INSTITUTE FOR DEFENSE ANALYSES

DATAWorks 2022: What Statisticians Should do to Improve M&S Validation Studies

John T. Haman, Project Leader

April 2022

Public release approved. Distribution is unlimited.

IDA Document NS D-32965

Log: H 2022-000042

INSTITUTE FOR DEFENSE ANALYSES
730 East Glebe Road
Alexandria, Virginia 22305

The Institute for Defense Analyses is a nonprofit corporation that operates three Federally Funded Research and Development Centers. Its mission is to answer the most challenging U.S. security and science policy questions with objective analysis, leveraging extraordinary scientific, technical, and analytic expertise.

About This Publication
This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-19-D-0001, Task C9082, "Statistics and Data Science Working Group," for the Office of the Director, Operational Test and Evaluation. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Acknowledgments
The IDA Technical Review Committee was chaired by Dr. V. Bram Lillard and consisted of Dr. Curtis Miller, Dr. Kelly Avery, and Dr. Thomas Johnson from the Operational Evaluation Division.

For more information:
John T. Haman, Project Leader
jhaman@ida.org • (703) 845-2132

Robert R. Soule, Director, Operational Evaluation Division
rsoule@ida.org • (703) 845-2482

Copyright Notice
© 2022 Institute for Defense Analyses
730 East Glebe Road, Alexandria, Virginia 22305 • (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 [Feb. 2014].

Rigorous Analysis | Trusted Expertise | Service to the Nation

INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS D-32965

DATAWorks 2022: What Statisticians Should do to Improve M&S Validation Studies

John T. Haman, Project Leader

Executive Summary

A. Purpose

This is a presentation on statistical best practices in modeling and simulation validation. The presentation will be given at the DATAWorks Conference 2022 at the Institute for Defense Analyses in Alexandria, Virginia.

B. Modeling and simulation validation

The Department of Defense (DOD) relies on modeling and simulation (M&S) for a variety of tasks. In particular, for operational test and evaluation, M&S is used to:

- Supplement live test data when experiments are cost- or safety-prohibitive;
- Examine threats incapable of being reproduced for testing;
- Easily plan data collection;
- Perform end-to-end mission evaluation; and
- Inform experimental design decisions.

While M&S brings several advantages compared to operational testing, testers know that M&S must be *validated*.

Model validation is the practice of measuring the degree to which the simulation disagrees with live test data. The practice of validation is *inherently statistical*, but it is up to programs and testers to determine the best design and analysis strategy for each program. Designs and analyses differ depending on intended use, operational space, operational outcomes, acceptability criteria, and many other factors.

This variation in designs and analyses is good because it gives programs and testers the freedom to choose test strategies that optimize for the specifics of their program. However, the freedom can be harmful, because suboptimal M&S validation may be applied to a program.

My opinion is that validation is a tremendously difficult undertaking, so the value of the freedom in design and analysis is a function of the statistical expertise available to a program.

C. Does faulty statistical practice affect M&S validation studies?

In 1995, Douglas Altman, a well-regarded biostatistician at Oxford University, wrote the controversial editorial *The Scandal of Poor Medical Research.*[1] It has since been called one of the most important pieces published by the British Medical Journal. In the piece, Altman claimed that many medical research papers are simply incorrect due to five fundamental statistical problems:

1. Inappropriate designs
2. Unrepresentative samples
3. Small samples
4. Incorrect analyses
5. Faulty interpretations of the analysis.

I provide some evidence that the situation has not improved in the 27 years since.

In this presentation, I suggest that several of Altman's statistical issues affect the work of data scientists and statisticians working in test and evaluation, and in M&S validation studies in particular.

D. Eight recommendations for statistical practice

My concern for statistical rigor motivated me to create a list of recommendations that I believe analysts should use to minimize the probability that an M&S validation study produces misleading results. My recommendations are:

1. Improve planning with experimental designs;
2. Improve planning by clarifying the estimand;
3. Clarify the experimental units;
4. Collaborate, do not consult;
5. Apply statistical solutions to new areas;
6. Defend against statistical inadequacies;
7. Advocate for and advance methods that are suitable for our work; and
8. Recognize that statistics is one component of model validation.

I group these recommendations into three themes, *Design*, *Process,* and *Analysis,* and discuss them in detail, and sometimes with examples.

---

[1] Altman, D.G. The Scandal of Poor Medical Research. *BMJ* 1994; 308:283 doi:10.1136/bmj.308.6924.283.

At the end of the talk, I create a map which shows how applying the recommendations can address Altman's five problems.

What Statisticians* Should do to Improve Modeling and Simulation (M&S) studies

Dr. John Haman

Presented at DATAWorks Conference
Alexandria, VA
April 27, 2022

* Statisticians, data scientists, and other data-minded individuals

Institute for Defense Analyses
730 East Glebe Road ● Alexandria, Virginia 22305

What is the point of this talk?

- Reflect on statistical practice in DOD testing
  - Focus on Modeling and Simulation
- Suggest potential issues with analysis and design
- Provide recommendations that statisticians and analysts can apply to their work

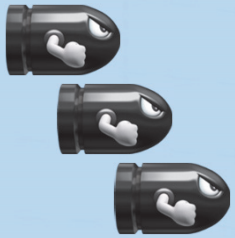
The many reasons to use models in testing and evaluation:

- Supplement live test data when experiments are cost- or safety-prohibitive
- Examine threats incapable of being reproduced for testing
- Ease of planning and logistical reasons
- Allow for end-to-end mission evaluation
- Inform experimental design decisions

Many good reasons to consider a model, but it's no substitute for a real test. And the models must be validated.

“Validation” is an assessment of the extent to which a model disagrees with live test data

Live Fire Test



Outcomes

Penetration

Modeling and Simulation

Penetration

Some perspective from another discipline…

Many papers published in medical journals are wrong (Altman, 1994). Why?

- Inappropriate designs
- Unrepresentative samples
- Small samples
- Incorrect analyses
- Faulty interpretations of the analysis

Altman, D.G., The Scandal of Poor Medical Research, *BMJ* 1994; 308 :283 doi:10.1136/bmj.308.6924.283

Sentiment has not changed much in ~25 years

"We have more data than ever, more good data than ever, a lower proportion of data that are good, a lack of strategic thinking about what data are needed to answer questions of interest, sub-optimal analysis of data, and an occasional tendency to do research that should not be done."

– Frank Harrell

Altman, D.G. The Scandal of Poor Medical Research, *BMJ* 1994; 308 :283 doi:10.1136/bmj.308.6924.283

Is faulty statistical practice a problem for M&S studies?

Many papers published in medical journals are wrong (Altman, 1994). Why?

- Inappropriate designs (unavoidable problem)
  - *Cannot test to measure improvement*
- Unrepresentative samples (big problem)
  - Ranges are old, simulation used for new threats, etc.
- Small samples (big problem)
  - *Obviously…*
- Incorrect analyses (unknown problem)
  - *Analyses classified, hidden*
- Faulty interpretations of the analysis (small problem)

My view of the extent of Altman's problems to defense testing and model validation.

Statisticians can improve the quality of M&S studies considerably

In this talk, I present 8 recommendations that statisticians can use to improve validation studies

Design

1. Experimental design
2. Clarify estimands
3. Experimental units

Process

4. Collaborate
5. Apply statistical solutions
6. Proactivity

Analysis

7. Advocate for methods
8. Multiple components

1. Improve planning with experimental designs

DOE = language to discuss trade-offs associated with planning

Good DOE =

- Optimizes the utility of the test
- Clarifies how we analyze the data
- Maximizes Pr(analysts agree on what to do with data)
- Minimizes Pr(analysis problems)

- **Specific recommendations for M&S:** Use space-filling designs for computer experiments
  - Underused in defense community

Planning by alison from NounProject.com

2. Improve planning by clarifying the estimand

- How do we link requirements with data?
- The primary objective of the validation study is to demonstrate that the simulation data is congruent with the live data
  - What that means *is the estimand*
  - How we do it *is the estimator*
- Can be very tricky in validation studies. Suppose we want to compare P_{hit} between sim and live data
  - Can use $P(hit,live)-P(hit,sim)$ or $P(hit,live)/P(hit,sim)$ or something else
  - Which factors should we control to make comparisons
  - We don't want to be in a situation where **"I'll know what's best when I get the data"**

Estimate Time by Eucalyp from NounProject.com

Example: JAGM

Two goals in validation concept:

1. Ensure that the aggregate miss distance distribution of sim shots agrees with live shots
   - ***Estimand:*** Maximum vertical distance between CDFs
2. Determine whether the M&S accounts for factors that affected miss distance of the live shots
   - ***Estimand:*** Change in mean miss distance between test factors

3. Improve analyses by clarifying the *experimental units*

- Experimental unit = smallest independent unit of data
  - In classroom: A coin flip or dice roll
  - In real life: everything correlated (Hard!!)

- When we know the units, we are on the same page about how much data is enough
- For many situations, experimental unit is ***The Mission***
- When units are not independent, good analysis considers correlation

Analysis by Orin zuu from NounProject.com

4. Collaborate, do not consult

- In collaboration, focus on minimizing Pr(Misleading Results).
  - Because analyses can be wrong!
- Statisticians should focus on quality control aspects of the acquisition pipeline
  - Consultants cannot work on the process, they only see a piece.
- Guard against the *"belief confirmation machine"*
- Collaboration gives statisticians a basis to suggest improvements and educate

Collaboration by Soremba from NounProject.com

5. Apply statistical solutions to new areas

- Do not wait for statistical problems to come along
- It is better to apply data-centric methods to all problems
- Statisticians have a background that colors their perspective in a unique way

Bar Graph by b farias from NounProject.com

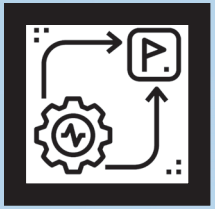
6. Proactively address statistical inadequacies

- Not speaking up when something is wrong = **You Fail**
- Explain assumptions
- Explain the concepts of statistical validation, create mock analyses
- Explain that risks influence how good is good enough, and that quantification is required

---

- **Specific Recommendation for M&S:** Keep track of which data are used for model development, and which are used for evaluation. Do not “double dip.”

Proactive by LAFS from NounProject.com

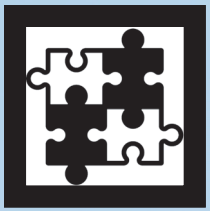


7. Advocate for and advance methods that are suitable for the test community

- Community does not like to make assumptions about data
- When methods are proposed, deduce the assumptions of those methods to determine what is palatable in your org.

- **Specific Recommendation for M&S:** Get around the "what if my model is mis-specified" problem in T&E.
  - Splines for efficient modeling factor effects
  - Ordinal data methods for non-parametric modeling
  - Markov or mixed models for correlated data

Method by Nithinan Tatah from NounProject.com



8. Recognize that statistics is one component of model validation

- Work to understand the perspectives of non-statisticians
- Many things bear on validation, such as intended use, but there is no way to incorporate these into a statistical model. Validation = thoughtful statistics + consideration of risks.
- Validation is a team-based exercise

Component by WEBTECHOPS LLP from NounProject.com

My eight recommendations for statisticians working on M&S

1. Improve planning with experimental designs
2. Improve planning by clarifying the estimand
3. Clarify the experimental units
4. Collaborate, do not consult
5. Apply statistical solutions to new areas
6. Defend against statistical inadequacies
7. Advocate for and advance methods that are suitable for our work
8. Recognize that statistics is one component of model validation

I propose recommendations that address these issues in our community

Altman's Statistical Problems

- Inappropriate designs
- Unrepresentative samples
- Small samples
- Incorrect analyses
- Faulty interpretations of the analysis

Statistical Recommendations

1. DOE
2. Clarify the estimand
3. Clarify the experimental units
4. Collaborate
5. Apply statistical solutions to new areas
6. Proactively address statistical inadequacies
7. Advocate for methods
8. Recognize that statistics is one component

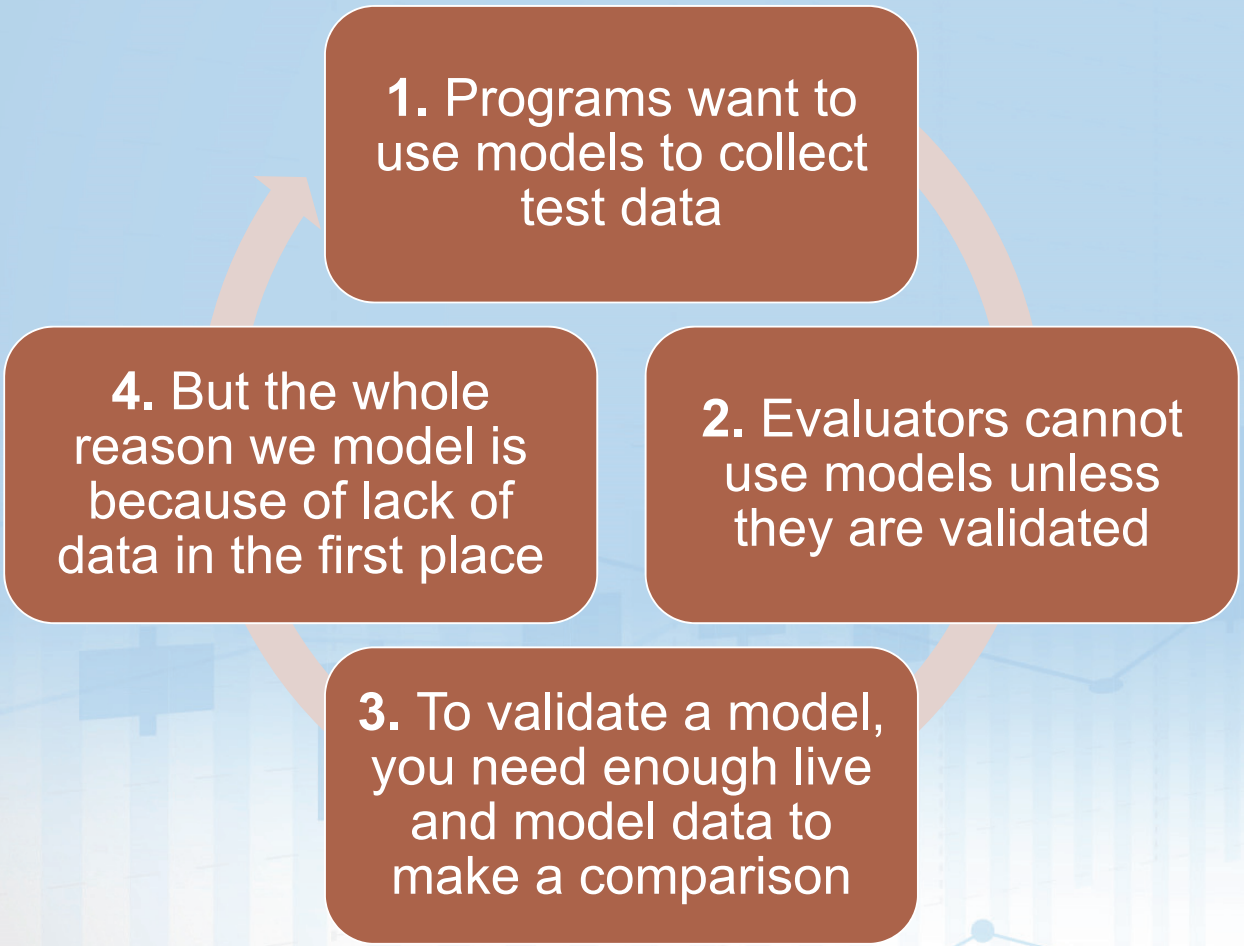
Design

Process

Analysis

Backups

Current state and the paradox of model validation



Currently, there is no policy on how to overcome this problem, so each project has a custom validation strategy

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
04-2022	Draft Final	START DATE	END DATE Apr 2022

4. TITLE AND SUBTITLE
DATAWorks 2022: What statisticians Should do to Improve M&S Validation Studies

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
Separate Contract		
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**
C9082	C9082	

6. AUTHOR(S)
Haman, John, T.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Institute for Defense Analyses
730 East Glebe Road
Alexandria, Virginia 22305

8. PERFORMING ORGANIZATION REPORT NUMBER
D-32965-NS
H 2022-000042

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER

12. DISTRIBUTION/AVAILABILITY STATEMENT
Public release approved. Distribution is unlimited.

13. SUPPLEMENTARY NOTES
Project Leader: Haman, John, T.

14. ABSTRACT
It is often said that many research findings -- from social sciences, medicine, economics, and other disciplines -- are false. This fact is trumpeted in the media and by many statisticians. There are several reasons that false research is published, but to what extent should we be worried about them in defense testing and modeling and simulation?

In this talk I will present several recommendations for actions that statisticians and data scientists can take to improve the quality of our validations and evaluations.

15. SUBJECT TERMS
Simulation; Statistics; Data Science; Modeling; Best Practices

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	c. THIS PAGE		
Unclassified	Unclassified	Unclassified	SAR	35

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER
John T. Haman	703-845-2132