Statistical Design and Validation of Modeling and Simulation (M&S) Tools Used in Operational Testing (OT)

Kelly McGinnity, Laura Freeman
Institute for Defense Analyses
October 27, 2015
Motivation

• Models and simulations are increasingly becoming an essential element of operational test and evaluation
  – Collecting sufficient data to evaluate system performance is often not possible due to time, cost, and resource restrictions, safety concerns, or lack of adequate / representative live threats

• There is currently little to no DoD guidance on the science of validating such models
  – Which / how many points within the operational space should be chosen for optimal ability to verify and validate the M&S?
  – What is the best way to statistically compare the live trials to the simulated trials for the purpose of validating the M&S?
  – How close is close enough?
Outline

• Examples of M&S in OT&E
  • Validating the Simulation
  • Designing the Simulation Experiment
Why do I need M&S to assess Operational Effectiveness and Suitability?

- Expansion of the operational space from what can be done live
  - High threat density (air and ground)

- Frame the operational space
  - Large number of factors contribute to performance outcomes

- Improve understanding of operational space
  - Limited live data available

- End-to-end mission evaluation

- Translation of test outcomes to operational impact
Expansion of the Operational Space: Air Combat Simulator F-22 Raptor

• **Why we need M&S:**
  – System is specifically designed to operate in higher threat densities and against more challenging threats than we can test open air (5th gen problems)

• **Expanding the Operational Space**
  – Higher air threat densities
  – Supports end-to-end missions with more fidelity than real time casualty assessments

• **M&S Solution:**
  – Complex, integrated simulation capability incorporating multiple simulation integration labs, operator-, hardware-, and software-in-the-loop
  – Allows for end-to-end mission conduct in a simulated environment
Expansion of the Operational Space: Air Combat Simulator (ACS) F-22 Raptor
• Leave behind benefits of high fidelity M&S
  - FOT&E - Large potential reductions in live flight testing if we understand the modeling capabilities
  - Training
  - Tactics Development
Frame the Operational Space: Weapons Assessment Facility (WAF)

- Hardware-in-the-loop simulation capability for lightweight and heavyweight torpedoes
- Creates simulated acoustic environment
  - Sonar propagation
  - Ocean features
  - Submarine targets
- Interfaces with torpedo guidance and control scenarios
- Why we need M&S?
  - Complex operational space where performance is a function of many environmental factors
- Limitations
  - Computer processing prohibits full reproduction of full ocean conditions which have limited prediction accuracy
Frame the Operational Space: Weapons Assessment Facility (WAF)

- Run WAF Simulations
  - Dozens of Factors
  - Examine Complex Space

- Characterize Operational Space
  - Determine most important factors from WAF
  - Highlight risk areas

- Plan Operational Testing
  - Use factors identified in WAF (subset of all possible)
  - Informed scope

Even when Modeling and Simulation has limited predictive ability it can still be used to inform operational testing.
Question to be addressed:
- Self-defense requirements for Navy combatants include a Probability of Raid Annihilation (PRA) requirement
- To satisfy the PRA requirement, the ship can defeat an incoming raid of anti-ship cruise missiles (ASCM) with any combination of missiles, countermeasures, or signature reduction

Why we need M&S:
- Safety constraints limit testing
- No single venue where missiles, countermeasures and signature reduction operate together in OT
Improve Understanding: PRA Test Bed

- PRA is a federation of models that is fully digital
  - Many system models are tactical code run on desktop computers
  - Uses high-fidelity models of sensors including propagation and environmental effects
  - Incorporates high-fidelity six-degree-of-freedom missile models

- Limited “live” data from the Self Defense Test Ship provides limited understanding of PRA

- Architecture will be useful for a variety of ship classes
  - LPD 17 was the first successful implementation – provided more information on PRA under the same conditions as live testing
  - LHA 6, DDG 1000, Littoral Combat Ship, CVN 78 will be examined
End-to-End Mission Assessment: Common Infrared Counter Measures (CIRCM)

• **System Overview:**
  – Multiband infrared (IR) pointer/tracker/laser jammer for small/medium rotorcraft and small fixed wing aircraft

• **Why we need M&S:**
  – Shooting live missiles at aircraft is difficult

• **M&S Solution**
  – Simulate end-to-end missile engagements by combining results from multiple test facilities using identical initial conditions
  – Allows the full suppression chain to be assessed
End-to-End Mission Assessment: Common Infrared Counter Measures (CIRCM)

- Integrated Threat Warning Lab
  - Assess flight path/geometry

- Threat Signal Processing in the Loop (T-SPIL)
  - Actual Threat Tracking

- Guided Weapons Evaluation Facility (GWEF)
  - Inclusion of actual seeker and countermeasures supports wider operational space

- Open Air Range, Missile Plum Simulators

- Free-Flight Missile Test
  - Non-representative targets

Acronyms this slide: Infrared (IR) Countermeasures (IRCM); Missile Warning System (MWS);
• For complex systems, the Services use several M&S tools based on discrete event simulations (e.g., Raptor, LCOM) to model Operational Availability ($A_o$). These digital simulations are based on:
  1. Reliability block diagrams
  2. Expected component reliability
  3. Expected maintainability

• Why we need M&S:
  – Operational Availability cannot be assessed across all mission types during live testing
  – Models are useful for assessing sensitivity of operational availability to changing conditions
Modeling Fidelity Terminology and the M&S Space

- Partial tasks
- Full Mission
- Testing Capabilities
- Features/Number of Simulations
- Physics-Based/Accurate/High Detail
- Effects-Based/Less Accurate/Low Detail
- Functional Fidelity
- Interface Fidelity
Key Questions

For each goal:

1. What is the best analysis method for *validating* the simulation?
2. What is the best technique for designing the *simulation* experiment?
3. What is the best technique for designing the *live* experiment?
Framework for M&S use in T&E

M&S Predictions → Statistical Emulator → Live Testing

Model Validation and Refinement

Informs Selection of Live Testing

Analyze Test Results, Consider inclusion of M&S Results

Evaluation

Common Parameter Space

Identify the common set of variables that spans the operational space

Controllable and Recordable Conditions

Full Factor Space

Operational Test Factors

Predictor Variables
Outline

• Examples of M&S in OT&E
• Validating the Simulation
• Designing the Simulation Experiment
Verification, Validation & Accreditation (VV&A)

- All M&S used in T&E must be accredited by the intended user. The Director, Operational Test and Evaluation (DOT&E) determines if a model has been adequately VV&A’d to use in Operational Testing.

- "Verification is the process of determining if the M&S accurately represents the developer's conceptual description and specifications and meets the needs stated in the requirements document."

- "Validation is the process of determining the extent to which the M&S adequately represents the real-world from the perspectives of its intended use."

- "Accreditation is the official determination that the M&S is acceptable for its intended purpose."

“A model should be developed for a specific purpose (or application) and its validity determined with respect to that purpose” (Sargent 2003)
Validation Methods Overview

• **Typically a combination of validation techniques will be used**
  – Comparison to other models
  – Event validity (does the simulation go through all necessary steps?)
  – Face validity (evaluation by subject matter experts)
  – Comparison to historical data
  – Extreme condition comparisons
  – Internal validity

• **Methods that should be used more frequently**
  – Sensitivity analysis – changes to inputs produce reasonable changes to outputs
  – Predictive validation – can the model predict live test outcomes
Challenges

- Approaches will likely be different depending on:
  - Type of model (deterministic vs. stochastic, continuous vs. discrete outcome, etc.)
  - Purpose of the model
  - Amount of data available

- What are the changes in outcomes as we move across test conditions? Do they match live testing? [Factor Effects]

- What is the variability within a fixed condition? Is it representative of live testing? [Run-to-run variation]

- What defines “matching live testing”? What is close enough? [Bias and Variance]

- How do we control statistical error rates? [Type I and Type II errors]
Existing Methods (DoD)

- **Graphical Comparison**
  - Graph test data vs. simulation data, is it a straight line?

- **Confidence Intervals**
  - Comparing confidence intervals about live data to those about sim data

- **Simple hypothesis tests**
  - Compare Means, Variances, Distributions

- **Limitations**
  - Averages over different conditions
    » Combine results and test aggregated data
  - Does not account for factor effects
  - No way to separate problems with bias vs. variance

- **Better Options:**
  - Fisher’s combined probability test
  - Regression modeling
  - Logistic regression model emulator for cross-validation and classification
Fisher’s Combined Probability Test

• **Applied to validation of missile miss distance**
  – 1 live shot per condition
  – Null hypothesis is that the live shot comes from the same distribution as the simulation “cloud”
  – Tail probabilities under each condition combined using a chi-squared test statistic
    » \( X = -2 \sum \ln(p) \) follows a chi-square distribution with \( 2N \) degrees of freedom

• **Strengths**
  – Intuitive way to handle limited data
  – Preferred to the t-test which ignores the variability of the “cloud”
  – Preferred to goodness-of-fit tests for most alternative hypotheses

• **Limitations**
  – Sensitivity to one failed test condition
  – Method requires adjustment if more than 1 live shot per condition is obtained
  – No formal test of factor effects
Regression Modeling

• Developed for validating the Probability of Raid Annihilation (PRA) Test Bed
  – The Navy’s modeling and simulation venue used to examine the ability of shipboard combat systems to defend a ship against a cruise missile attack
  – Only 1 live shot per test condition (4 threat types)
  – Build a statistical model to compare the M&S results to the live test results and test for significant differences
  – Detection Range = $\beta_0 + \beta_1 \text{TestType} + \beta_2 \text{TestThreat} + \beta_3 (\text{TestType} \times \text{TestThreat}) + \epsilon$

• Strengths
  – PRA Testbed runs can be formally compared to the live test events, even when there is limited live data
  – The model allows analysts to test for a Test Type effect, a Test Threat effect, and an interaction effect
    » If the Test Type effect is not statistically significant then the PRA Testbed runs are providing meaningful data
    » If the interaction term is significant, there may be a problem with the simulation under some conditions but not others

• Limitations
  – Relatively weak test
  – Limited data; cannot differentiate between problems with bias vs. variance
  – Parametric model assumptions questionable
Model Emulator for Cross-Validation and Classification

• Build an empirical emulator (e.g. a logistic regression model) from the simulation
  – As a new set of live data becomes available, compare each point with the prediction interval generated from the emulator under the same conditions
    » If a live point falls within the prediction interval, that is evidence that the simulation is performing well under those conditions
  – Compare/model the live points that do vs. don’t fall within the emulator prediction intervals and test for any systematic patterns
    » Will help explain where / why the simulation is failing in certain cases
  – Once the live data is classified or “tested”, it can then be used to update the simulation and continue to “train” the model

• Strengths
  – Applicable to any amount of live data
  – Can test for factor effects, as well as differentiate between problems with bias and variance (in the case of >1 live shot per condition)
  – Live data serves dual purposes of validating and updating the model
  – Emulator can help inform the live test

• Limitations
  – Not reasonable in the case of 1 or very few simulation runs per condition
Validation Recommendations

- Avoid using basic hypothesis tests or averaging results across conditions.

- Given limited data and no real factors, Fisher’s Combined Probability Test is a reasonable and intuitive approach.

- Otherwise, one of the modeling approaches is recommended:
  - Allows for rigorous testing of factor effects.

- More advanced methods may become feasible as statistics in the DoD advances and M&S test designs are developed appropriately.
Outline

• Examples of M&S in OT&E

• Validating the Simulation

• Designing the Simulation Experiment
Design of Experiments (DOE) provides a framework for selecting:
- Which simulation runs?
- Which live runs?
- How to validate?

Facilitates answering the key validation questions

1. What are the changes in outcomes as we move across test conditions? Do they match live testing? [Factor Effects]
2. What is the variability within a fixed condition? Is it representative of live testing? [Run-to-run variation]
3. What defines “matching live testing”? What is close enough? [Bias and Variance]
4. How do we control statistical error rates? [Type I and Type II errors]
Types of Designs – Overview

- Classical Factorials
- Fractional Factorial Designs
- Response Surface Method Designs
- Optimal Designs
- Combinatorial Designs
- Software Testing/Deterministic Processes

Number of Factors

# Levels Per Factor Needed

Constraints/Complexity of Surface
How do we choose the best design?

- Most appropriate design choice depends on:
  - The purpose of the M&S / goal of the validation analysis
  - The type of simulation (deterministic vs. stochastic)
  - The nature of the data (categorical vs. discrete)
  - The model terms desired to be estimated (e.g. what the “emulator” should look like)

- Various selection criteria for design evaluation:
  - High statistical power for important effects
  - Robustness to missing data
  - Low correlation between factors
  - Maximize the number of estimable main effects, two factor interactions and other higher order terms (depending on the goal of the test)
  - Minimize correlation between two-factor interactions and main effects
DOE for Deterministic M&S

- **Space Filling Designs**
  - An efficient way to search or cover large continuous input spaces
  - Algorithms spread out test points using tailored optimality criteria
  - Analyzed via Gaussian process models

- **Factor Covering Arrays**
  - Type of combinatorial design; used to find problems
  - An efficient way to test when the space is large and made up of combinations of selections (categorical / binary input)

- **Computer simulation experiments**
  - Many recent methods in academic literature
  - Parameter calibration using Gaussian Stochastic Process Models
  - Bayesian techniques
DOE for Stochastic M&S

- **Classical Factorial Designs**
  - Full coverage
  - Highest fidelity
  - All model terms estimable

- **Screening Designs (e.g. Fractional Fact.)**
  - Good for testing many factors at once
  - Lower fidelity
  - Some aliasing / inestimable terms

- **Response Surface Designs**
  - Best for characterizing a few continuous factors
  - Allows testing for curvature

- **Optimal Designs**
  - Most efficient and flexible
  - Allows for constrained spaces, disallowed combinations, etc.
Design Recommendations by Goal

• Expansion of the operational space from what can be done live
  – Need to facilitative extrapolation across the space
  – Classical factorial designs, Response Surface, Optimal
  – Ensure there is some overlap (anchor points) between live test and simulation experiment if possible

• Frame the operational space
  – Many potential factors
  – Screening or Optimal designs

• Improve understanding of operational space
  – Limited live data
  – Replicate live points
  – Space Filling (if deterministic), Response Surface or Optimal otherwise

• End-to-end mission evaluation
  – Design must be repeatable across venues
  – Factorial or Response Surface

• Translation of test outcomes to operational impact
  – Test for sensitivity to changing conditions
  – Space Filling / Covering Arrays (if deterministic), Response Surface or Optimal otherwise
Conclusions

- Statistical rigor of M&S validation in OT needs improvement

- The goal of the M&S and its role in OT evaluations should inform both the design of the simulation experiment and the analysis method used to validate it

- Design of experiments techniques can improve the efficiency of testing and optimize the information gained
  - The dual purpose of live testing (characterization and validation) needs to be considered

- Rigorous statistical analyses can characterize the extent to which the simulation matches the live data
  - Process should be iterative

- More work to be done via future research, case studies, and policy guidance
References


• Thomas, Dean and Dickinson, R. “Validating the PRA Testbed Using a Statistically Rigorous Approach.” IDA Document NS D-5445, 2015.


References II


Existing Advanced Statistical Methods

  – Use physical data to calibrate the computer experimental data and estimate unknown parameters
  – Uses basis functions for computing mean and variance

• Modified calibration of models (Rui Tuo & C.F. Jeff Wu 2013)
  – Modified Kennedy & O’Hagan (2001) – Kernel based, not Bayesian
  – Find parameter which minimizes L2 distance between computer model and “reality”
  – Estimate “real” model from Kernel interpolation and Gaussian Process Prediction

• Recursive Bayesian Hierarchical Modeling (Shane Reese et al 2004)
  – Use computer model outputs and expert opinion to improve estimation and prediction of a physical process

• Hierarchical linear models
  – Remove the variation due to covariates first, then test live vs. sim

• Limitations
  – Complex methodologies limit DoD application
  – Current M&S designs do not support Gaussian Stochastic Process models
  – Focus is on improving prediction, we simply need to validate and state limitations