Army Test and Evaluation Command

Incorporating DoE Analytic Techniques and Test-execution Lessons-Learned to increase Credibility of T&E

NDIA Presentation on 2 March 2010

Nancy Dunn, DA Civilian
Analysis Division, US Army Evaluation Command
nancy.dunn@us.army.mil
(410) 306-0454

Rick Kass, GaN Corporation
Contract Technical Support to –
US Army Operational Test Command (USAOTC)
rick.kass@us.army.mil
(254) 286-5572

Army Proven
Battle Ready

What is a Credible T&E?

To justify recommendations ….
…need "credible T&E"

Propose Two General Characteristics for "Credible T&E"

Robustness: -- "breadth"

Robust T&E Strategy/Design-- ***systematically assesses all important factors and conditions that could impact system performance*** across the full expected operational environment.

Rigor: -- "depth"

Rigorous Test Event – ***provides convincing evidence to support system-performance conclusion*** by eliminating threats to test validity.

…. some overlap between techniques…

T&E Robustness -- Central Challenge

Factors	Conditions
Time of Day	day, night
Type of C2	voice, digital
SUT Activity	stationary, move
Threat Intensity	Hi, low
Operational Environment	urban, rural
Threat ECM	benign, ECM
…..	…..

System-Under-Test (SUT)

System Performance

(MOE/MOP

- Percent of detections
- Probability of kill
- Message completion rate
- ……

T&E Primary Issue – What impact do operational factors and conditions have on system performance?

(Under what conditions does the SUT meet requirements?)

1….given a SUT and outcome measure (MOE or MOP) of interest …

2….and a large potential number of factors and conditions that could impact SUT performance…

3….what is the most <u>scientifically defensible</u> and <u>efficient</u> way to examine the largest number of factors and conditions with the fewest number of test trials.

Robustness – Shaping the T&E (1 of 5)

1. **Define critical response variables** (MOE/MOP)
   – missed distance & time-to-way-point
2. **Determine all factors** that could affect response variables
3. **Determine levels of factors** that can be implemented

…determine what conditions to test in which event….

4. **Determine availability of assessment Events** (Tests and M&S)
5. **Determine Factors and Levels** to be evaluated **in each event**

Factors	# of Levels	Conditions	DT-1	DT-2	LUT	EW	IOT Ph 1	IOT Ph 2	IOT Ph 3	M&S
System Under Test	2	2 Venders (A, B)	2	2	2	2	2	2	2	2
Mission Type	2	Attack, Defense,	None	None	2	2	2	Attack	2	2
Terrain Type	2	Flat, Hill	2	2	2	2	2	2	2	2
Light Condition	2	Day, Night	2	2	2	2	2	2	2	2
Blue Echelon	4	BN, CO, PLT, SQD	individual	individuals	CO	CO	PLT	SQD	BN	4
Network Load	2	High, Low	Low	2	2	Medium	2	Low	2	2
EW Environment	2	Benign, Jammed	Benign	Benign	Benign	2	Benign	Benign	Benign	Benign
IW Environment	2	Benign, Threat-CNO	Benign	Benign	Benign	Benign	Benign	Benign	2	Benign

6. **Determine most efficient Design for each event**

Robustness -- Shaping the T&E (2 of 5)

6. Determine most efficient Test Design for a particular event (1 of 3)

Limited User Test (LUT) Test Design Matrix

Vendor	Mission	Net Load	Flat		Hilly	
			Day	Night	Day	Night
A	Attack	Lo				
		Hi				
	Defend	Lo				
		Hi				
B	Attack	Lo				
		Hi				
	Defend	Lo				
		Hi				

…how much testing is enough?

…to examine each combination only once would take 32 test trials….

….too much or too little?

Robustness -- Shaping the T&E (3 of 5)

6. Determine most efficient Test Design for a particular event (2 of 3)

If all combinations important,
but can't do 32 trials (16 trials per Vendor)…

- DWWDLT – "Do what we did last time."
- OFAT -- Examine "one factor at a time"
- Select worst-case combinations
- Select most-likely combinations
- Ask someone – ask the "oldest evaluator/tester"

- **Use DoE Factorial techniques …….**

Robustness -- and Traditional DOE

Robust Test -- ***systematically assesses all important factors and conditions that could impact system performance***

Design of Experiments (DoE) provides ....

....scientific credibility/justification test design

....explicit way to determine test sample size – how much testing is enough

....most efficient method to examine large number of conditions with fewest test trials

…test design now becomes a science…

…base on 100+ years of methodological development

…new computer DoE software allows Statistician to fit design to the experiment

Factorial Designs and ANOVA are DOE. DOE was first developed and used in farm trials by Sir R. A. Fisher (1925), a mathematician and geneticist

From Greg Hutto's presentation to OTA Conference, Oct08

Robustness -- Shaping the T&E (4 of 5)

6. Determine most efficient Test Design for a particular event (3 of 3); based on ……

…desired resolution of factors (alias structure)

…power analysis requirements (sample size -- # of test trials)

…available time/resources to execute # of trials

Vendor	Mission	Net Load	Flat Day	Flat Night	Hilly Day	Hilly Night
A	Attack	Lo	1	1	1	1
		Hi	1	1	1	1
	Defend	Lo	1	1	1	1
		Hi	1	1	1	1
B	Attack	Lo	1	1	1	1
		Hi	1	1	1	1
	Defend	Lo	1	1	1	1
		Hi	1	1	1	1

Full-factorial

32 test trials; Res-VI

99% power (1-β)

Vendor	Mission	Net Load	Flat Day	Flat Night	Hilly Day	Hilly Night
A	Attack	Lo		1	1	
		Hi	1			1
	Defend	Lo	1			1
		Hi		1	1	
B	Attack	Lo	1			1
		Hi		1	1	
	Defend	Lo		1	1	
		Hi	1			1

Half-factorial

16 test trials; Res-IV

93% power (1-β)

Quarter-factorial

8 test trials; Res-III

36% power (1-β)

Vendor	Mission	Net Load	Flat Day	Flat Night	Hilly Day	Hilly Night
A	Attack	Lo		1		
		Hi			1	
	Defend	Lo			1	
		Hi		1		
B	Attack	Lo				1
		Hi	1			
	Defend	Lo	1			
		Hi				1

Robustness -- Lessons Learned thus far for DOE implementation in T&E Planning

T&E Strategy and Design

- Requires good understanding of DoE to examine alternative designs
- Are all critical factors considered?
- Balancing act between resources and sufficient sample size

Post-test Data Production

- Need quick-look results capability on test site
  - Too late to understand why anomalies/trends occurred after everyone goes home
- Need to associate trial conditions (factors/levels) with response variables

Test Planning & Execution ….

…now that we have a Robust T&E Strategy and Design…

…how do we ensure we will have **a valid test execution and valid data** to analyze**?**

Rigor: -- depth

Rigorous **Test Planning & Execution –** ***provides convincing evidence to support system-performance conclusion*** by eliminating or reducing threats to test validity.

Army Proven Battle Ready

Test Rigor -- 4 General Requirements

	Requirement	Evidence for Validity	Threat to Validity
1	ability to **employ treatment** (test system and planned factors)	Treatment successfully implemented	System and test architecture did not work
2	ability to **detect change** in response (MOE/MOP)	Response changed as Treatment changed	Too much noise, can not detect any change
3	ability to **isolate reason** for change	Treatment alone caused Response	Alternate explanations of change available
4	ability to **relate results** to actual operations	Response magnitude is expected in actual operations	Observed change may not be applicable

ATEC

Test Rigor

-- 5 Test Components to Consider

Treatment

Possible Cause

Independent Variable

Examples

- new sensor
- new C2 process
- day/night

System and Test Factors

Player Operators/Unit

Smallest Unit Assigned to Treatment

Examples

- sensor operator
- sensor management cell
- Company A

Effect

Response Variable MOE/MOP

Dependent Variable

Examples

- targets detected
- time from sensor to shooter
- percent objectives met

Trial

- **Execute Treatment to observer Response**
- **Includes constant and random conditions**
  - **Weather, free play, etc**

Analysis

Document CHANGE in Response

Examples

-Response Variable compared to:

- -different factors/conditions
- -System requirements

Test Rigor -- 21 Threats to Test Validity

Five Test Components	Four Test Validity Requirements				
	1. Ability to Employ System and Test Factors	2. Ability to Detect Change	3. Ability to Isolate Reason for Change: Single Group	3. Ability to Isolate Reason for Change: Multiple Groups	4. Ability to Relate Test Results to Operations
1. Treatment	(1) System functionality does not work.	(5) System functionality varies in performance.	(11) System functionality changes across …	NA	(18) System functionality does not represent future capability.
2. Players	(2…				…9) Players do not …ent …nal unit.
3. Effect					…es do … …t effects.
4. Trial	(4) Factor… conditions not adequately implemented.		…	…under different trial conditions.	(21) Scenario is not realistic.
5. Analysis		(9) Low statistical power (10) Statistical assumptions violated.			

These 21 Threats

need to be considered during test planning …

… so that they are <u>controlled, reduced, or eliminated</u>

<u>during test execution</u>

Test Rigor –

Guidelines for Designing Test Execution

… by eliminating threats to meet 4 Validity Requirements

Test Rigor -- 21 Threats to Test Validity

Five Test Components	Four Test Validity Requirements				
	1. Ability to Employ System and Test Factors	2. Ability to Detect Change	3. Ability to Isolate Reason for Change: Single Group	3. Ability to Isolate Reason for Change: Multiple Groups	4. Ability to Relate Test Results to Operations
1. Treatment	(1) System functionality does not work.	(5) System functionality varies in performance.	(11) System functionality changes across trials.	NA	(18) System functionality does not represent future capability.
2. Players	(2) Players are not adequately prepared.	(6) Test players vary in proficiency.	(12) Player proficiency changes across trials.	(15) Groups differ in player proficiency.	(19) Players do not represent operational unit.
3. Effect	(3) Measures are insensitive to capability impact	(7) Data collection accuracy is inconsistent.	(13) Data collection accuracy changes across trials.	(16) Data collection accuracy differs for each group.	(20) Measures do not reflect important effects.
4. Trial	(4) Factors & conditions not adequately implemented.	(8) Trial conditions fluctuate.	(14) Trial conditions change across trials.	(17) Groups operate under different trial conditions.	(21) Scenario is not realistic.
5. Analysis		(9) Low statistical power (10) Statistical assumptions violated.			

Internal Validity -- "Ability to…

1. **…Employ Test System in Planned Conditions**
2. **…Detect Change in Response** MOE/MOP
3. **…Isolate Reason for Change** in Response

External Validity -- "Ability to…

4. **…Relate Test Results** to Military Operations

Rigorous test – *provides evidence to support system-performance conclusion* by eliminating or reducing threats to test validity

Army Proven Battle Ready

1. Ability to Employ Test System

in Planned Conditions

Most consistent "lessons learned" reported after test completed:

- ***New System did not function as designed.***
- ***Players did not know how to employ it properly.***
- ***Response Measures (instrumentation) not sensitive to its use.***
- ***Trial Conditions not adequately implemented to impact system employment***

Threats

Treatment

1. System functionality does not wo
   Does the HW/SW work?

Unit

2. Players not adequately prepare
   Do the players have the training and

Effect

3. Measures insensitive to system impact
   Is the response varia

Trial Conditions

4. Factors and Conditi
   test conditions sufficient to i

PREVENTION examples

of capability Materiel Readiness
tement.

training, TTP, and sufficient
aining Readiness Statement

- SMEs and data collectors ability to "see" differences
  rtification

and monitor

Requires **full-up Pilot-Test** with **adequate time** prior to Record Trials…

….. to examine results and implement fixes

Test Rigor –

Ensuring that the system-under-test is used and can make a difference ….

…..is the first logical step in designing a valid test.

Army Proven Battle Ready

Test Rigor –

Guidelines for Designing Test Execution

… by eliminating threats to meet 4 Validity Requirements

Test Rigor -- 21 Threats to Test Validity

Five Test Components	Four Test Validity Requirements				
	1. Ability to Employ System and Test Factors	2. Ability to Detect Change	3. Ability to Isolate Reason for Change: Single Group	3. Ability to Isolate Reason for Change: Multiple Groups	4. Ability to Relate Test Results to Operations
1. Treatment	(1) System functionality does not work.	(5) System functionality varies in performance.	(11) System functionality changes across trials.	NA	(18) System functionality does not represent future capability.
2. Players	(2) Players are not adequately prepared.	(6) Test players vary in proficiency.	(12) Player proficiency changes across trials.	(15) Groups differ in player proficiency.	(19) Players do not represent operational unit.
3. Effect	(3) Measures are insensitive to capability impact	(7) Data collection accuracy is inconsistent.	(13) Data collection accuracy changes across trials.	(16) Data collection accuracy differs for each group.	(20) Measures do not reflect important effects.
4. Trial	(4) Factors & conditions not adequately implemented.	(8) Trial conditions fluctuate.	(14) Trial conditions change across trials.	(17) Groups operate under different trial conditions.	(21) Scenario is not realistic.
5. Analysis		(9) Low statistical power (10) Statistical assumptions violated.			

Internal Validity -- "Ability to…

1. **…Employ Test System in Planned Conditions**
2. **…Detect Change in Response** MOE/MOP
3. **…Isolate Reason for Change** in Response

External Validity -- "Ability to…

4. **…Relate Test Results** to Military Operations

Rigorous test – *provides evidence to support system-performance conclusion* by eliminating or reducing threats to test validity

Army Proven Battle Ready

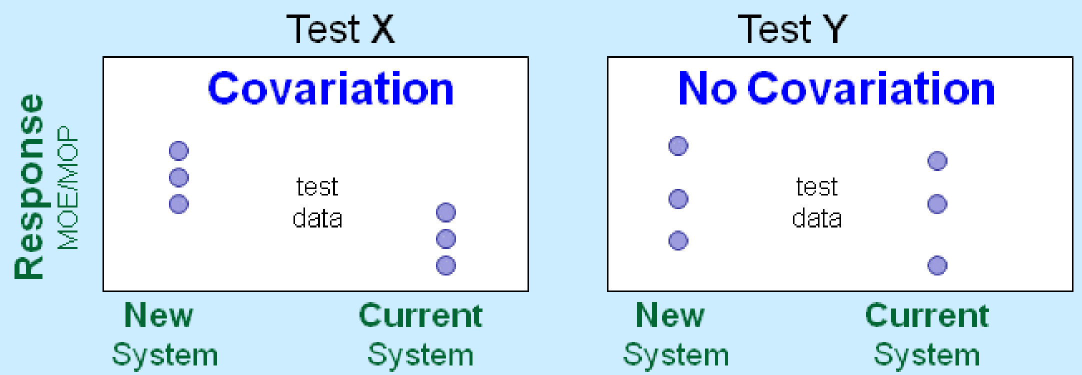
2. Ability to Detect Change in Response

•Given that System and Test Factors are adequately employed

•Next Question: Did Response change when Test Factors were changed?

Ability to detect change in Response

"Detect Change" = Detect COVARIATION --- Response changes as Treatment changes



Two Groups of Threats to Detecting Change

- **Fail to Detect Real Change**
  - Incorrectly see no covariation (**Type II Error, Producer Risk, Beta Error**)
- **Incorrectly Detect Change--**
  - Incorrectly see covariation (**Type I Error, Consumer Risk, Alpha Error**)

Test Rigor – 2. Ability to Detect Change-- statistical validity

Threats

PREVENTION examples

Treatment — **5. Test Systems vary in performance** Continual fluctuation in functionality

- Continually monitor
- Use mature system

Unit — **6. Players vary i**
- Differe
- Dif

…level performance

Fail to Detect Change

Type II Error

Effect — **7. De**

…data collectors

Trial — **8. Trial co**
- Inadvertent

…conditions

…to see "effect" …

•reduce "noise" in test architecture

•run sufficient Sample Size

-- less variation in test architecture reduces sample-size requirement

Analysis — **9. Low Statistical Power**
- **Small sample**
- Too stringent alpha risk (1%, 5%, 10%)
- **Inefficient statistical test**

- Increase number of replications
- Increase alpha risk
- Use paired comparisons
- Use appropriate statistical test for data assumptions

Incorrectly Detect Change

Type I Error

10. High Consumer Risk
- **High alpha risk**
- **Error rate problem (fishing)**
  - Large number of statistical tests
- Violating statistical technique assumptions

- Evaluate impact/tradeoffs of alpha-beta levels
- Select fewer, more meaningful MOPs

Test Rigor –

Guidelines for Designing Test Execution

… by eliminating threats to meet 4 Validity Requirements

Test Rigor -- 21 Threats to Test Validity

Five Test Components	Four Test Validity Requirements				
	1. Ability to Employ System and Test Factors	2. Ability to Detect Change	3. Ability to Isolate Reason for Change: Single Group	3. Ability to Isolate Reason for Change: Multiple Groups	4. Ability to Relate Test Results to Operations
1. Treatment	(1) System functionality does not work.	(5) System functionality varies in performance.	(11) System functionality changes across trials.	NA	(18) System functionality does not represent future capability.
2. Players	(2) Players are not adequately prepared.	(6) Test players vary in proficiency.	(12) Player proficiency changes across trials.	(15) Groups differ in player proficiency.	(19) Players do not represent operational unit.
3. Effect	(3) Measures are insensitive to capability impact	(7) Data collection accuracy is inconsistent.	(13) Data collection accuracy changes across trials.	(16) Data collection accuracy differs for each group.	(20) Measures do not reflect important effects.
4. Trial	(4) Factors & conditions not adequately implemented.	(8) Trial conditions fluctuate.	(1[illegible]) Trial conditions change across trials.	(17) Groups operate under different trial conditions.	(21) Scenario is not realistic.
5. Analysis		(9) Low statistical power (10) Statistical assumptions violated.			

Internal Validity -- "Ability to…

1. **…Employ Test System in Planned Conditions**

2. **…Detect Change in Response** MOE/MOP

3. **…Isolate Reason for Change** in Response

External Validity -- "Ability to…

4. **…Relate Test Results** to Military Operations

Rigorous test – *provides evidence to support system-performance conclusion* by eliminating or reducing threats to test validity

Army Proven
Battle Ready

3. Isolating the Reason for Change

- Given that **System and Test Factors are adequately employed**
- Given that **Response change when Test Factors were changed?**
- **Next Question: What really produced change in Response** MOE/MOP**?**

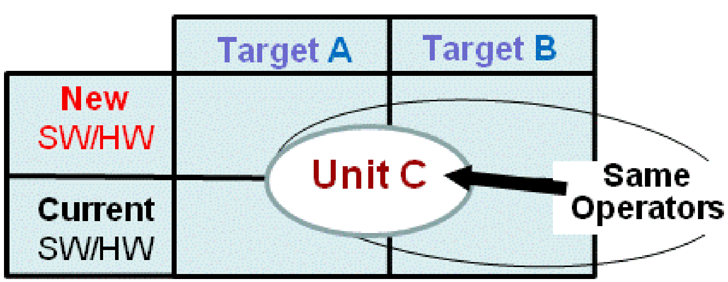
Validity -- <u>Treatment alone</u> caused change in Response

Threat -- Something else caused change in Response -- ***confounded results***

-- Threat depends on type of experimental design

Single Group Design

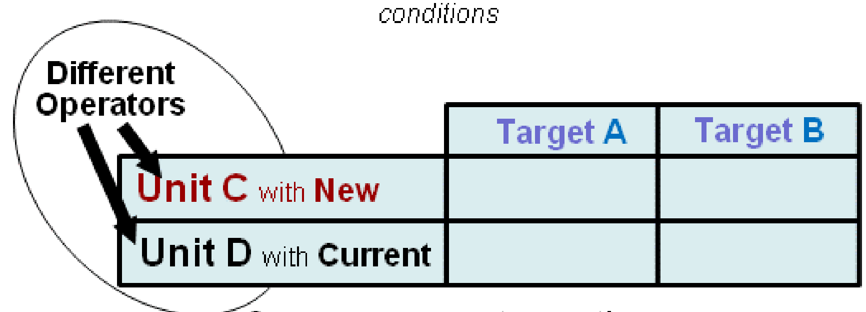
One unit receives all treatment conditions



Compare group under different conditions

Multiple Group Design

Different units receive different treatment conditions



Compare group to another group
- Side-by-side baseline
- Side-by-side "shoot off"

3. Isolating the Reason for Change

In SINGLE-GROUP DESIGNS

Sequence of trial presentation is critical consideration

Sequence 1: **Sequenced**

Mon	Tue	Wed	Thu
Current	Current	New	New

(1+0=1) (1+1=2) (1+2=3) (1+3=4)

Current =3 New =7

Sequence 2: **Mixed**

Mon	Tue	Wed	Thu
Current	New	Current	New

(1+0=1) (1+1=2) (1+2=3) (1+3=4)

Current =4 New =6

Sequence 3: **Counterbalanced**

Mon	Tue	Wed	Thu
Current	New	New	Current

(1+0=1) (1+1=2) (1+2=3) (1+3=4)

Current =5 New =5

(1 + 0 = 1)

Treatment Effect — **Learning Effect** — **Observed Effect**

In single-group design, order effect generates greatest threat to Isolating Reason for Change

3. Isolating the Reason for Change

SINGLE-GROUP DESIGN ORDER EFFECTS

Order effect impacts all 4 components of test execution

Treatment

Threats

11. System Functionality changes across trials
System functionality improves or degrades over time

Unit

12. Player Profici
Performance improves
than treatment prese

Effect

13. Data Collect
trials Data collectors or ins
over time ---artificially changing resu

Trial

14. Factors & Conditions change across trials
Implementation of factors levels or controlled and uncontrolled trial conditions (weather, OPFOR) improve or degrade over time

PREVENTIONS examples

- Use fixed configuration
- No fix-test-fix
- ...ze or counterbalance
- ...rmance prior to start
- ...aximum performance prior to start
- ...heck and recalibrate instrumentation after each trial
- Train OPFOR to maximum performance prior to start
- Randomize or counterbalance trials

Continually monitor

for increases or decreases in all 4 test components…

…to prevent/control **unintended changes** across test trials

3. Isolating the Reason for Change

Multiple-Group Designs – "unintended difference"

Previous Order-Effect threats are neutralized

- if same sequence given to both groups, and
- all comparisons are between groups

(Compare Unit C with current systems to Unit D with future systems)

	Target A	Target B
Unit C with Future		B_1
Unit D with Current		B_2

While Multiple-Group designs alleviate Order-Effect threats *…for between-group comparisons…*

A new set of threats arises…

- **…because different treatments are intertwined with different groups**
- **…difficult to separate treatment effects from group effects** (confounding)

Threats

Unit **15. Player Groups differ in Proficiency**

- **Initial group** differences
- **Design group** differences
- **Motivational** differences

Effect **16. Data Collection** …

Group Different instrument…

Trial **17. Player Groups operate under di…**

Conditions Different OPFOR tactics or environmental conditions

Multiple-group design validity

is enhanced ….

….as **unintended differences** between treatments are **controlled**

PREVENTION examples

- Use randomization or matching.
- Report similarities and differences.
- Use no-treatment control group.
- …nalyze data with/without outliers.
- …tion flow between group.
- …t-trial comparability.
- …tors between groups.
- Use simultaneous presentation when possible.
- Measure trial conditions for comparability.

Test Rigor –

Guidelines for Designing Test Execution

… by eliminating threats to meet 4 Validity Requirements

Test Rigor -- 21 Threats to Test Validity

Five Test Components	Four Test Validity Requirements: 1. Ability to Employ System and Test Factors	2. Ability to Detect Change	3. Ability to Isolate Reason for Change: Single Group	3. Ability to Isolate Reason for Change: Multiple Groups	4. Ability to Relate Test Results to Operations
1. Treatment	(1) System functionality does not work.	(5) System functionality varies in performance.	(11) System functionality changes across trials.	NA	(18) System functionality does not represent future capability.
2. Players	(2) Players are not adequately prepared.	(6) Test players vary in proficiency.	(12) Player proficiency changes across trials.	(15) Groups differ in player proficiency.	(19) Players do not represent operational unit.
3. Effect	(3) Measures are insensitive to capability impact	(7) Data collection accuracy is inconsistent.	(13) Data collection accuracy changes across trials.	(16) Data collection accuracy differs for each group.	(20) Measures do not reflect important effects.
4. Trial	(4) Factors & conditions not adequately implemented.	(8) Trial conditions fluctuate.	(14) Trial conditions change across trials.	(17) Groups operate under different trial conditions.	(21) Scenario is not realistic.
5. Analysis		(9) Low statistical power (10) Statistical assumptions violated.			

Internal Validity -- "Ability to…

1. **…Employ Test System in Planned Conditions**

2. **…Detect Change in Response** MOE/MOP

3. **…Isolate Reason for Change** in Response

External Validity -- "Ability to…

4. **…Relate Test Results** to Military Operations

Rigorous test – *provides evidence to support system-performance conclusion* by eliminating or reducing threats to test validity

ATEC

Test Rigor –

4. Ability to Relate Test Results to Actual Operations

- Given that **System and Test Factors are adequately employed**
- Given that **Response change when Test Factors were changed?**
- Given that **the Treatment alone probably produced change in the Response**

Next Question: **Are these test findings related to actual operations?**

Threat - - magnitude of System Effectiveness in the Test may not be effectiveness in actual operations

Threats

Treatment

18. System Functionality does not represent future capability Not functionally representative

Unit

19. Players do not represent operational w

- Level of training –under-trained or over-trained (golden crew)
- Nonrepresentative players.

Effect

20. Measures do not repres

- Use of SME instead of
  Observer opinion vs
- Inadequate data source
  Single data collector
  Qualitative measures onl

Trial

21. Unrealistic scenario

- **Blue** operations inappropriate
- **Threat** unrealistic
- Unrealistic setting
- Player familiarity with scenario

PREVENTION examples

- Ensure functionality of experimental "surrogate" capability is present.

...ctual end users.
...periment "practice time."
...ined" units

...nission effect (lasers,

...llectors.
...elated quantitative measures

...combat developer accreditation
- Provide adaptive independent accredited threat
- Provide appropriate civilian and military background
- Adaptive "free play" threat enhances scenario setting and uncertainty

Realism in …

…System Functionality,
…Test Players,
… Response Measures,
…Trial Scenario & Execution …..

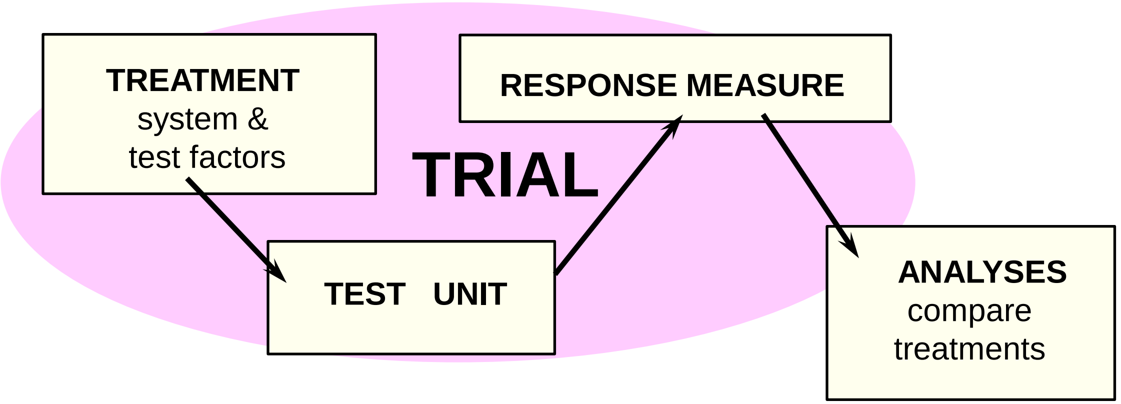
….key to Operational Validity

Army Proven
Battle Ready

Test Event Rigor – Summary

Design 5 Test Components to **reduce/eliminate**

21 Threats to Validity



If as a Result of Test Execution -- the following is demonstrated

- System & test conditions successfully employed
- Response variable changed as factors and conditions changed
- Change in factors and conditions alone caused change in Response Variable
- System performance occurred under operationally relevant conditions

Then, there is **convincing Evidence** that the **test produced Valid conditions & data for DoE analysis.**

Army Proven Battle Ready

Summary

Designing Credible T&E

Doing the
right thing….

Design a Robust T&E Strategy to address the appropriate problem space efficiently

- **Identify all factors/conditions that could affect system performance**
- **Distribute across available evaluation events (DT, OT, M&S)**
- **Design each individual event – using formal DOE techniques**

Design a Rigorous Test to produce valid evidence

- **Design execution of test to …….**
  - … meet the **4 Test Validity Requirements**
  - … by reducing/controlling the **21 Threats to Validity**

Doing the
thing right….