National Aeronautics and Space Administration

How to Develop and Interpret Credibility Assessments of Numerical Models for Human Research: NASA-STD-7009 Demystified

Emily Nelson[1], Lealem Mulugeta[2], Marlei Walton[3], and Jerry Myers[1]

1. NASA Glenn Research Center
2. Universities Space Research Association, DSLS
3. Wyle Science, Technology & Engineering Group

IMM

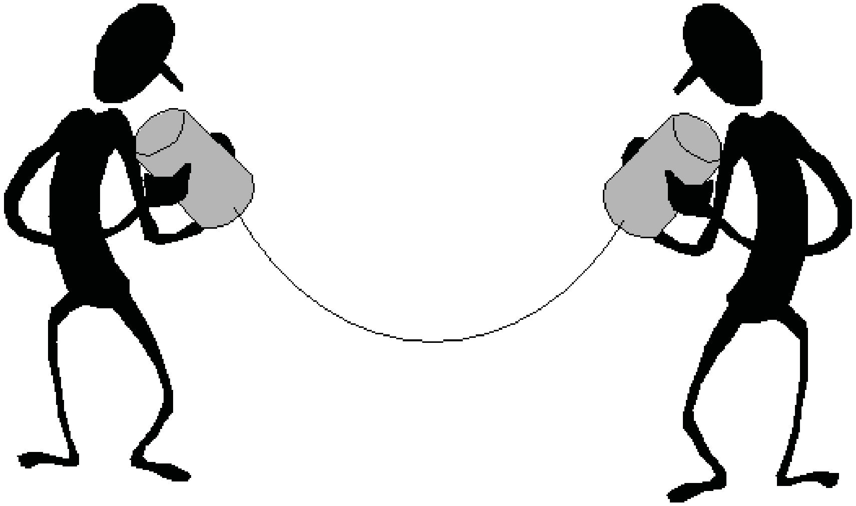
Why use NASA-STD-7009?

Numerical models continue to

- **Increase in complexity and capability**
  - Significant expertise is required for understanding them
  - Need for uncertainty quantification is recognized across all disciplines
- **Become more important in decision-making**
  - Answer questions that can't be tested except in a virtual environment
  - Code limitations and bounds of applicability must be well understood

→ **We need clear communication between modelers and customers/users**

Why use NASA-STD-7009 (cont'd)?



NASA-STD-7009 was

- **Developed after the Columbia accident to evaluate engineering systems**
  - Rich history of use
  - Abundance of documentation
- **Adopted by the Human Research Program for biomedical models**
  - Encourages the use of best practices
  - Adaptation is required to keep it relevant
  - Approach demonstrated in many biomedical applications
  - Guidance document will be available soon

Modeling and Simulation (M&S) Risk Assessment

M&S Results Influence	5: Controlling	🟩	🟨	🟥	🟥
	4: Significant	🟩	🟨	🟥	🟥
	3: Moderate	🟩	🟨	🟨	🟥
	2: Minor	🟩	🟩	🟨	🟨
	1: Negligible	🟩	🟩	🟩	🟩
M&S Risk Assessment Matrix		IV: Negligible	III: Moderate	II: Critical	I: Catastrophic
		Decision Consequence			

The M&S Risk Assessment

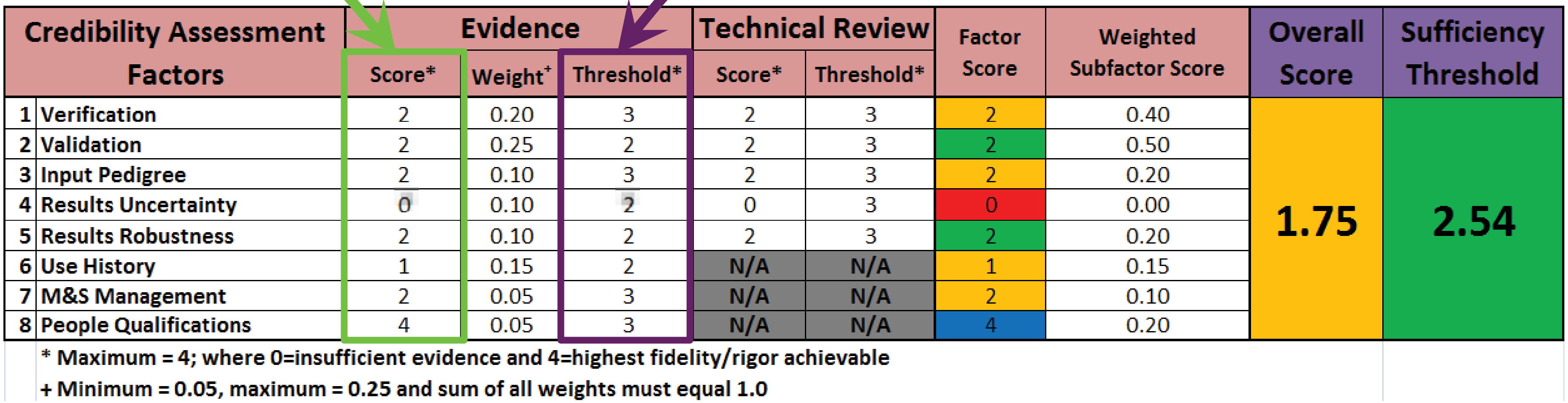
- Each new use of a numerical model should undergo risk assessment
- Zone color dictates the need for NASA-STD-7009

Zone color	NASA-STD-7009 use
Red	Required
Yellow	Typically Required
Green	Not required

Credibility Assessment Matrix: Factor Scores

NUMERICAL MODEL SCORES

TARGET SCORES



Credibility Assessment Factors		Evidence			Technical Review		Factor Score	Weighted Subfactor Score	Overall Score	Sufficiency Threshold
		Score*	Weight+	Threshold*	Score*	Threshold*				
1	Verification	2	0.20	3	2	3	2	0.40	1.75	2.54
2	Validation	2	0.25	2	2	3	2	0.50		
3	Input Pedigree	2	0.10	3	2	3	2	0.20		
4	Results Uncertainty	0	0.10	2	0	3	0	0.00		
5	Results Robustness	2	0.10	2	2	3	2	0.20		
6	Use History	1	0.15	2	N/A	N/A	1	0.15		
7	M&S Management	2	0.05	3	N/A	N/A	2	0.10		
8	People Qualifications	4	0.05	3	N/A	N/A	4	0.20		

* Maximum = 4; where 0=insufficient evidence and 4=highest fidelity/rigor achievable

+ Minimum = 0.05, maximum = 0.25 and sum of all weights must equal 1.0

Credibility assessment matrix

- Scores represent both **customer/end user** ***and*** **supplier** for each new application
- Is a living assessment that changes as the M&S evolves (for better or worse)

HRP modifications include

- High emphasis on technical reviews and
- Weighting factors appropriate to the type of M&S (deterministic, probabilistic and statistical models)

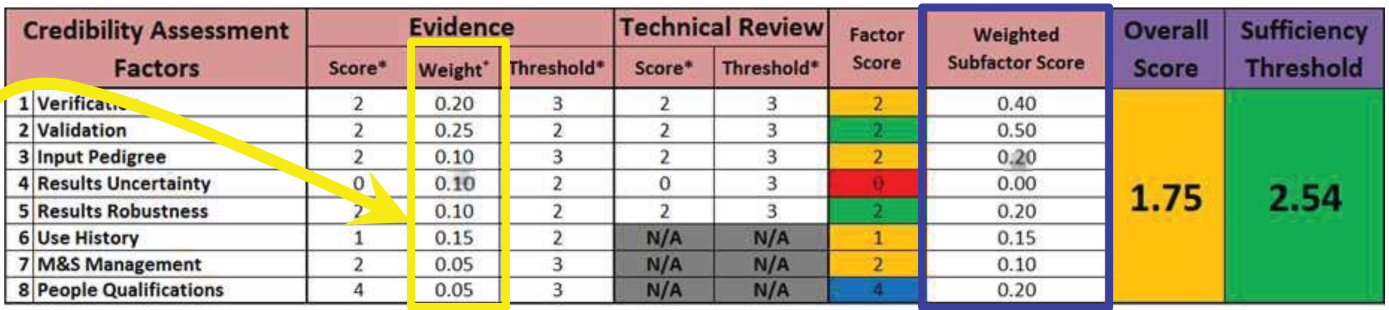
Credibility Assessment Factors

Credibility Factor		Description
1	Verification	Is the problem solved correctly? Are there bugs in the code?
2	Validation	Does the model prove itself against real-world data?
3	Input Pedigree	How much confidence is placed in the data used and the approach taken to build the model? How well does the model capture the real-world scenario?
4	Results Uncertainty	How is error assessed? Is it quantified? How much uncertainty is due to demographic/situational variation? Parameter uncertainty?
5	Results Robustness	What is the model sensitivity to key parameters? Can it be quantified over the region of M&S application?
6	Use History	Has the model been used for decision-making? Was it used in the area of application?
7	M&S Management	What are the processes/documentation developed during M&S planning, development and maintenance?
8	People Qualifications	Who is providing the guiding vision? Who is performing the implementation? What experience and background do they have?

Factor scores range from 0 (insufficient evidence) to 4 (highest fidelity/rigor)

Credibility Assessment Matrix: Proposed Weighting Strategy

WEIGHT



Credibility Assessment Factors		Evidence			Technical Review		Factor Score	Weighted Subfactor Score	Overall Score	Sufficiency Threshold
		Score*	Weight*	Threshold*	Score*	Threshold*				
1	Verification	2	0.20	3	2	3	2	0.40	1.75	2.54
2	Validation	2	0.25	2	2	3	2	0.50		
3	Input Pedigree	2	0.10	3	2	3	2	0.20		
4	Results Uncertainty	0	0.10	2	0	3	0	0.00		
5	Results Robustness	2	0.10	2	2	3	2	0.20		
6	Use History	1	0.15	2	N/A	N/A	1	0.15		
7	M&S Management	2	0.05	3	N/A	N/A	2	0.10		
8	People Qualifications	4	0.05	3	N/A	N/A	4	0.20		

WEIGHTED SUBFACTOR

Factor Weight (Proposed)		Deterministic	Probabilistic
1	Verification	0.2	0.075
2	Validation	0.25	0.15
3	Input Pedigree	0.1	0.275
4	Results Uncertainty	0.1	0.2
5	Results Robustness	0.1	0.15
6	Use History	0.15	0.15
7	M&S Management	0.05	0.05
8	People Qualifications	0.05	0.05
	TOTAL	**1.0**	**1.0**

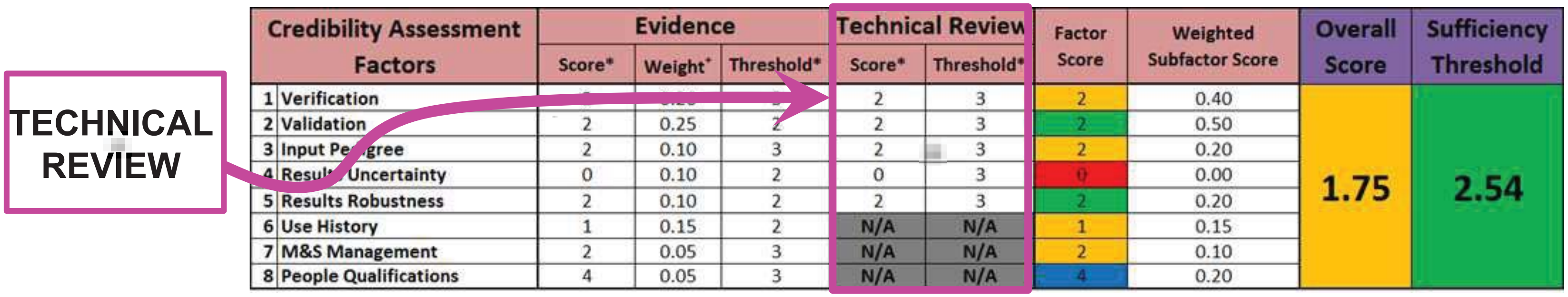
$0.05 < W_i < 0.25$

Subfactor	Weight
Evidence Weighting	0.7
Technical Review*	0.3
TOTAL	1.0

Factor and subfactor weights are assigned by the customer

*(Maximum weight is 0.3)

Credibility Assessment: Technical Review



TECHNICAL REVIEW

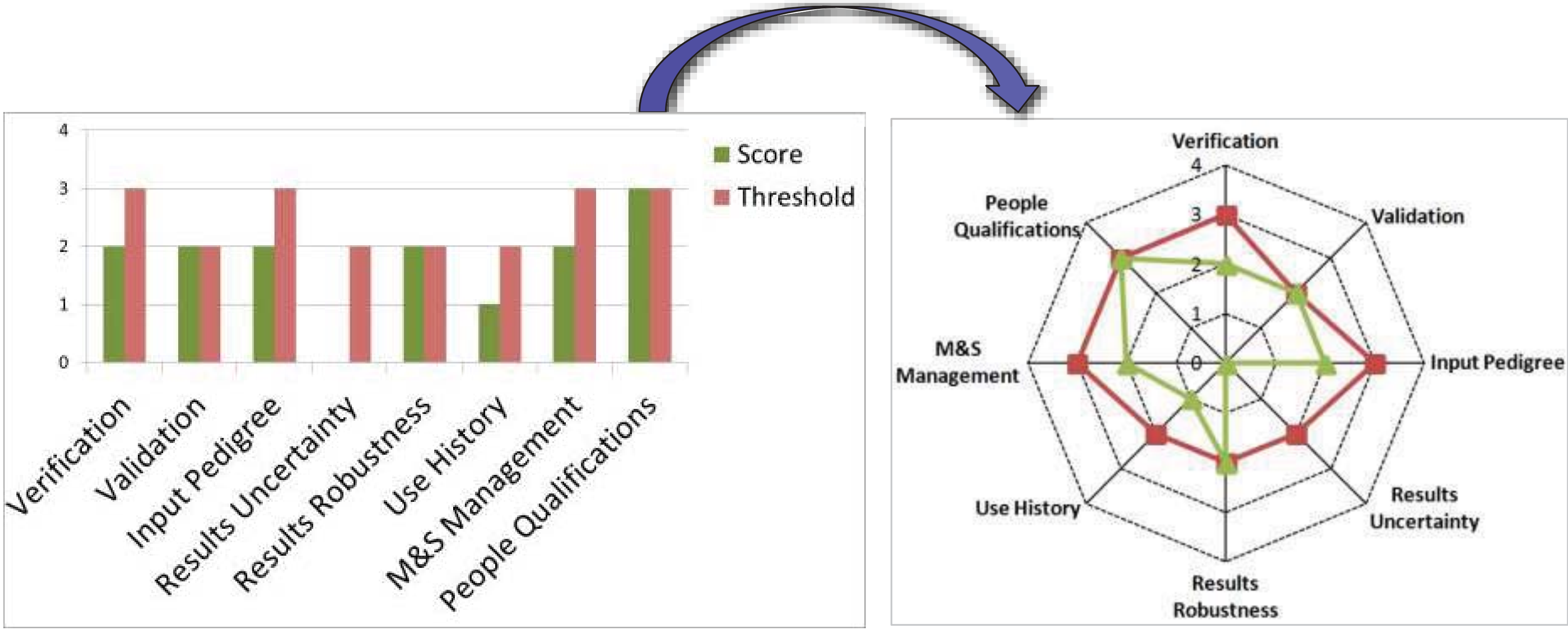
Credibility Assessment Factors		Evidence			Technical Review		Factor Score	Weighted Subfactor Score	Overall Score	Sufficiency Threshold
		Score*	Weight*	Threshold*	Score*	Threshold*				
1	Verification	2	0.20	3	2	3	2	0.40	1.75	2.54
2	Validation	2	0.25	2	2	3	2	0.50		
3	Input Pedigree	2	0.10	3	2	3	2	0.20		
4	Results Uncertainty	0	0.10	2	0	3	0	0.00		
5	Results Robustness	2	0.10	2	2	3	2	0.20		
6	Use History	1	0.15	2	N/A	N/A	1	0.15		
7	M&S Management	2	0.05	3	N/A	N/A	2	0.10		
8	People Qualifications	4	0.05	3	N/A	N/A	4	0.20		

- Technical Review also provides input on the required threshold and M&S readiness for some of the factors
- Customer specifies the level of technical review that is required for the application

Level	Technical Review
4	Favorable external peer review with independent factor evaluation
3	Favorable external peer review
2	Favorable internal peer review
1	Favorable informal internal peer review
0	Insufficient evidence

Credibility Assessment Matrix: The Spider Plot

The spider plot is essentially a multidimensional histogram to display the model and threshold scores for each factor.

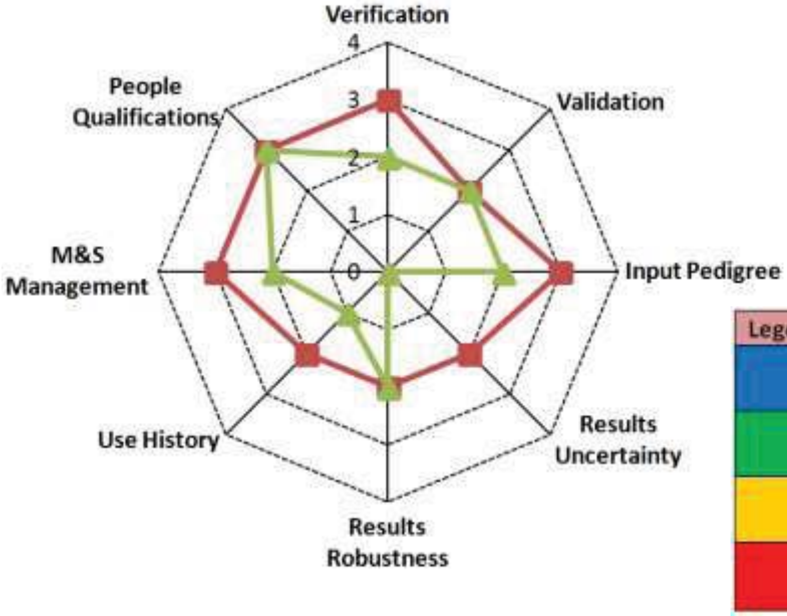


Example of Credibility Scoring – With Factor Weighting (NASA HRP Implementation)

	Credibility Assessment Factors	Evidence			Technical Review		Factor Score	Weighted Subfactor Score	Overall Score	Sufficiency Threshold
		Score*	Weight+	Threshold*	Score*	Threshold*				
1	Verification	2	0.20	3	2	3	2	0.40	1.75	2.54
2	Validation	2	0.25	2	2	3	2	0.50		
3	Input Pedigree	2	0.10	3	2	3	2	0.20		
4	Results Uncertainty	0	0.10	2	0	3	0	0.00		
5	Results Robustness	2	0.10	2	2	3	2	0.20		
6	Use History	1	0.15	2	N/A	N/A	1	0.15		
7	M&S Management	2	0.05	3	N/A	N/A	2	0.10		
8	People Qualifications	4	0.05	3	N/A	N/A	4	0.20		

* Maximum = 4; where 0=insufficient evidence and 4=highest fidelity/rigor achievable

+ Minimum = 0.05, maximum = 0.25 and sum of all weights must equal 1.0



Subfactors	Weight
Evidence	0.7
Technical Review	0.3

Legend	
🟦	CAS Score > Threshold *Exceeds credibility requirements*
🟩	Threshold ≥ CAS Score ≥ (Threshold-0.5) *Ready for use*
🟨	(Threshold-0.5) > CAS Score ≥ (Threshold-1.0) *Use with caution*
🟥	CAS Score < (Threshold-1.0) *Use not recommended or to be used with EXTREME CAUTION by subject matter experts only*

Thank you!

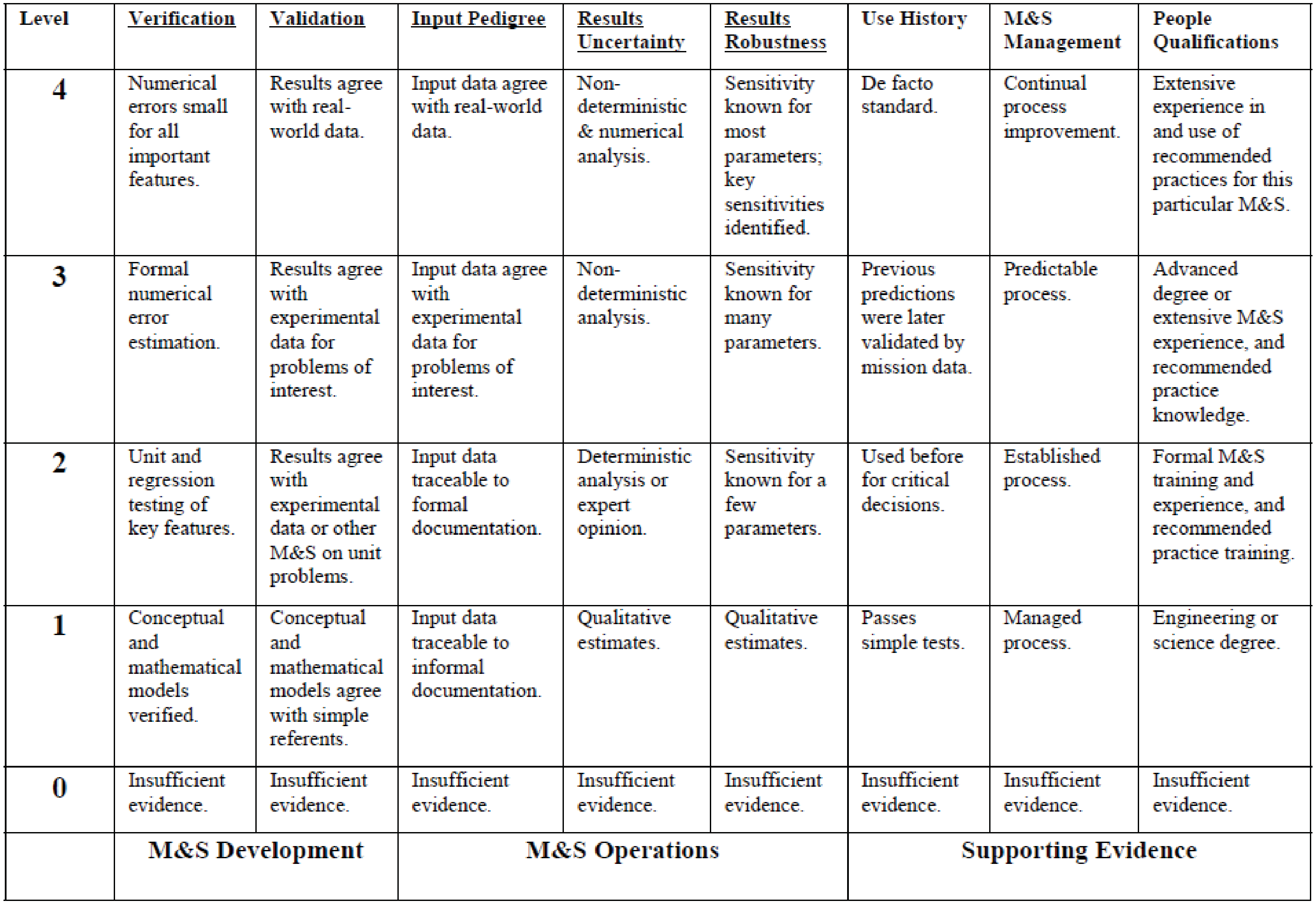
Questions?

IMM

Backups

Establish Credibility Thresholds

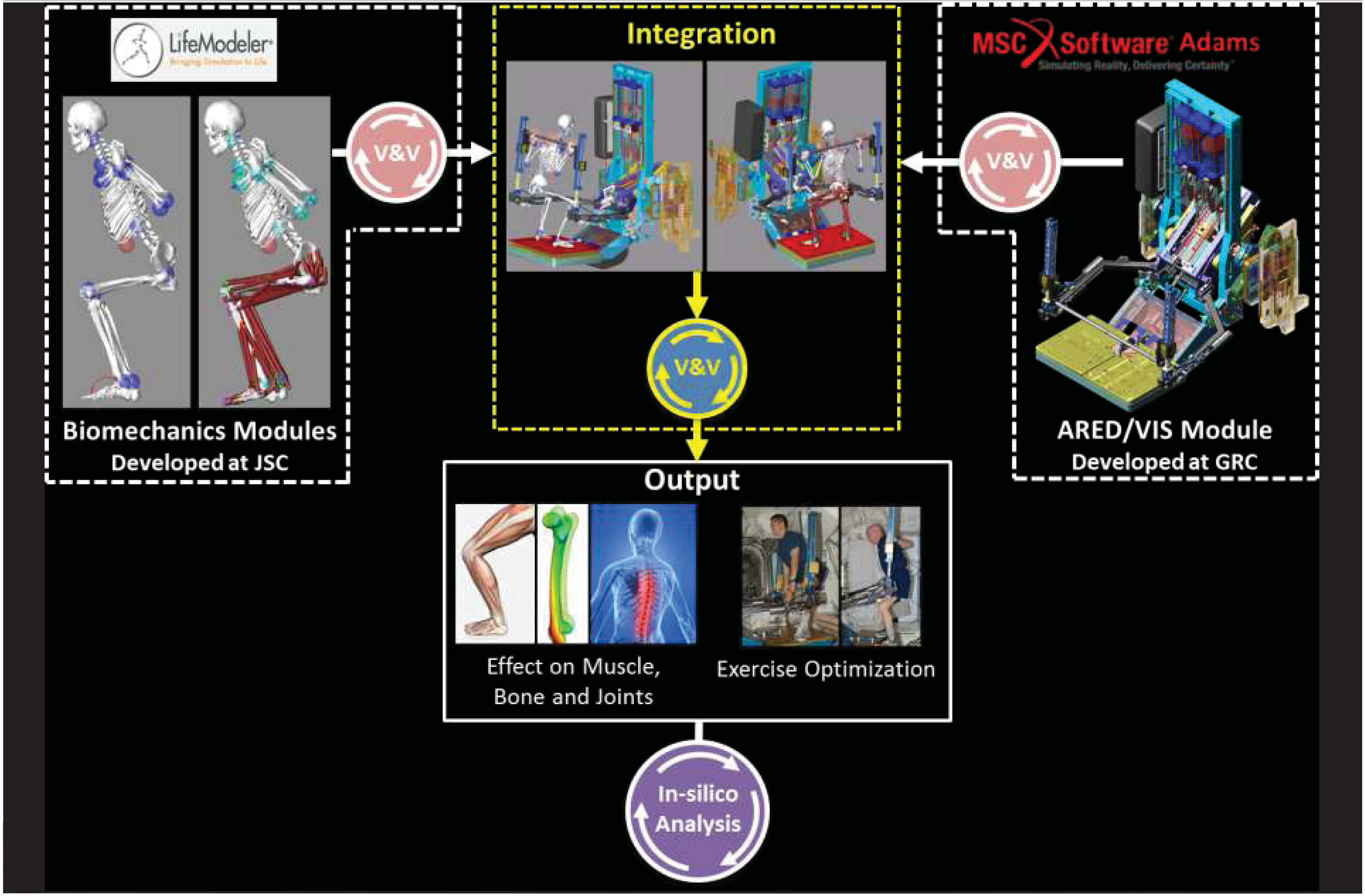
Sufficiency Thresholds



Level	Verification	Validation	Input Pedigree	Results Uncertainty	Results Robustness	Use History	M&S Management	People Qualifications
4	Numerical errors small for all important features.	Results agree with real-world data.	Input data agree with real-world data.	Non-deterministic & numerical analysis.	Sensitivity known for most parameters; key sensitivities identified.	De facto standard.	Continual process improvement.	Extensive experience in and use of recommended practices for this particular M&S.
3	Formal numerical error estimation.	Results agree with experimental data for problems of interest.	Input data agree with experimental data for problems of interest.	Non-deterministic analysis.	Sensitivity known for many parameters.	Previous predictions were later validated by mission data.	Predictable process.	Advanced degree or extensive M&S experience, and recommended practice knowledge.
2	Unit and regression testing of key features.	Results agree with experimental data or other M&S on unit problems.	Input data traceable to formal documentation.	Deterministic analysis or expert opinion.	Sensitivity known for a few parameters.	Used before for critical decisions.	Established process.	Formal M&S training and experience, and recommended practice training.
1	Conceptual and mathematical models verified.	Conceptual and mathematical models agree with simple referents.	Input data traceable to informal documentation.	Qualitative estimates.	Qualitative estimates.	Passes simple tests.	Managed process.	Engineering or science degree.
0	Insufficient evidence.	Insufficient evidence.	Insufficient evidence.	Insufficient evidence.	Insufficient evidence.	Insufficient evidence.	Insufficient evidence.	Insufficient evidence.
	M&S Development		**M&S Operations**			**Supporting Evidence**		

See NASA-STD-7009 for more info

DAP's Development and Implementation Process for Spaceflight Exercise M&S



Integration

Biomechanics Modules
Developed at JSC

ARED/VIS Module
Developed at GRC

Output

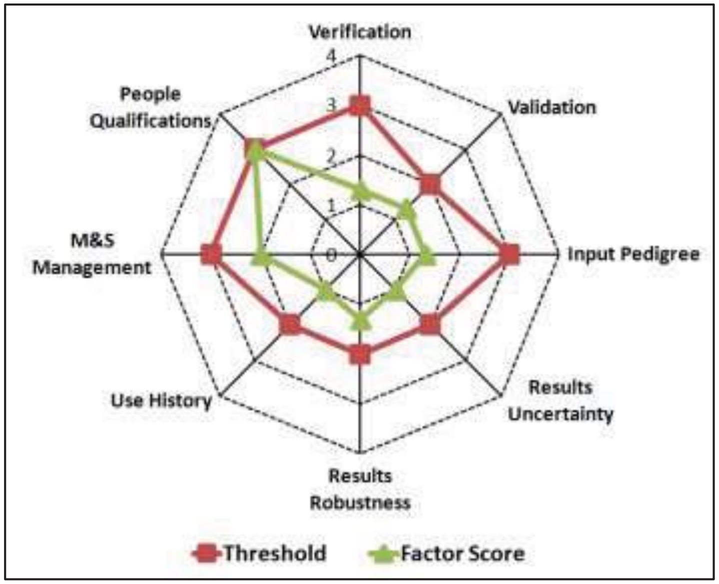
Effect on Muscle, Bone and Joints

Exercise Optimization

In-silico Analysis

Results – Credibility Summary for 0g Simulations

Credibility Assessment Factors		Evidence			Technical Review		Factor Score	Weighted Subfactor Score	Overall Score	Sufficiency Threshold
		Score*	Weight^{+}	Threshold*	Score*	Threshold*				
1	Verification	1	0.20	3	2	3	1.3	0.26	1.35	2.54
2	Validation	1	0.25	2	2	3	1.3	0.33		
3	Input Pedigree	1	0.10	3	2	3	1.3	0.13		
4	Results Uncertainty	1	0.10	2	1	3	1	0.10		
5	Results Robustness	1	0.10	2	2	3	1.3	0.13		
6	Use History	1	0.15	2	N/A	N/A	1	0.15		
7	M&S Management	2	0.05	3	N/A	N/A	2	0.10		
8	People Qualifications	3	0.05	3	N/A	N/A	3	0.15		



Credibility estimated via face value and subject matter expert inference from 1g results and knowledge of 0g exercise with ARED

Legend	
🟦	CAS Score > Threshold *Exceeds credibility requirements*
🟩	Threshold ≥ CAS Score ≥ (Threshold-0.5) *Ready for use*
🟨	(Threshold-0.5) > CAS Score ≥ (Threshold-1.0) *Use with caution*
🟥	CAS Score < (Threshold-1.0) *Use not recommended or to be used with <u>EXTREME CAUTION</u> by subject matter experts only*

Subfactors	Weight
Evidence	0.7
Technical Review	0.3

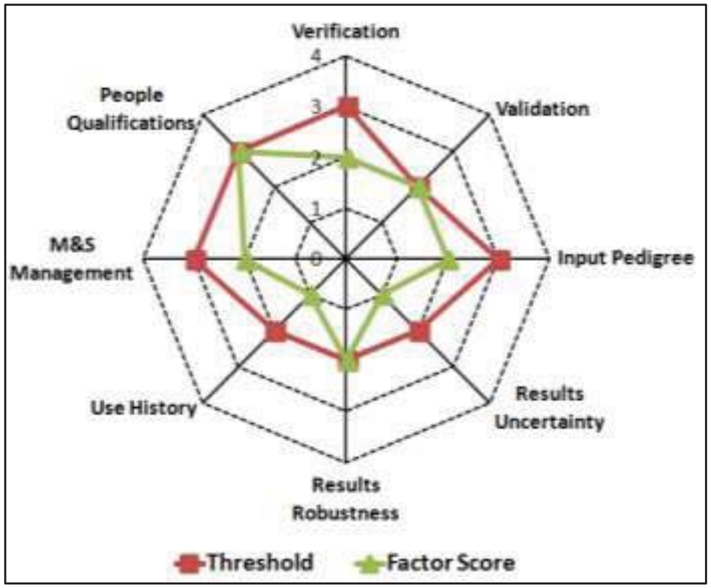
Unweighted – Model would have a CS = 1!

Results – Credibility Summary for 1g Simulations

Credibility Assessment Factors		Evidence			Technical Review		Factor Score	Weighted Subfactor Score	Overall Score	Sufficiency Threshold
		Score*	Weight+	Threshold*	Score*	Threshold*				
1	Verification	2	0.20	3	2	3	2	0.40	1.80	2.54
2	Validation	2	0.25	2	2	3	2	0.50		
3	Input Pedigree	2	0.10	3	2	3	2	0.20		
4	Results Uncertainty	1	0.10	2	1	3	1	0.10		
5	Results Robustness	2	0.10	2	2	3	2	0.20		
6	Use History	1	0.15	2	N/A	N/A	1	0.15		
7	M&S Management	2	0.05	3	N/A	N/A	2	0.10		
8	People Qualifications	3	0.05	3	N/A	N/A	3	0.15		

* Maximum = 4; where 0=insufficient evidence and 4=highest fidelity/rigor achievable

+ Minimum = 0.05, maximum = 0.25 and sum of all weights must equal 1.0



Legend	
🟦	CAS Score > Threshold *Exceeds credibility requirements*
🟩	Threshold ≥ CAS Score ≥ (Threshold-0.5) *Ready for use*
🟨	(Threshold-0.5) > CAS Score ≥ (Threshold-1.0) *Use with caution*
🟥	CAS Score < (Threshold-1.0) *Use not recommended or to be used with EXTREME CAUTION by subject matter experts only*

Subfactors	Weight
Evidence	0.7
Technical Review	0.3

Unweighted – Model would have a CS = 1!