The "Power" of Risk-Based VV&A

Presented to:

Presented by:

Dr. James N. Elele, NAVAIR VV&A SME
David H. Hall, SURVICE Engineering Company

Outline

- What is M&S Credibility?
- M&S Validation and Credibility
  - Relation to System Testing
  - Some Statistical Analysis Techniques
  - A Thought Experiment
- Risk associated with use of M&S
- Risk-Based VV&A Process
- Risk-Based VV&A as Hypothesis Testing
  - The “Power” of VV&A
- Conclusions & Recommendations

What is M&S Credibility?

Credibility = f (Capability, Accuracy, Usability)

Capability: Does it do what you need it to do?

Functional and Fidelity Characteristics

Accuracy: Does it do it well enough for you?

Software (Verification), Data V&V, Outputs (Validation)

Usability: What's out there to keep you from misusing it?

Training, Documentation, User Support

Appropriate Hardware & Software

VV&A as Statistical Testing

- Risk-based VV&A is based on the statistical hypothesis testing process
  - Test for "null hypothesis, H_o" that the M&S represents the system under evaluation well enough for the intended use
- Types of Errors associated with hypothesis testing
  - Type I error – α - rejecting a valid model
    - model builder's risk*
  - Type II error – β - accepting an invalid model
    - model user's risk*
  - Type III error – γ – depending on the source, either asking the wrong question, using the wrong null hypothesis, or getting the right answer for the wrong reason
    - rarely used

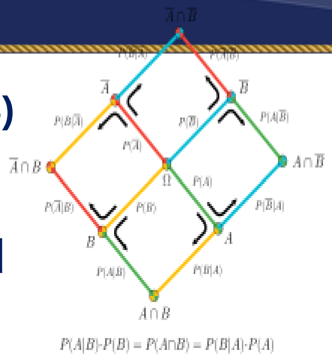
*** Balci and Sargent, 1981**

Relation of M&S Credibility to Validation

- "M&S Validation – the degree to which M&S outputs match the Real World from the perspective of the intended use"
  - There are three ways to represent the real world for comparison to M&S outputs:
    - Benchmarking (or Registration) against another proven M&S
    - Face Validation via SME review (SME interpretation of the real world)
    - Results Validation via SUT testing (facility/laboratory/field test)

- Statistical analysis of test data comparisons (Results Validation) generally is considered the preferred way to establish M&S credibility
  - Statistical techniques are very useful for analyzing test data to eliminate: Biases, Autocorrelation, Errors in instrumentation, etc.
  - Statistical comparison with test data can't guarantee that there are no M&S errors
    - The most that statistics can say is that there are insufficient data to reject the null hypothesis that the model matches the data
    - Which means that often we accept a "bad" model if we rely <u>only</u> on comparison to test data
    - Often we don't have enough test data to reject a "bad" model

Bayesian Statistics



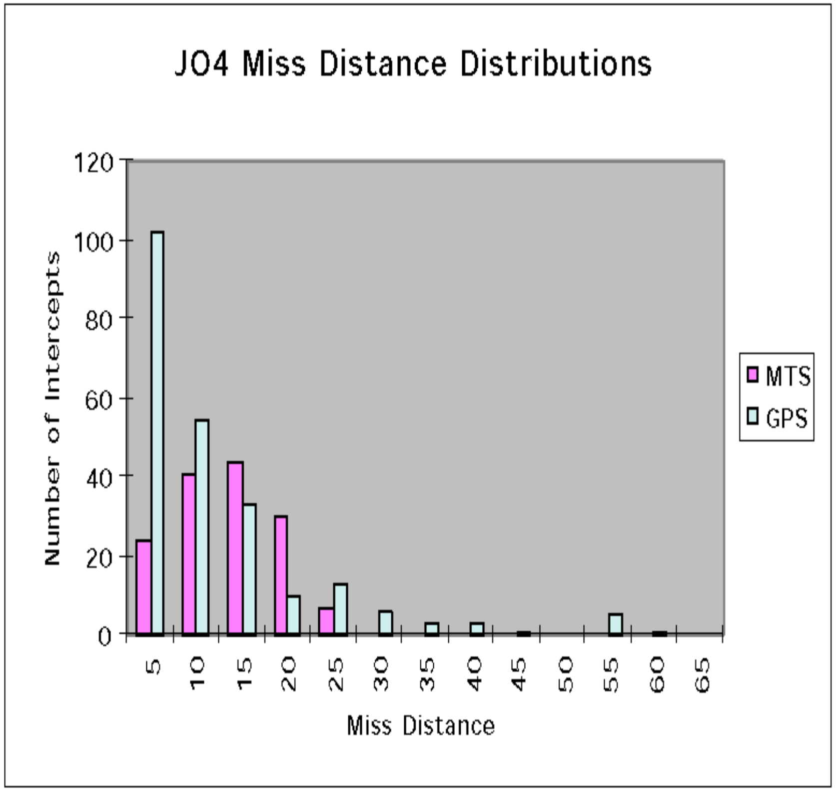
P(A/B) = [P(B/A) * P(A)] / P(B)

- Bayesian statistics are based on Bayes' Theorem
  - Approach allows for use of a prior (assumed) probability and distribution in statistical analysis of observed data
  - Can use SME opinion, prior test data, M&S results, etc. to develop the prior distribution – helpful with small test samples
- A good example is given in a recent ITEA Journal*
  - For analysis of system reliability in support of OT&E
  - An example where often there is little data available from testing to support reliability claims
- Bayesian statistics could prove a useful tool for M&S Results Validation where there may be insufficient data for standard statistical techniques
  - Due to the cost of testing, such as ship survivability estimates
  - Due to the difficulty of testing, such as nuclear systems

***The ITEA Journal of Test and Evaluation* 2016; 37: 298-305, Kassandra Fronczyk, Ph.D. and Laura Freeman, Ph.D., Institute for Defense Analyses (IDA)**

Complicating Things: Analytical vs. Statistical Significance

- **Example: Miss distance distributions for the same missile system from two different sources**
- **<u>Statistically different</u> - Rayleigh vice Poisson**
- **<u>Analytically Equivalent</u> - Overlapping Effectiveness Confidence intervals**
  - **E_{MTS} = (0.72, 0.76), μ= 0.74**
  - **E_{GPS} = (0.74, 0.80), μ= 0.77**



- **Based on statistical comparisons we would have rejected the null hypothesis that they were from the same distributions**
  - **But in terms of the intended use they are equivalent**
- **Statistical significance is not the same as analytical significance**

Results Validation Must be Done in Terms of the Intended Use

Testing for Intervals

- For validating simulation and analytic stochastic models*
  - Hypothesis testing when the desired amount of model accuracy is specified
  - Provides for the model to be accepted if the difference between the system and the model outputs are within a specified range of accuracy
    - H_0: Model is valid for the acceptable range of accuracy under the set of experimental conditions
    - Asks: Is the difference, D, between sample mean and population mean within the required accuracy?
    - Sargent used Student-t distribution to test H_0
  - May have limitations where sample range is close to acceptable accuracy bounds: In that case a larger sample is required
- This technique would link analytical significance to statistical significance
  - By using a range of accuracy determined by the intended use

*A New Statistical Procedure for Validation of Simulation and Stochastic Models
Robert G. Sargent, *Syracuse University, Department of Electrical Engineering and Computer Science*, rsargent@syr.edu
SYR-EECS-2010-06

Issue: Validation Data are Expensive!

- Many DOD programs do not have funds to generate statistically significant testing for M&S validation
  - Some programs can only afford one test!
- Some programs cannot conduct “all-up” testing at all
  - Due to regulatory or other constraints
- System tests are focused on demonstrating system performance and not on M&S validation
  - Most programs do not coordinate with modelers to ensure test data are adequately instrumented for comparison to M&S
  - Programs that do coordinate validation needs with testing often do not collect a sufficient sample size for statistical significance
- DOT&E has the “hammer”
  - Systems must pass OT&E muster to be fielded, so if they want to use M&S as part of the OT&E process they will have to pony up the funds for M&S validation

A Thought Experiment

- Let's say a missile development program wants to use a missile flyout simulation to demonstrate the system's capability across its launch envelope
  - Usual metric is "miss distance" for comparison to test data, although a number of other factors can be as important in determining effectiveness
- A test program is designed to fire the missile at various points within the envelope and compare to M&S results
  - Test program is naturally limited by cost, scheduling, system availability, test range availability, etc.
    - Almost always results in a smaller test sample than the program would like to have from a statistical standpoint
- What are the possible outcomes of such a test program?
  - H_o: the M&S miss distribution is the same as the test data distribution

Possible Experiment Outcomes

M&S Result	Test Result	Result	Comments
Large σ	Small σ	Accept Ho	test data fall within M&S distribution (depends on statistical test used, but confidence intervals clearly overlap)
Large σ	Large σ	Accept Ho	
Small σ	Small σ	Accept Ho	
Small σ	Large σ	Reject Ho	depending on the size of the difference and the overall sample size – smaller samples would likely have us <u>accept Ho</u>

ISSUES:

- Depending on the intended use, we probably would have wanted to reject Ho in case 1 – M&S as "catch-all" distribution – plus calculation of missile effectiveness would be wrong
- We accept Ho in three out of the four cases, and probably in all four cases if the test sample size is small
  - We're minimizing the likelihood of rejecting a good model, **while increasing our likelihood of accepting a bad one**
  - Testing of this nature may not be giving us the result we're hoping for (rejecting bad models)

Goodness of Fit Approaches

- A number of techniques exist to compare distributions for “goodness of fit”
  - Chi-Square, Kolmogorov-Smirnov (non-parametric), etc.
  - Fisher’s combined probability test
- Fisher procedure has been used for situations similar to our “thought experiment”*
  - Fisher is essentially a Chi-Square test on a natural-log transform of the original distribution
  - Fisher test can have higher power than others, but that is not uniformly the case
  - It has advantages and disadvantages (one being that a single outlier data point can cause H_o to be rejected)
- GOF tests can allow for data that are sparsely distributed across a multi-dimensional space*
  - Somewhat alleviate the problem of sparse data, but still designed to minimize the likelihood of rejecting a “good” model (α)

**** Another ‘New’ Approach For ‘Validating’ Simulation Models*, Arthur Fries, PhD, Institute for Defense Analyses**

The "Power" of Hypothesis Testing

DECISION	TRUTH	
	H_0	H_A
Accept H_0	Correct $P = 1 - \alpha$	Incorrect (Type II error) $P = \beta$
Reject H_0	Incorrect (Type I error) $P = \alpha$	Correct $P = 1 - \beta$

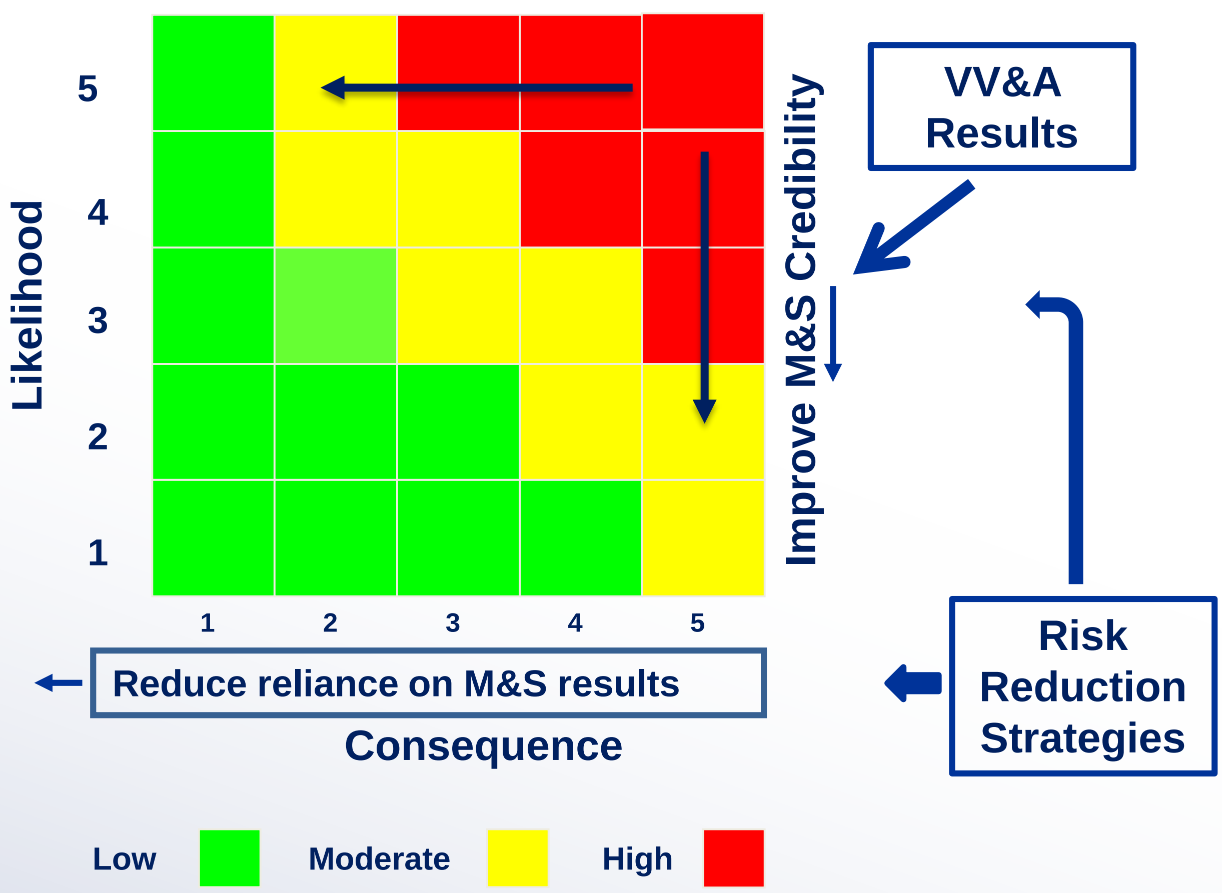
- **You can't drive down both Type I and II errors at the same time – push down one, the other pops up**
  - **Like our jury system – "innocent until proven guilty beyond a reasonable doubt" is designed to minimize the likelihood of convicting an innocent person (Type I), but raises the likelihood of releasing a guilty one (Type II)**
  - **Reducing the likelihood of rejecting a credible model increases the likelihood of accepting an invalid one**
- **Since statistical tests minimize the probability of rejecting a true hypothesis (α), the "Power" of the test is defined as the probability that we correctly reject a false hypothesis: Power = $1 - \beta$**

So What Do We Do About the Risk of Using “Bad” Models?

- VV&A Is a Risk Reduction Process
- Risk associated with M&S:
  - Loss incurred when an erroneous M&S result is used to make a decision
    - Technically, Risk = Likelihood x Consequence (DOD Definition)
    - **Risk = expected value of loss**
- Nature and extent of information required to support accreditation decision ***<u>is based on an assessment of risk to the intended use</u>***
  - Role of M&S results in decision making process
  - Importance of decision that M&S is supporting
  - Severity of the consequences of making incorrect decisions because M&S results were wrong
  - Likelihood that M&S results are wrong

V&V Is a Process; Accreditation Is the Decision; Risk Is the Metric

Reducing the Risk of M&S Use



Risk = Likelihood x Consequence

VV&A as Minimizing β

- Standard Statistical Hypothesis Testing is based on techniques designed to minimize a Type I error (α)
- Issue: the tester cannot generally control (or often even estimate) Type II error (β) at the same time
  - Type II error is accepting an inadequate ("bad") M&S
  - The VV&A process aims at reducing these errors
  - VV&A becomes the "hands-on" practical way of minimizing β
  - Some V&V results are objective, but most are subjective judgements of SME
- The "Power" of a VV&A process, therefore, is how well it ensures that we <u>reject an inadequate M&S </u>= 1 - β

Risk-Based VV&A Process

- Determine risk associated with using wrong M&S outputs
  - Based on the intended use of the M&S
  - Rejecting a “valid” M&S (Type I error) results in opportunity lost and additional cost
  - Accepting an “invalid” M&S (Type II error) can result in substantial risks depending on how much the user relies on its output
- Conduct V&V focused on M&S intended use <u>based on the risk involved</u>
  - Develop Information (capability, accuracy, usability info) needed to reduce risk of M&S use for the application
- NOTE: Validation (comparison to test data) is only one piece of the puzzle
  - What do you do if you can’t get validation data?
  - How do you compare validation data with M&S outputs?
  - What do you do if you can’t get enough validation data for a statistically significant comparison?
  - Who decides you’ve done enough, and how?

M&S Risk Characteristics & Criteria

10 Risk Characteristics	Risk Acceptance Criteria	Rating
	Capability	
Intended Use & Acceptability Criteria	Clearly Articulate requirements and criteria	G
Conceptual Model Validation	The conceptual model is complete and documented	Y
Model Fidelity	The model's Functions, Entities and Data are documented and appropriate for the intended use.	R
	Accuracy	
Design Validation	The algorithms and their applications are correct and valid.	Y
Input and Embedded Data	Data are credible, and subject to review and revision.	G
System Verification	M&S demonstrated to accurately represent the specific intended use(s) and requirements.	Y
Output Accuracy	**The M&S outputs have been compared with known or expected behavior and are sufficiently accurate for the specific intended use(s).**	**Y**
Configuration Management	A sound written Configuration Management (CM) Plan.	Y
	Usability	
Documentation	Documentation is readily available, up-to-date, and complete.	Y
User Community	User support and documentation is adequate to ensure proper use	G

The "Power" of Risk-Based VV&A

- Statistical analysis of test data comparisons to M&S results is only one component of establishing M&S credibility situations
  - SME review of M&S <u>sensitivity analyses</u>, Benchmarking against other M&S and comparisons with graphical representations of test results
  - Assessment of documentation, conceptual model validation, user support functions, etc.
- Statistical techniques are most useful for analyzing test data to eliminate biases, autocorrelation, errors in instrumentation, etc.
- By focusing on <u>all</u> aspects of M&S Credibility (capability, software accuracy, data accuracy, output accuracy, and usability) <u>we can help minimize β as well as α</u>
  - Comparison to test data helps to make sure we don't reject a "good" model
  - <u>Everything else</u> helps to make sure we don't accept a "bad" one <u>and</u> highlights where it is "bad"

Recent DOT&E Guidance*

- In addition to 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.
- M&S validation is a complex process and there are <u>many important elements that provide useful information</u> and can be used in conjunction with statistical modeling:
  - Documentation review
  - Face validation
  - Subject matter expert (SME) evaluation
  - Comparison to other models (benchmarking)

***Clarifications on Guidance on the Validation of Models and Simulation used in Operational Test and Live Fire Assessments, Jan 17, 2017, Director Operational Test and Evaluation**

Conclusions & Recommendations

- Risk-based M&S VV&A is a practical, hands-on method for minimizing the probability of Type II errors (β)
  - Identifies and documents M&S requirements, acceptability criteria and metrics for the application and prioritizes V&V activities that are most cost-effective
- Uses statistical techniques for comparison to test data where feasible
  - Bayesian, testing for intervals, and GOF tests are promising when test data are sparse
  - Still need to focus on requirements of the intended use
- Uses other methods to augment comparisons to test data
  - Sensitivity analyses, graphical methods, benchmarking
  - These methods meet recent DOT&E guidance

Overall, Risk-Based VV&A is a (sometimes subjective) statistical analysis approach to ensuring that we not only accept a “good” model but also reject a “bad” one

Questions?

An Aside: A Conceptual Statistical View of Risk-Based VV&A

- The 10 risk-characteristics can be viewed as a sample from the total number of characteristics that describe risks of M&S use
- Since there are three possible ratings for each, and without any prior information we assume they are all equally likely, the null hypothesis is that the risk ratings follow a binomial distribution with probability of a correct rating p=1/3:

$$P(\text{Type I error}) = \sum_{x=S^*}^{n} \binom{n}{x} \left(\frac{1}{3}\right)^{x} \left(\frac{2}{3}\right)^{n-x} = [1-\mathbf{P}(S \le S^* \mid H_0 : \mathbf{P} = \tfrac{1}{3})]$$

  - Where S* is the number of characteristics we can get wrong and still get the correct overall result (total overall risk assessment)
- From the binomial distribution, we can compute the power of the test from standard formulas, given any other assumed true value for p
  - β is a function of n, p and α