INSTITUTE FOR DEFENSE ANALYSES

Best Practices for Statistically Validating Modeling and Simulation (M&S) Tools Used in Operational Testing

Kelly McGinnity
Laura Freeman
Rebecca Dickinson

August 2015

Approved for public release;
distribution is unlimited.

IDA Document NS D-5582
Log: H 15-000824

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

The Institute for Defense Analyses is a non-profit corporation that operates three federally funded research and development centers to provide objective analyses of national security issues, particularly those requiring scientific and technical expertise, and conduct related research on other national challenges.

About This Publication

In many situations, collecting sufficient data to evaluate system performance against operationally realistic threats is not possible due to cost and resource restrictions, safety concerns, or lack of adequate or representative threats. Modeling and simulation tools that have been verified, validated, and accredited can be used to supplement live testing in order to facilitate a more complete evaluation of performance. Two key questions that frequently arise when planning an operational test are (1) which (and how many) points within the operational space should be chosen in the simulation space and the live space for optimal ability to verify and validate the M&S; and (2) once that data is collected, what is the best way to compare the live trials to the simulated trials for the purpose of validating the M&S? This conference presentation addresses various strategies for addressing these two questions. The best methodologies for designing and analyzing will vary depending on the goal of operational test, the type of model used in the simulation, and the amount of live and simulated data available.

Copyright Notice

© 2015 Institute for Defense Analyses
4850 Mark Center Drive, Alexandria, Virginia 22311-1882 • (703) 845-2000.

This material may be reproduced by or for the U.S. Government pursuant to the copyright license under the clause at DFARS 252.227-7013 (a)(16) [Jun 2013].

INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS D-5582

Best Practices for Statistically Validating Modeling and Simulation (M&S) Tools Used in Operational Testing

Kelly McGinnity
Laura Freeman
Rebecca Dickinson

Best Practices for Statistically Validating Modeling and Simulation (M&S) Tools Used in Operational Testing

Kelly McGinnity, Laura Freeman, Rebecca Dickinson

Institute for Defense Analyses

August 11, 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 <u>science</u> 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?

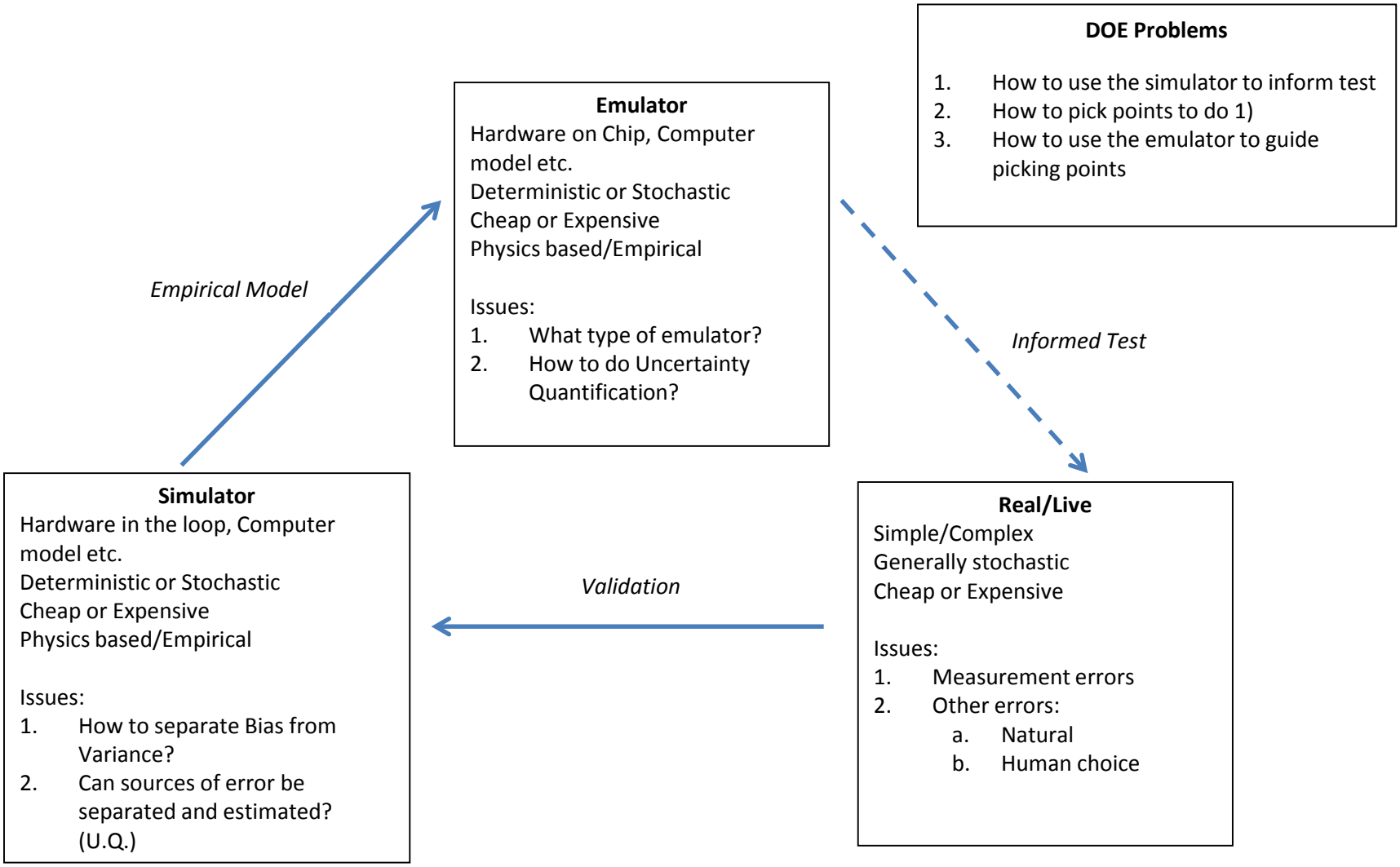
Verification, Validation & Accreditation (VV&A)

- "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."

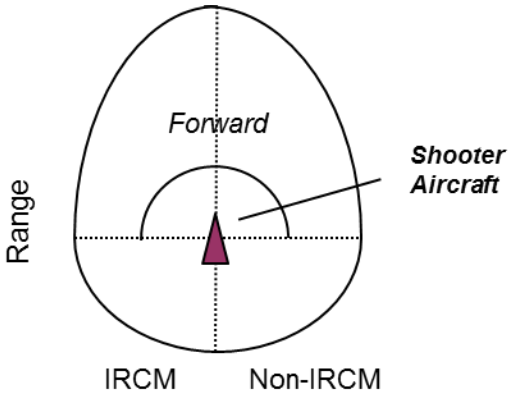
- "Accreditation is the official determination that the M&S is acceptable for its intended purpose."

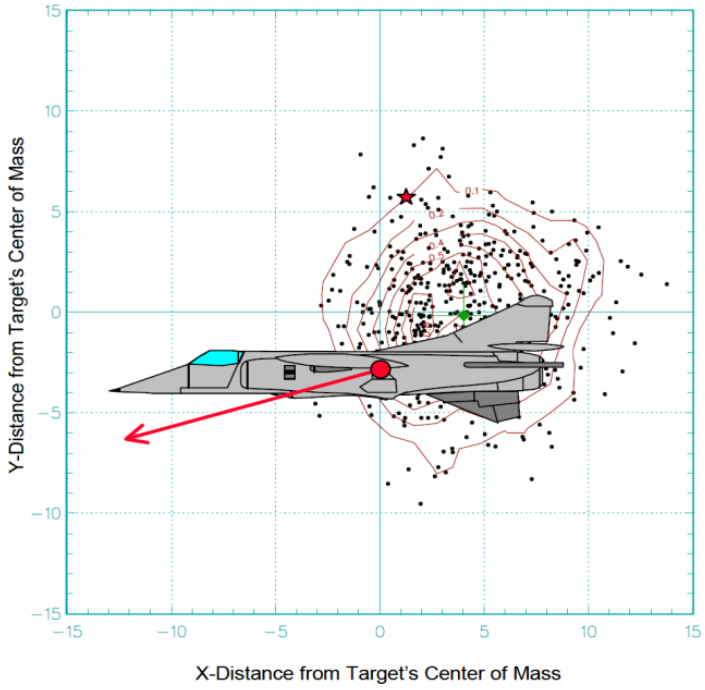
Framework for Validation



Weapons System Validation Example

- **Consider a generic missile program**
  - M&S is heavily relied upon due to a limited number of live fire shots available
- **Interested in the probability of kill (p_k) for reporting**
  - Binary Response
  - P_k requirements are defined in terms of an “egg”. The bins of the egg are defined by Range, Infrared Counter Measures (IRCM) and off-boresight angles (placing you in the foreword or rear hemisphere)
  - P_k requirements in 4 quadrants of the “egg”
- **Interested in miss distance for validation**
  - Continuous response
  - X- and Y- distance from target
- **Factors of interest could include range, IRCM, off-boresight angle, background (weather), time of day, target type, etc.**



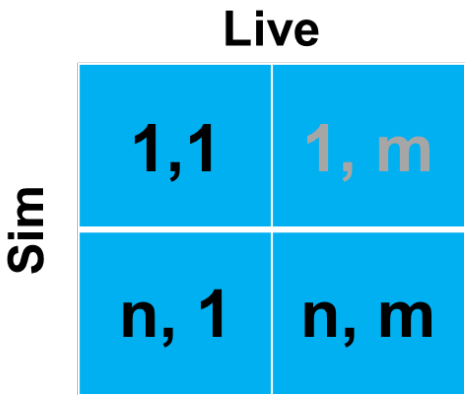


Key Questions

1. What is the best technique for designing the *simulation* experiment?
2. What is the best technique for designing the *live* experiment?
3. What is the best analysis method for *validating* the simulation?

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)

1,1	1, m
n, 1	n, m

- **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 (DoD)

- **Fisher's combined probability test**
  - Combines tail probabilities under each condition using a chi-squared test
- **Regression modeling**
  - Use live vs. sim as a factor in the model and test for significance
- **Logistic regression model emulator for cross-validation and classification**
  - Build a logistic regression model to emulate the simulation; test and update model as live data becomes available
  - Compare prediction intervals from emulator to live data and test for systematic failures

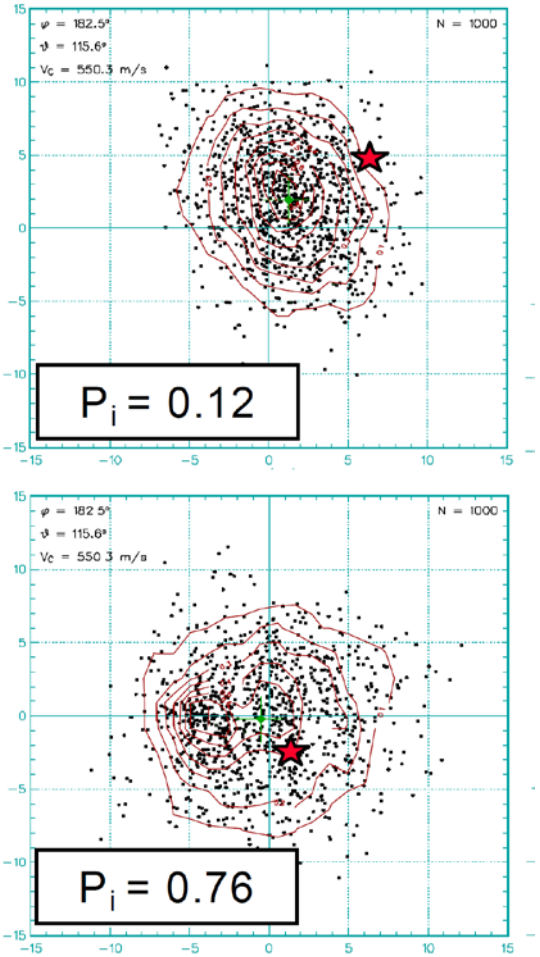
Fisher's Combined Probability Test

1,1	1, m
n, 1	n, m

- **Developed for validating 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
    - » $X = -2 \Sigma \ln(p)$ 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

1,1	1, m
n, 1	n, m

- **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* $= \beta_0 + \beta_1 TestType + \beta_2 TestThreat + \beta_3 (TestType * TestThreat) + \epsilon$

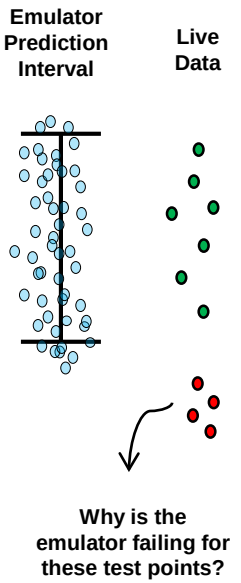
- **Strengths**
  - PRA Testbed runs can be formally compared to the live test events, even when there is limited live data
  - The model allows analysts to test for a Test Type effect, a Test Threat effect, and an interaction effect
    - » If the Test Type effect is not statistically significant then the PRA Testbed runs are providing meaningful data
    - » If the interaction term is significant, there many 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

1,1	1, m
n, 1	n, m

- **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

- **Limitations**
  - Not reasonable in the case of 1 or very few simulation runs per condition

Existing Advanced Statistical Methods

1,1	1, m
n, 1	n, m

- **Gaussian Stochastic Process Models (Johnson et al. 2008, Bates et al. 2006)**
- **Bayesian parameter calibration using GASP (Kennedy and O'Hagan 2001)**
  - 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 predication of a physical process
- **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

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**

Future Work

- **Validation:**
  - Perform rigorous simulation studies to further justify the best analysis method under various situations
- **Design of simulation experiments:**
  - Compare various design types using metrics such as power, prediction variance, and correlation between factors
  - Screening designs (i.e. fractional factorial) may be a good start if there are a lot of factors of interest
    - » Emulator would be a linear model
    - » Can add runs to support a better characterization of the most important factors
  - Space filling designs may be better if higher order terms are of interest or a more detailed characterization of a few factors is needed
    - » Emulator could be a response surface or Gaussian process model
  - Amount of replication depends on the goal
    - » Is it necessary to ensure that the variance of the simulation closely matches the live variation?
- **Design of live experiments:**
  - Need to link analysis method with design of live experiment
  - Must consider the dual objectives of the experiment: model validation and live testing characterization

References

- Sargent, Robert G. "Verification and validation of simulation models." *Proceedings of the 35th conference on Winter simulation*. IEEE Computer Society Press, 2003.
- Oberkampf, William L., and Timothy G. Trucano. "Verification and validation in computational fluid dynamics." *Progress in Aerospace Sciences* 38.3 (2002): 209-272.
- Rao, Lei, Larry Owen, and David Goldsman. "Development and application of a validation framework for traffic simulation models." *Proceedings of the 30th conference on Winter simulation*. IEEE Computer Society Press, 1998.
- Kleijnen, Jack PC, and Robert G. Sargent. "A methodology for fitting and validating metamodels in simulation." *European Journal of Operational Research* 120.1 (2000): 14-29.
- Kleijnen, Jack PC, and David Deflandre. "Validation of regression metamodels in simulation: Bootstrap approach." *European Journal of Operational Research*170.1 (2006): 120-131.
- Rivolo, A. Rex, Fries, Arthur, Comfort, Gary C. "Validation of Missile Fly-out Simulations", IDA Paper p-3697, 2004.
- Thomas, Dean and Dickinson, R. "Validating the PRA Testbed Using a Statistically Rigorous Approach." IDA Document NS D-5445, 2015.
- Law, Averill M. *Simulation modeling and analysis*. Vol. 5. New York: McGraw-Hill, 2013.
- Rolph, John E., Duane L. Steffey, and Michael L. Cohen, eds. *Statistics, Testing, and Defense Acquisition:: New Approaches and Methodological Improvements*. National Academies Press, 1998.
- Easterling, Robert G., and James O. Berger. "Statistical foundations for the validation of computer models." *Computer Model Verification and Validation in the 21st Century Workshop*, Johns Hopkins University. 2002.

References II

- Kennedy, M. C., and O'Hagan, A. "Bayesian Calibration of Computer Models" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 63, 425–464, 2001
- Reese, C. Shane et al. "Integrated Analysis of Computer and Physical Experiments." *Technometrics.* Vol 46 Issue 2: 153-164, 2004.
- Bates, Ron A., et al. "Achieving robust design from computer simulations." *Quality Technology and Quantitative Management* 3.2: 161-177, 2006.
- Johnson, Rachel T., et al. "Comparing designs for computer simulation experiments." *Proceedings of the 40th Conference on Winter Simulation*. Winter Simulation Conference, 2008.
- Tue, Rui and Wu, C. F. Jeff. "A Theoretical Framework for Calibration in Computer Models: Parametrization, Estimation and Convergence Properties." *Preprint (*submitted to *Annals of Statistics*), 2013.

Contact Info

Kelly McGinnity

kmcginni@ida.org

(850) 582-6738