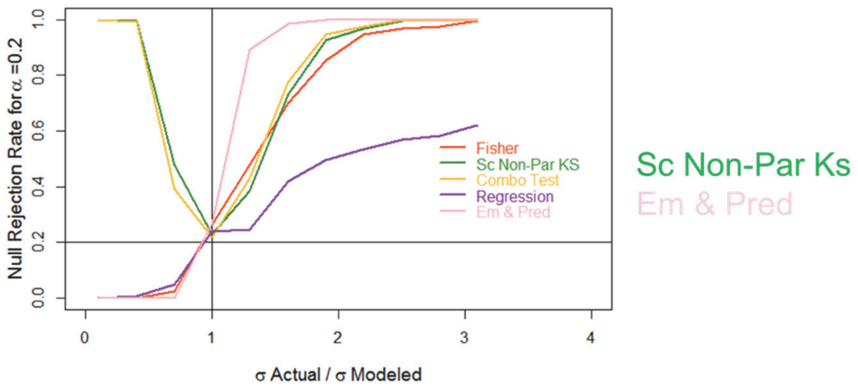
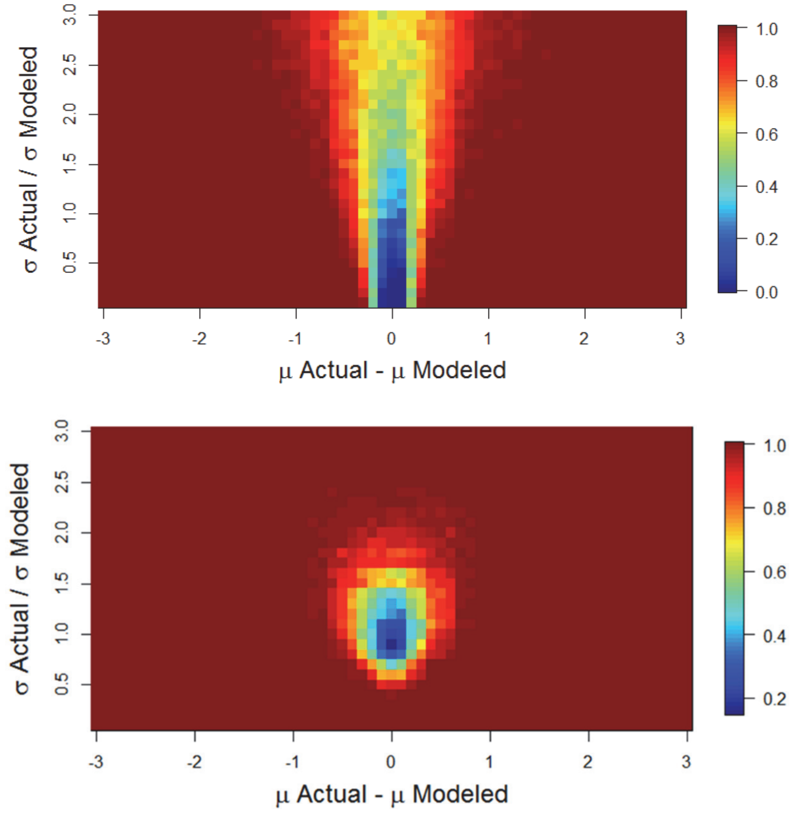


Figure 4.1. Power curves for mean (top graph) and variance (bottom graph) for symmetric data, designed experiment, moderate live sample size.

Figure 4.2 shows three power heat maps, one for each of the recommended methods from Figure 4.1. The ideal heat map has a tiny blue dot (power of 0.2) at no change in mean or variance and is dark red everywhere else. The figure shows that the scaled nonparametric KS test can detect changes in the mean and variance, while the regression analysis does poorly at detecting changes in variance. This is somewhat by design as the regression analysis does not include a formal test of the variance. This figure also re-emphasizes that emulation and prediction only works well when the actual data has larger variance than the M&S.



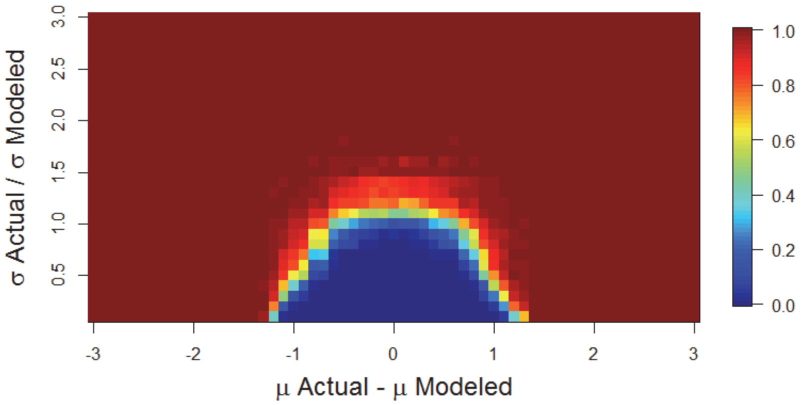


Figure 4.2. Power heat maps for regression (top graph), scaled nonparametric KS test (center graph) and emulation and prediction (bottom graph) for symmetric data, designed experiment, moderate live sample size.

5. Conclusion and Future Work

Ensuring that M&S capabilities on which operational evaluations rely adequately represent reality is an essential step before acquiring and deploying a new system to our nation's warfighters. To the extent possible, quantitative and defensible statistical methods should be used to both design efficient experiments and make meaningful comparisons of test data to M&S output.

Validation is a complex process and the most appropriate techniques depend on the goals of the analysis and the type and amount of data that can be collected. Based on a simulation study of statistical power, robust statistical approaches that are widely applicable in the M&S context include the KS test and regression modeling.

Our study is another step in helping to improve the quality and rigor of model validation in the DoD. Previous work in this area [10] [11] tend to ignore the limited live data problem that is prevalent in OT. However, much more work can and should be done to explore the applicability and performance of other techniques. Bayesian methodologies that allow prior information to be merged with empirical results have the potential to be very helpful in the context where little live data are available and thus are worthy of future study. In addition, systems and M&S capabilities that output continuous, functional responses over time could be analyzed more effectively using statistical process control or time series techniques.

6. References

[1] Department of Defense: "DoD Modeling and Simulation (M&S) Verification, Validation, and Accreditation (VV&A)," DoD Instruction 5000.61, December, 2009.

[2] Robert G. Sargent: "Verification and Validation of Simulation Models," Proceedings of the 2003 Winter Simulation Conference, pp. 37-48

[3] Douglas C. Montgomery, "Design and analysis of experiments," John Wiley & Sons, 1990.

[4] Jeong-Soo Park, "Optimal Latin-hypercube designs for computer experiments," Journal of statistical planning and inference 39.1, pp. 95-111, 1994.

[5] George E. P. Box, William G. Hunter and J. Stuart Hunter, "Statistics for experimenters," 2nd edition, John Wiley & Sons, 2005.

[6] M.A. Stephens, “EDF Statistics for Goodness of Fit and Some Comparisons,” Journal of the American Statistical Association 69.347, pp. 730-737, 1974.

[7] Ronald A. Fisher, “Statistical Methods for Research Workers,” Oliver and Boyd, 1925.

[8] Alan Agresti, “Categorical Data Analysis,” 2nd edition, John Wiley & Sons, 2002.

[9] Christian P. Robert: “Monte Carlo Methods” John Wiley & Sons, Ltd, 2004.

[10] Jack PC. Kleijnen, "Verification and validation of simulation models," European journal of operational research 82.1, pp. 145-162, 1995.

[11] S. Y. Harmon and Simone M. Youngblood, "A proposed model for simulation validation process maturity," The Journal of Defense Modeling and Simulation 2.4, pp. 179-190, 2005.

Author Biographies

KELLY MCGINNITY AVERY is a Research Staff Member at the Institute for Defense Analyses. She is a statistical consultant for the test and evaluation of tactical aircraft, satellite systems, and software-intensive systems, and provides guidance to the Director, Operational Test and Evaluation on the use of rigorous statistics in test and evaluation. Her areas of expertise include statistical process control, design of experiments, and statistical modeling. Dr. Avery has a B.S. in Statistics, a M.S. in Applied Statistics, and a Ph.D. in Statistics, all from Florida State University.

LAURA J. FREEMAN is currently an Assistant Director with the Institute for Defense Analyses, Alexandria, Virginia, where she leads the Test Science Task providing support to the Director, Operational Test and Evaluation on the use of statistics in test and evaluation. Her areas of statistical expertise include designed experiments, reliability analysis, and industrial statistics. She focuses on operational tests for the F-35 Program. In addition, she has a background in aerospace engineering. She received a B.S. degree in Aerospace Engineering, and M.S. and Ph.D. degrees in Statistics from the Virginia Polytechnic Institute and State University, Blacksburg, Virginia.