SAND2009-5430P

Risks and Metrics in Influence Ops Modeling

For Air Force Influence Operation Modeling

George Backus
gabacku@sandia.gov
Rich Hills
rghills@sandia.gov
Tim Trucano
tgtruca@sandia.gov

August 26, 2009

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Sandia National Laboratories

Definitions

Verification:

- "Verification is the process of determining that a model implementation accurately represents the developer's conceptual description and specification …"

Validation:

- "Validation is the process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended use of the model …"

[DoD, DOE, AIAA, ASME, IEEE, etc …]

Accreditation:

- "The official determination that an M&S application and its associated data are acceptable for use for a specific purpose …"

[DoD]

V&V in support of Individual, Organizational, and Social (IOS) modeling: broad issues

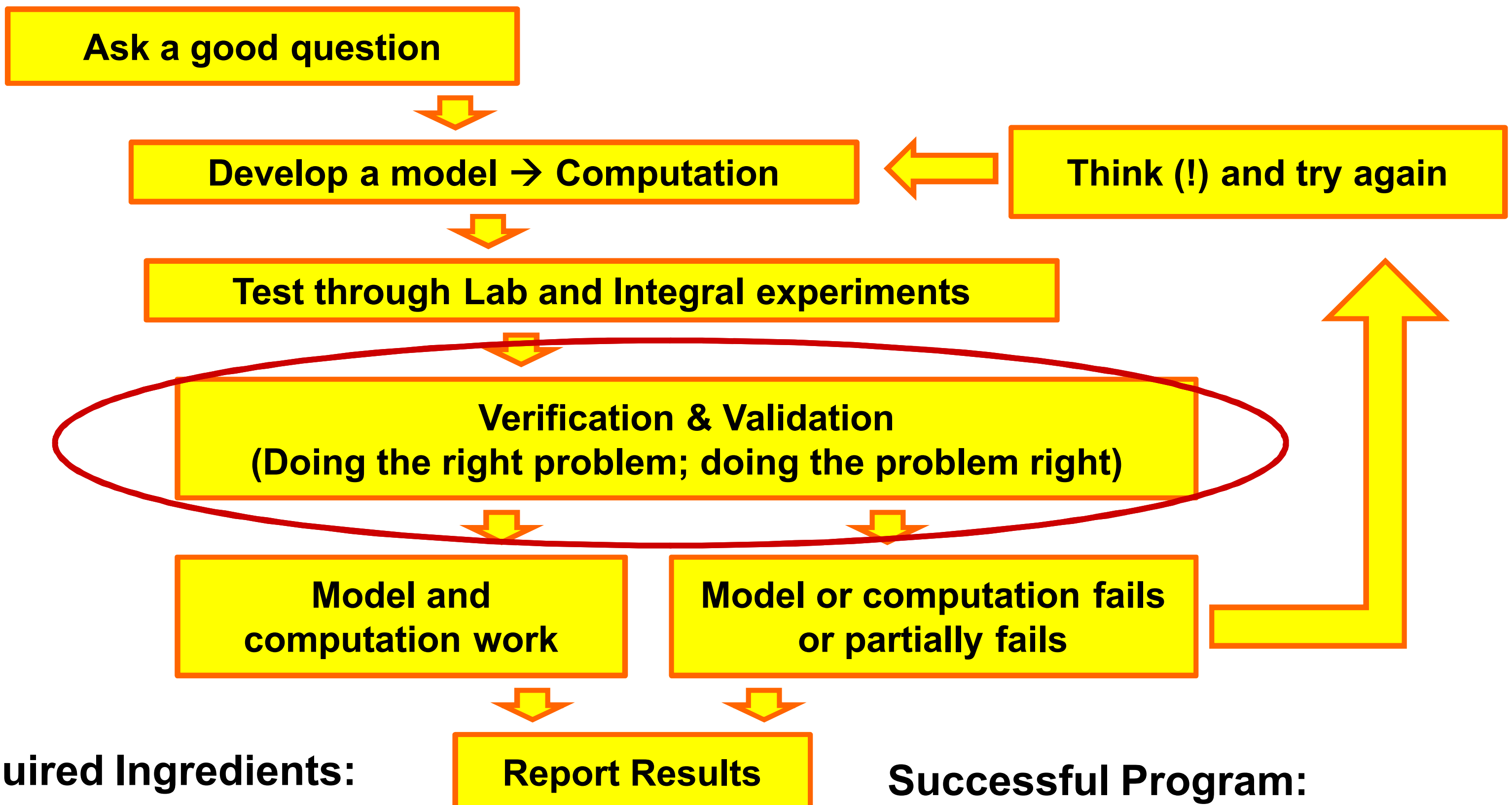
Taxonomy of errors of concern:

- Type 1 – Model "right," believed wrong
- Type 2 – Model "incorrect," believed right
- Type 3 – Wrong problem solved
- Type 4 – Model results used incorrectly

Sharpening thinking about V&V:

- Clarify and emphasize the need for pragmatism relative to the intended application
- Rigorous and clear decisions about appropriateness of identified benchmark observational data for IOS model validation
- Rigorous decisions about IOS model credibility for intended applications
- Programmatic integration of V&V with IOS model development

Koonin – "[IOS] Computation as a tool in <u>science</u>"



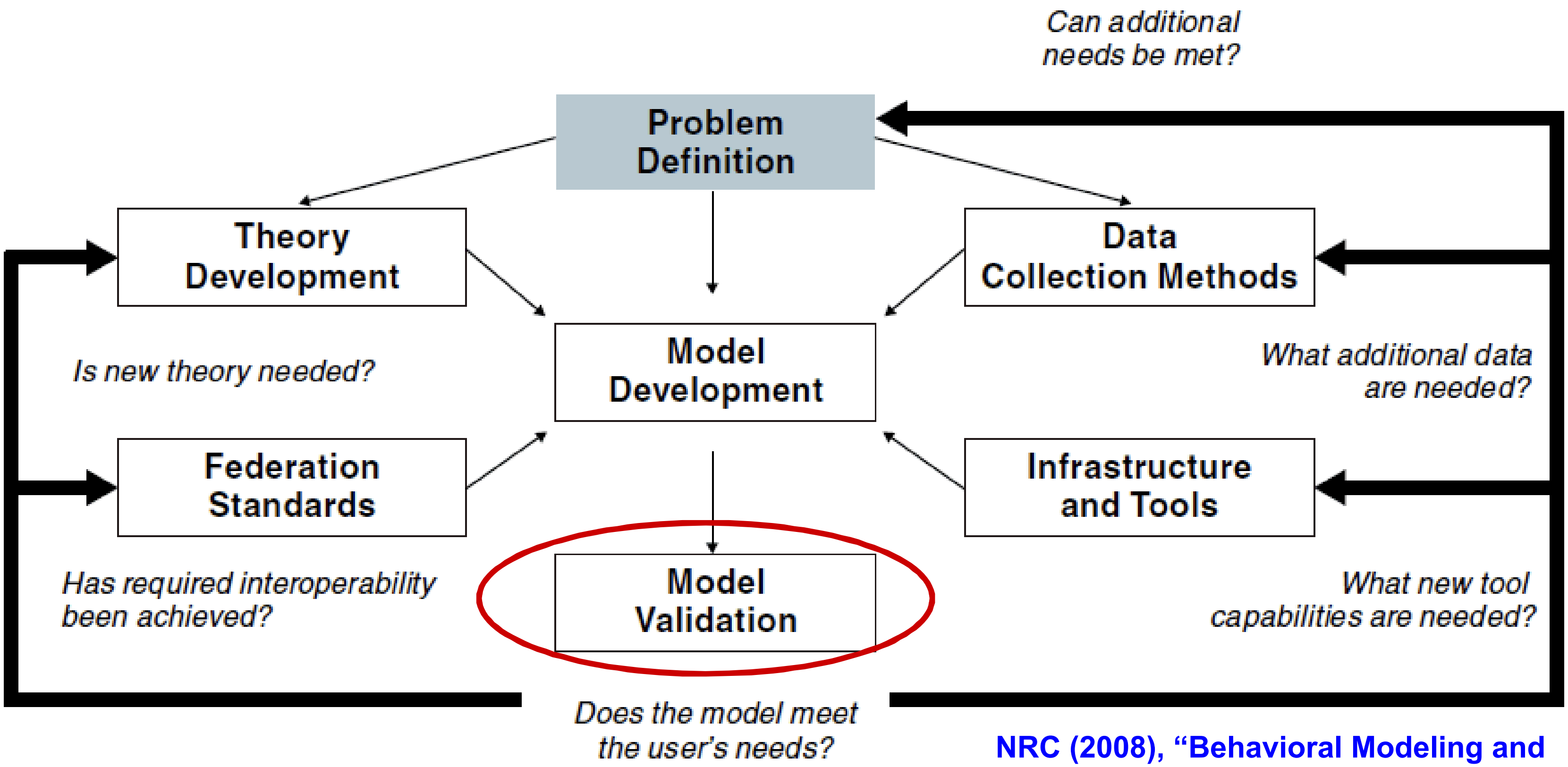
Required Ingredients:

1) Theorists
2) Computational Scientists
3) Experimentalists
4) Applied Mathematicians
5) Computer Scientists

Successful Program:

1) Guides experiments
2) Quantifies uncertainties
3) Yields solutions/insights
4) Eliminates tunable parameters

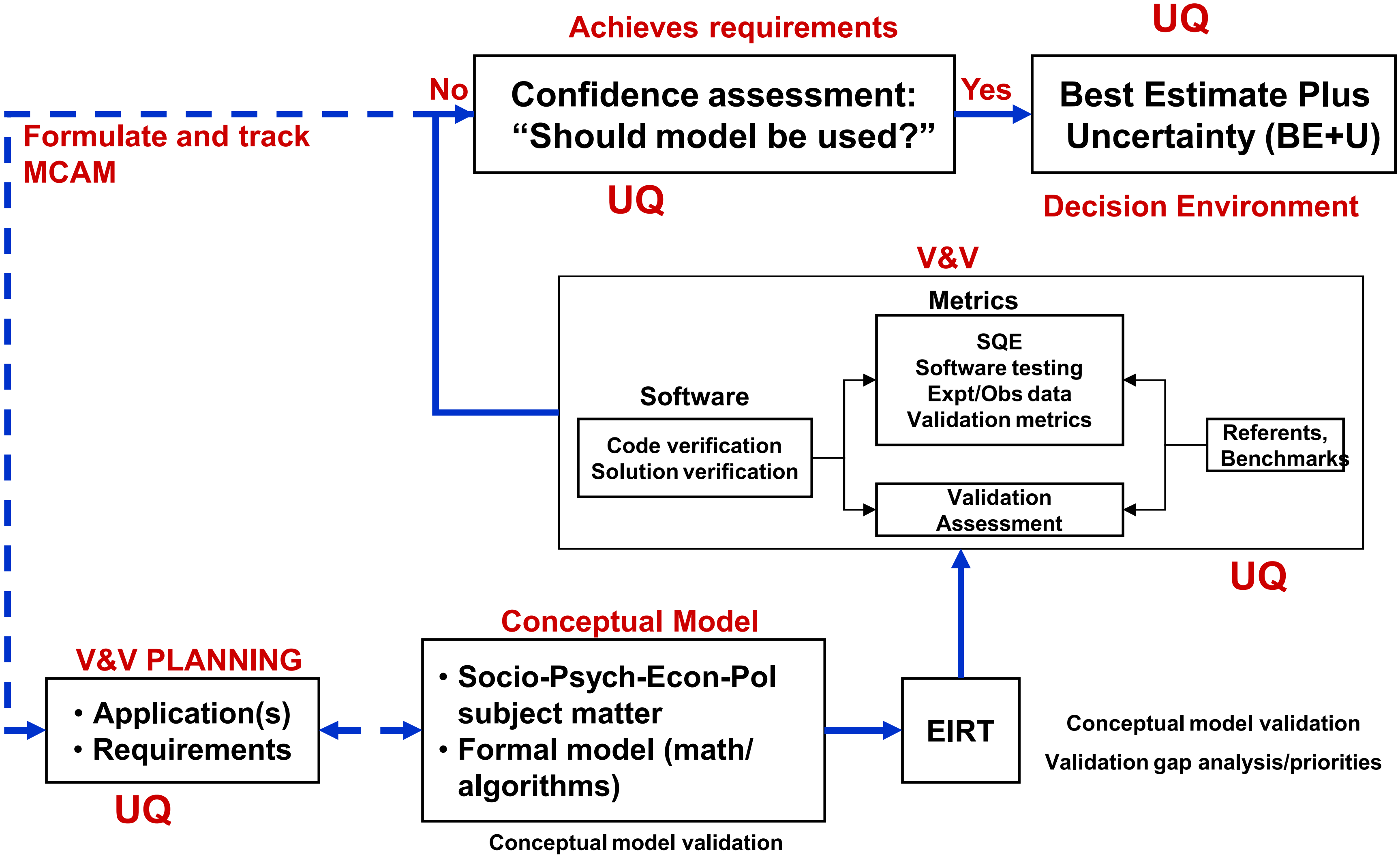
NRC – "IOS Computation as a tool for Action"



NRC (2008), "Behavioral Modeling and Simulation: From Individuals to Societies," National Academies Press

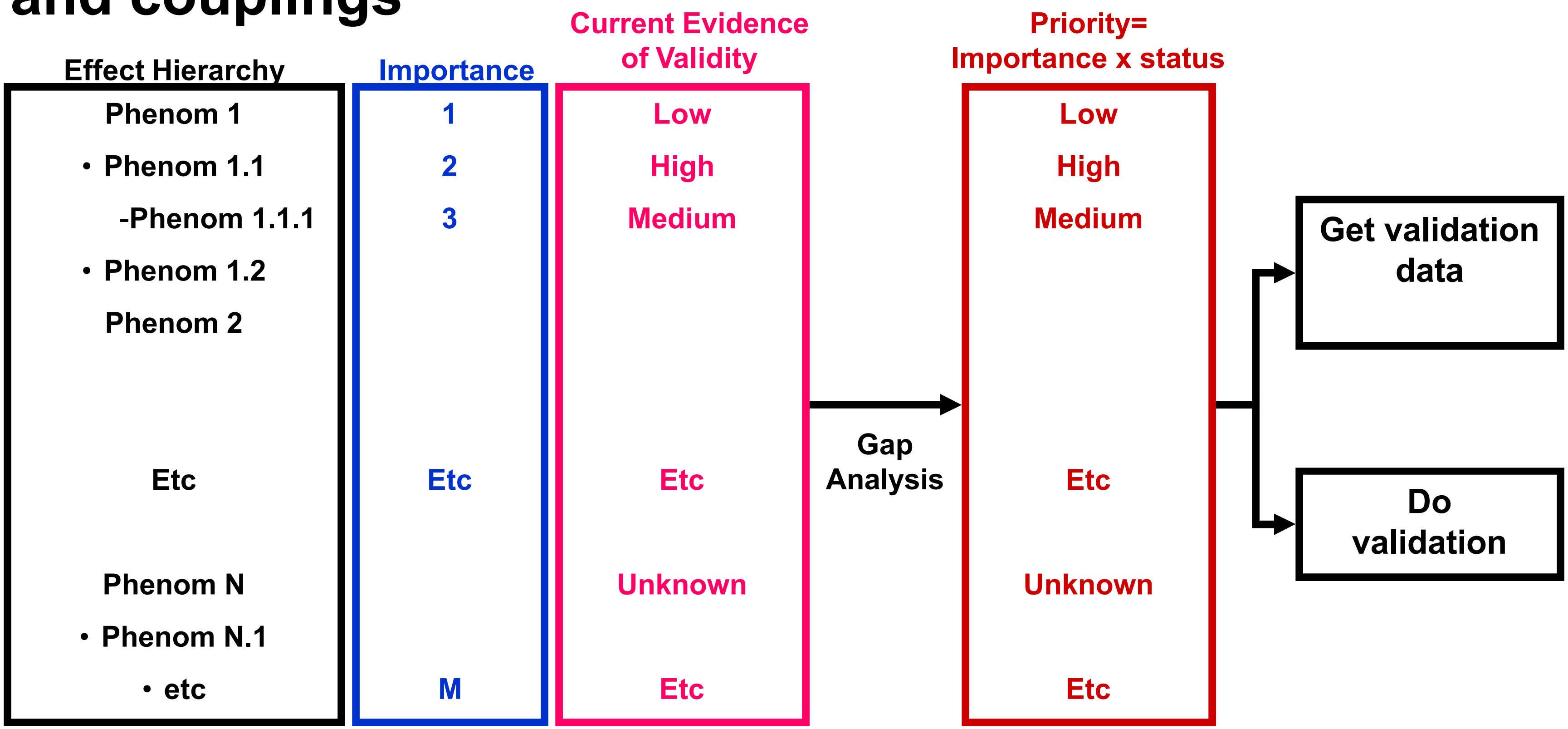
V&V is a community of practice involving "customers," users, domain experts, and model developers. Rarely are all these roles embodied in one individual or one center of expertise.

Embedded V&V for IOS: A Notional Methodology



L. A. McNamara, et al. (2008), "R&D for Computational Cognitive and Social Models: Foundations for Model Evaluation through Verification and Validation," SAND2008-6453.

Effects Identification and Ranking Table (EIRT): Social-Economic-Psychological-Political mechanisms and couplings



"Hierarchical" assumes effects AND couplings are identified.

These priorities are relevant for verification as well

- The EIRT is an adaptable V&V planning element (called a "PIRT" elsewhere) that should be periodically reviewed and modified as needed
- THE EIRT also guides V&V of the conceptual model

Uses of EIRT in effects modeling and simulation

- EIRT helps specify research directions, test regimes, and support of conceptual model
- Generalized (or generalizable) EIRT can accommodate new phenomena under evolving realization of conceptual limitations
- EIRT is useful in research as well as production applications
- EIRT aids project management and planning
- Prominence of EIRT is NOT BASED on the assumption that IOS "reality" is hierarchical – it IS BASED on the assumption that the IOS MODEL (equations, algorithms, software) is hierarchical

Uncertainty Quantification (UQ) is pervasive

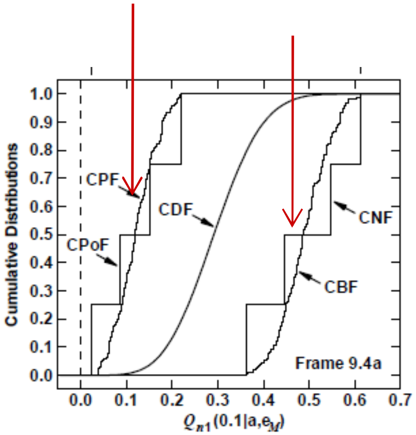
UQ in generation, reliability assessment (V&V), delivery and application of model results is a requirement for high-consequence applications

- UQ is a technical challenge
- The goal is to achieve reliable Best Estimate Plus Uncertainty in all modeling, which is difficult even in "well-understood" modeling applications (like computational physics)
- UQ is a dominant factor when modeling results are delivered for decision support (this does not imply the decisions are "model-based")

UQ is unavoidable in specifying the domain of application

UQ is required to avoid bad answers and bad decisions, and to understand the inadequacy of "point" predictions.

UQ has major technical complexities, certainly for IOS models.



Ex. Non-probabilistic UQ may be required for qualitative inference (epistemic uncertainty dominated)

Sandia National Laboratories

Metrics align with "Validation of the model for the intended application"

Given a specified Domain of Application (DA) of the intended models for Influence Ops, metrics conform to evolving V&V characterization of model implementation and results.

All V&V is driven by the DA

- Typical phrase: "Validating a model for its intended application" (which is more precisely interpreted as the phrase "Acquiring evidence that the model is valid for its intended application)

The Domain of Validity (DV) is usually a subset of DA

- The relationship of DV to DA is crucial. For example, this relationship helps clarify the meaning of extrapolation and going beyond the validation evidence base in application of models.

Detailed specification of the DA defines the requirements basis for V&V.

- V&V is requirements centered: No requirements = no V&V!

Sandia National Laboratories

Users' views of "successful models" are key:

DA reflects the importance of users in at least three model dimensions that must be acknowledged in the project as a whole and V&V in particular:

- "Useful," (the user thinks its useful) for example:
  - Improved situational awareness
  - Expands knowledge structure of users
  - Effective in supporting decision making
- "Usable," (the user will use it) for example:
  - Compatible with the user environment
  - Provides usable and useful results given the ability to acquire data
- "Reliable," (the user will trust it) for example:
  - Provides repeatable, robust, credible, etc results
  - Clearly communicates prediction accuracy relative to limitations and accumulated uncertainties of the model

Metrics: Measuring the implementation, progress, and impact of V&V

The tool we intend to develop and implement for this purpose is the Model Capability Assessment Matrix (MCAM – also called the Predictive Capability Maturity Model in the computational physics and engineering context).

- W. L. Oberkampf, et al. (2007), "Predictive Capability Maturity Model for Computational Modeling and Simulation," Sandia National Laboratories, SAND2007-5948.
- NASA (2008), *Standard For Models And Simulations, National Aeronautics and Space Administration Technical Standard*, NASA-STD-7009.

This tool is a component of the suggested surety framework of Peercy, Shaneyfelt, et al. that contributes to risk management.

This tool is also compatible with existing approaches, such as assessment templates and other formal records.

- MCAM can organize information for use in these records or vice versa.

Model Capability Assessment Matrix (MCAM)

Increasing evidence (reliability for intended application) →

	1	2	3	4
Subject Matter Completeness				
Verification				
Validation				
UQ				
Socio-political-economic-psychological sophistication				
"Transitions"				

Measures completeness and quality of IOS model capabilities

- The MCAM table acts as a "confidence" measure, via the numerical metrics.
  - Benchmark levels and required targets can be defined, progress tracked, etc.
  - For example: "1" – Exploratory; "4" – "High Consequence Use" (accredited)
- "Subject Matter Completeness" links to the EIRT (Effect Identification and Ranking Table), and therefore to (1) V&V planning; (2) articulation and validity of conceptual model (an SME point of contact also)
- "UQ" is a separate element in the table because of its link to the application of the model (for example to robust decision making).
- The DA directly imprints on the elements, but especially "Subject Matter Completeness" and "S-P-E-P sophistication"

Ex. Detailed form of EIRT from Hills' work

	1	2	3	4
Subject Matter Completeness				
Data				
Physically-based models				
HSCB-Based models				
Verification				
Federated Model Validation				
Uncertainty quantification				
Documentation/archiving				

Domain of application incompletely and informally defined. Significant potential for mission creep for the model during development and assessment process.

The SQE process is "managed." <u>Managed</u>: Detailed measures of software process and product quality are collected. Both the software process and products are quantitatively understood and controlled.

Limited comparisons of models to validation data.

- Level 1 – Low IOS model impact, e.g. scoping studies, research models
- Level 2 – Some IOS model impact, e.g. preliminary product use, classroom training
- Level 3 – High IOS model impact, e.g., significant influence on planning, decision-making
- Level 4 – High IOS model impact, e.g., strong influence on planning of high risk operations, decision-making based on qualification or certification of product use

A level profile (green) could be variable over MCAM characteristics

The primary purpose of V&V (and A) is risk management

Risk management character of V&V centers on:

- Using V&V considerations to guide project planning
- Acknowledge the constraints R&D risks imply for delivered V&V results and approaches
- Recognizing the specification of the DA drives V&V risks

Much risk management in V&V centers on defined answers to key questions associated with the use of benchmarks (“referents”):

- What is the technical specification of comparing models to the benchmark?
- What does “good” and “good enough” agreement mean?

Approaches for V&V-centric risk management:

- D. E. Peercy, W. L. Shaneyfelt, E. O. Caldera, and T. P.Caudell (2008), “A Surety Engineering Framework to Reduce Cognitive Systems Risks,” SAND2008-7968 – LDRD-funded research project.
- L. A. McNamara, et al. (2008), "R&D for Computational Cognitive and Social Models: Foundations for Model Evaluation through Verification and Validation,“ SAND2008-6453.

Risks in Specific Project Tasks:

1. Phase I, Task 1: There is no technical risk, but there is potential timing risk if SMEs are not incorporated early.
   – "Background research … understand problem domain"
2. Phase I, Task 2: No V&V-related technical risk, but highly dependent on timely SME review.
   – "Initial software infrastructure … exploratory implementation … sensitivity analyses and confidence metrics …"
3. Phase I, Task 3: The primary risk is establishing benchmark criteria associated with real observational data.
   – "Develop and document V&V methodology"
4. Phase II, Task 1: Indirectly related to V&V.
   – "… collect information … create representative models …"
5. Phase II, Task 2: The primary risk is establishing strong benchmarks.
   – "… apply V&V methodology to several model components …"

V&V Risks From a Broader Perspective:

1. Poor specification of the Domain of Application of the intended models for Influence Ops
2. Poor V&V specification, implementation and/or results
3. Poor UQ specification, implementation and/or results
4. Lack of adequate data to calibrate, validate, and perform UQ, including the characterize of data uncertainties

The surety framework that Peercy et al. have developed is relevant to managing risks in general as well as V&V-related risks.

- **Our approach to V&V, in particular the use of the MCAM (constructed as "Cognitive System Maturity Model" in the surety framework), specifically integrates with the surety framework.***

Surety Methods

Safety Principles*
Reliability*
Risk Analysis*
Quality Methodology*
Monitoring Efforts
Cryptographic Security
System Security
Design
Public Communication
Redundancy
System Backups
Open Discussions

Cognitive System Maturity Model

Maturity Level → / Attribute ↓	Level 0 Low Consequence, Minimal Impact, Scoping Studies & Research Models for Understanding	Level 1 Moderate Consequence, Some Impact, Preliminary Product Experimental Use	Level 2 High-Consequence, High Impact, Decision Making Based on Controlled Product Operational Use	Level 3 High-Consequence, Decision-Making Based on Qualification or Certification of Product Use
Psychological Representation Are important functional features neglected because of simplifications or stylizations?				
Physiological Representation How fundamental are the physics and material models and what is the level of model calibration?				
Environmental Representation Are normal, abnormal, hostile environments represented?				
System Surety Engineering Are reliability, safety, security, and V&V methods applied to identify potential areas of risk?				
Ethics, Legal, Societal How are ELS issues understood, analyzed, and addressed?				
System Risk Mitigation How are gaps/vulnerabilities analyzed and risk mitigations implemented?				

Sandia National Laboratories

Summary thoughts

There is no V&V if there are no application REQUIREMENTS.

At best, a model is valid only within its Domain of Application.

- In reality, a model is valid only on a subset of its Domain of Application.

We view V&V as an integral collaborator in the modeling project from Day One, rather than an “assessor” at some delivery point in the future.

- V&V facilitates management and execution of model development, implementation and application.

Independent V&V (and A) “assessment” as specific project milestones is supported within this perspective.

Targets in the MCAM are important.

Preconceived, opinion-based benchmarks are useless

- Evidence-based benchmarks are required.

Modeling success requires observational, statistically significant data benchmarks.

Supplemental Slides

Complexity of Validation Regimes

Exploratory Processing	R&D Extensions	High-Consequence Application
Falsifiability Operations*		Multi-Attribute Metrics
Establish Legitimacy		Establish Confidence
Structural Assessment	Econometric Tests	Bayesian Convergence
Visual-Turing Comparison	Theil Statistics	Distributional Convolution
Dimensional-Consistency	Cointegration Analysis	BE+U Evaluation
Integral/Derivative Check	Epistemic/Aleatory Characterization	Engineering Index
Variable/Parameter Independence	Uncertainty Quantification	Data Validation
Doman Review	Performance Thresholds	Predictive Validation
Data Validation	Models Comparison	Schellenberger's Criteria
Black Box Validation	White Box Validation	Boundary Adequacy
Degenerate Test	Historical Data Validation	
Events Validity	Traces	
Face Validity	Confidence Intervals	
Extreme Condition Tests	Sensitivity Analysis	
Behavioral Anomaly	Replicability	
Sensitivity Analysis	Documentation Integration	
Documentation Integration	Boundary Adequacy	
Parameter Assessment	Structural Variation	
Family Member	Surprise Behavior	

*Falsifiability, as used here, implies a metric to guide conceptual modification to maximize validity.

Validation for "exploration" versus validation for "action"

Poor specification of the Domain of Application (DA)

All V&V is driven by the DA

- Typical phrase: "Validating a model for its intended application" (which is more precisely interpreted as the phrase "Acquiring evidence that the model is valid for its intended application)

The Domain of Validity (DV) is usually a subset of DA

- The relationship of DV to DA is crucial. For example, this relationship helps clarify the meaning of extrapolation and going beyond the validation evidence base in application of models.

Detailed specification of the DA defines the requirements basis for V&V.

- V&V is requirements centered: No requirements = no V&V!

DA plays out in at least three model dimensions that must be acknowledged in the project as a whole and V&V in particular:

- "Useful"
- "Usable"
- "Reliable"

Risk Management for DA:

Specification and documentation of the Domain of Application must be accomplished as early as possible in the project.

- The initial specification of DA is not fixed in stone, so a process should be developed to periodically review and evolve the DA specification
- Specific requirements must be identified and documented

Poor V&V:

V&V can be poor because:

- It is specified (planned, defined) poorly
- It is implemented (executed) poorly
- It achieves poor results
- Or all three

Example: V&V can be well-planned, and executed well, and be poor because the resulting domain of validity is too small to lend significant confidence to the reliable application of the model

In general, for Influence Ops the risk of poor V&V centers on technical risks associated with conducting rigorous V&V for complex human interaction models.

- These risks are partially characterized as R&D risks

Risk Management for V&V:

Risk management for V&V centers on:

- PLANNING, PLANNING, PLANNING
- Acknowledging the R&D risks and the constraints they imply on delivered V&V results
- V&V risks are magnified by risks in specifying the domain of application

Much risk management in V&V centers on precise answers to key questions associated with the use of benchmarks (“referents”):

- What is the technical specification of comparing model with benchmark?
- What does “good” agreement mean?
- What does “good enough” agreement mean?

Poor UQ:

UQ is a critical factor in V&V

UQ is an even more critical factor in the form of modeling results for the project as a whole

- UQ is a technical challenge
- The goal is to achieve Best Estimate Plus Uncertainty in all modeling, which is difficult even in well-understood modeling applications (like computational physics)
- UQ is dominant when modeling results are delivered for decision support

Understanding of the role of UQ in model results delivery and application is a requirement for pragmatic V&V

UQ is an essential element in specifying the domain of application

Risk Management for UQ:

This starts with specification of the domain of application

– Identifying the form of model result delivery and especially the use for decision support defines constraints and requirements for UQ

Technical UQ challenges embedded in V&V are naturally managed as part of V&V

Lack of Sufficient Data

In this domain, all model development, calibration, validation, and UQ is heavily impacted by the quality, quantity, and the scope of data

Validation/UQ can be poor because

- There is no clear understanding of the limitations and scope of available data for use by the project
- The risk associated with missing or poor data are not identified and addressed
- The use of the data (i.e., characterizing of it's uncertainty and incorporation into the process) is not adequate

Risk Management for Data:

Specify Domain of Application
- Include specification Model Capability Level targets

Identify the sources, types, and quality of data needed and available

Assess quality and sufficiency of data

Refine project efforts/deliverables based on what is reasonable and identify what is not reasonable without additional data

Characterize data uncertainty for use in Validation/UQ

Plan approach to include these types of uncertainty in the model architecture