Framework
For
Understanding Experiment Design Validity
(..with potential applicability to OT?)

Presentation to DOTE/OTA DoE Working Group
20 February 2009

Rick Kass
GaN Corporation
Contract Technical Support to –
Transformation Technology Directorate (TTD)
US Army Operational Test Command (USAOTC)
US Army Test and Evaluation Command (USATEC)

Validity fair

Control others, as you would have them control you

Freudian **Analysis**

Rebel without a **Causality**

"Nightmare on **Bias** street"

Discovery channel

"It was the best of **trials**. It was the worst of trials."

Data collection

"**Hypothesis** now"

Realism

Significance other

Who Wants to Be an Experimenter?

Bite-size **sample size**

"I think, therefore I **experiment**"

I **experimented**, but did not inhale.

Testing, 1, 2 , 3…

Serendipity do da

In search of Bobby **Findings**

Science fiction

Free play

Welcome back **Concepts**

Independent, **Dependent**, and Republican **Variables**

Risk-free **Alpha** and **Beta** bonds

Precision

$E=MC^2$ (Experiment equals Methodology Controlled by Confusion)

What is "Experiment Rigor"

To justify recommendations ….
…need "credible experiments"
…experiments with high degree of rigor (…validity…)

How to increase experiment rigor…?

…according to some..

…more realistic scenarios…
…better portrayal of unrestricted and adaptive threats…
…more quantifiable results…

…according to others…

…fewer variables with better control…
…more trials to increase sample size
…use of randomization

Experiment (..and Test …) Rigor is all of these and more…

Outline

Experiment References

Experiment Logic: 2-3-4-5-21

Experiment Rigor Requirements

Threats to Rigor
Good practices to counter threats

Implications of Logic

Design of individual experiments
Campaign of experiments

Useful Definition of Experiment

35 different definitions at "WWW. One-Look Dictionary Search"

Common Themes:

A test done in order to learn something or to discover whether something works or is true (Cambridge Advanced Learning Dictionary). An operation carried out under controlled conditions in order to discover an unknown effect or law, to test or establish a hypothesis, or to illustrate a known law (Merriam-Webster Dictionary)

Experiment –

"To explore the effects of manipulating a variable."

Shadish, Cook, & Campbell,. Experimental and Quasi-Experimental Designs for Generalized Causal Inference (p. 507)

Warfighting Experiment —

"A systematic process to explore the effects of manipulating warfighting capabilities or conditions."

Experiment Rigor References

William R. Shadish, Thomas D. Cook and **Donald T** **mpbell**,. **Experimental and Quasi-Experimental Desi** **eralized Causal Inference** (Houghton Mifflin Co; **200**

Thomas D. Cook and **Don** **. Quasi-Experimentation: Design and Analysis Issue** **ngs** Rand McNally, **1979**)

Donald T. **nd** Julian Stanley. **Experimental and Quasi-Experimen** **esigns for Research** (Rand McNally, **1963**)

Applicable to OT?

Experiment rigor requirements based on 40 years of writing about non-laboratory experiment requirements.

Adapted ideas and terminology for warfighting experiments

Apply traditional scientific principles to Warfighting Experimentation in innovative ways

Outline

Experiments References

Experiment Logic: 2-3-4-5-21

Experiment Rigor Requirements

Threats to Rigor
Good practices to counter threats

Implications of Logic

Design of individual experiments
Campaign of experiments

Experiment Hypotheses

"educated guesses of what might happen"

Useful:

- Help to clarify what experiment is about
- Identify logical thread of the experiment
- Guide experiment design and data collection

Nothing magic:

If _____; then _____.

2

"Two parts to experiment hypotheses"

proposed solution(s) → **problem to be overcome**

independent variable → dependent variable

potential cause → possible effect

Sea Basing → Rapid deployment
Collaboration → Adaptive planning
Global Cell → Inter-theater coordination
Robust ISR → Deny sanctuaries

Different Levels of Hypotheses

Capability Level
(overarching)

If Robust ISR is employed…;
then the threat will have no sanctuaries...

Experimental Level
(measurable-MOE/MOP)

If the Advanced XX System is employed…;
then threat will be continuously tracked.

MOE/MOP

Statistical Level

Ho: T ≥ YY
Ha: T < YY

Logic of Hypothesis Resolution

A **B**

If _____*proposed solution*_____: then _____*problem to be overcome (effect)*_____.

Logic of hypothesis resolution

1. **Did <u>A</u> occur?**
2. **Did <u>B</u> occur?**
3. **Was <u>B</u> due to <u>A</u> ?**

"Three parts to resolving hypotheses"

3

Internal Validity of an experiment

4 *Four Requirements for Rigorous (valid) Experiment*

If New Capability (**A**) ; **Then** Effect (**B**) .

Requirement	Evidence for Validity	Threat to Validity
1 ability to **use new capability**	**A occurred**	**Asset did not work or was not used**
2 ability to **detect change** in effect	**B ch**	**Too much noise, can not detect any change**
3 ability to **isolat** for change	**A alone caused B**	**Alternate explanations of change available**
4 ability to **relate results** to actual operations	**Change in B due to A is expected in actual operations**	**Observed change may not be applicable**

Applicable to OT?

5 *Five COMPONENTS of any EXPERIMENT*

(…5 components of any OT….?)

Treatment (A)

Possible Cause (A)
Independent Variable

System Under Test (SUT)

- new C2 process
- new JTF organization

Effect (B)

Possible Effect (B)
Dependent Variable
Measure of Performance (MOP)
Examples
- targets detected or not
- time from sensor to shooter
- percent objectives met

Trial

Experimental Unit

Smallest Unit Assigned to Treatment
Examples
- sensor operator
- sensor management cell
- Joint Task Force

Analysis

Document CHANGE in B
Examples
- Outcome (B) compared to:
  - different treatments
  - different conditions

21 Threats to a Rigorous Warfighting Experiment

5 Experiment Components	4 Experiment Requirements				
	1. Ability to Use the Capability	2. Ability to Detect Change	3. Ability to Isolate the Reason for Change: Single Group	3. Ability to Isolate the Reason for Change: Multiple Groups	4. Ability to Relate the Results to Operations
1. Treatment	(1) Capability functionality does not work.	(5) Capability systems vary in performance.	(11) Functionality changes across trials.		(18) Functionality does not represent future capability.
2. Players	(2) Players are not adequately prepared.	(6) Experiment players vary in proficiency.	(12) Player proficiency changes [illegible] trial[illegible]	[illegible]ffer	(19) Players do not represent operational unit.
3. Effect	(3) Measures are insensitive to capability impact.	(7) Data collectio[illegible] accuracy is incons[illegible]	[illegible] changes [illegible]oss trials.	(16) Data collection accuracy differs for each group.	(20) Measures do not reflect important effects.
4. Trial	(4) Capabilit[illegible] no oppor[illegible] perform.	[illegible]ns	(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 are violated.	•Experiment Hypothesis: if A, then B. •Purpose of an experiment: verify that A causes B. •Valid experiment allow conclusion "A causes B" to be based on evidence and sound reasoning… -By reducing or eliminating 21 threats to validity.		

Applicable to OT?

21

Outline

Experiments References

Experiment Logic: 2-3-4-5-21

Experiment Rigor Requirements
- Threats to Rigor
- Good practices to counter threats

Implications of Logic
- Design of individual experiments
- Campaign of experiments

Four Requirements To Design Rigorous Warfighting Experiments

Internal Validity

1. **Capability Used**
2. **Detection of Change in Effect**
3. **Isolation of Reason for Change**

External Validity

4. **Relating Results to Military Operations**

1. Ability to Use the New Capability

Most consistent "lessons learned" reported after warfighting experiments completed:

- *New Capability did not work as well as promised.*
- *Players did not know how to use it properly.*
- *The measure (instrumentation) was not sensitive to its use.*
- *The scenario play did not give the players the opportunity to use.*

…sounds familiar for OT?

Ensuring that the experimental capabilities are used and can make a difference is the first logical step in designing a valid experiment.

Threats to the Ability to Use the Capability

THREAT	PREVENTION
Treatment **1. Capability functionality does not work.** Does the HS/SW work?	• Ensure functionality of experimental capability is present.
Unit **2. Players are not adequately prepared.** Do the players have the training and TTP to use the capability?	• Ensure player organized, equipped, and trained for capability use. • Provide sufficient SOPs for capability use. • Provide pre-experiment "practice time."
Effect **3. Measures are insensitive to impact** Is the output sensitive to capability use?	• Pilot-test impact on experiment outcome • "Verify" model input-output logic
Trial **4. Capability has no opportunity to perform.** Does the scenario and MESL call for capability use?	• Pilot-test scenario and MSEL • "White cell" specific scenario injects and monitor for use

Applicable to OT?

Four Requirements To Design Rigorous Warfighting Experiments

Internal Validity

1. **Capability Used**
2. **Detection of Change in Effect**
3. **Isolation of Reason for Change**

External Validity

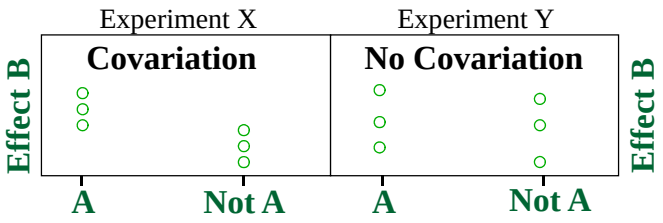
4. **Relating Results to Military Operations**

2. Ability to Detect Change in the Effect

- Given that **A was employed**
- Next Question: **Did B (effect) change when A was applied ?**

Ability to detect change in B: Statistically Valid Experiment

"Detect Change" = "Detect COVARIATION:" B changes when A applied



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**)

2. Ability to Detect Change-- statistical validity

Ability to detect change is enhanced as variability is reduced

Threats

Fail to Detect Change (**Type II Error**)	Treatment	**5. Capability Systems vary in performance** • Continual fluctuation in reliability	• Hold constant
	Unit	**6. Players vary in performance** • Different levels of training • Different reasons for use	• Examine only subset of pop
	Effect	**7. Data collection accuracy inconsistent** • Variation in collectors	ation versus data ctors • Experienced data collectors
	Trial	**8. Trial conditi** • Inadverte	• Set boundary conditions
	Analysis	**9. L ower** sent alpha risk (1% or 5%) **ficient statistical test**	• More repetitions • Increase alpha risk (to 10%) • Use paired comparisons
Incorrectly Detect Change (**Type I Error**)		**10. Statistical Test Assumptions Violated** • Some statistical techniques have sensitive assumptions • Error rate problem (fishing) • Large number of statistical tests	• Use appropriate statistical tool • Select fewer, more meaningful MOPs

Applicable to OT?

Four Requirements To Design Rigorous Warfighting Experiments

Internal Validity

1. **Capability Used**
2. **Detection of Change in Effect**
3. **Isolation of Reason for Change**

External Validity

4. **Relating Results to Military Operations**

3. Isolating the Reason for Change

- Given that **A was employed**
- Given that **B changed as A was applied**
- Next Question: **What really produced the change in B?**

Design Validity -- A alone caused change in B

- Threat -- Something

...also a good 2-type division for OT?

...depends on type of experimental design

Single Group Design

One unit receives all treatment conditions

Same unit

	Scenario 1	Scenario 2
Unit C with Current		
Unit C with Future		

Compare group under different conditions

Multiple Group Design

Different units receive different treatment conditions

Different units

	Scenario 1	Scenario 2
Unit C with Current		
Unit D with Future		

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: Unbalanced

Mon	Tue	Wed	Thu
Current	Current	**Future**	**Future**

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

Current=3 **Future=7**

Sequence 2: Balanced

Mon	Tue	Wed	Thu
Current	**Future**	Current	**Future**

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

Current=4 **Future=6**

Sequence 3: Counterbalanced

Mon	Tue	Wed	Thu
Current	**Future**	**Future**	Current

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

Current=5 **Future=5**

(1 + 0 = 1)

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

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

3. Isolating the Reason for Change
SINGLE-GROUP DESIGN ORDER EFFECTS

THREAT		PREVENTION
Treatment	**11. Capability Functionality changes across trials** System or process improves or degrades over time	• Use fixed configuration
Unit	**12. Player Proficiency changes across trials** Performance improves during later trials due to experience rather than treatment present	• T… aximum … prior to start
Effect	**13. Data Collection Accuracy across trials** Data collect… improve or degrade o… …anging results	…rain data collectors to maximum performance prior to start • Check and recalibrate instrumentation after each trial
Trial	**14. Trial conditi… …nge across trials** Weather, OPFOR, and simulations improve or degrade over time	• Train OPFOR to maximum performance prior to start

Applicable to OT?

Single-group design validity is enhanced as **unintended changes** over time are **controlled**

General prevention/check
- Counterbalance presentation sequence
- Check for increase/decrease over time

3. Isolating the Reason for Change
in
MULTIPLE-GROUP DESIGNS

- **Different player units receive different treatments**

	Scenario 1	Scenario 2
Unit C with Current		B_1
Unit D with Future		B_2

- ***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)

- ***While Multiple-Group designs alleviate Order-Effect threats***

….. A new set of threats arise

- …because now, different treatments are intertwined with different groups
- …difficult to separate treatment effects from group effects
- **…now the differences between capability's might be due to…**
  - inherent personnel differences
  - differences in data collection accuracies
  - differences of trial conditions

…between groups

3. Isolating the Reason for Change
MULTIPLE-GROUP DESIGN UNINTENDED DIFFERENCES

	THREAT	PREVENTION
Unit	**15. Player Groups differ in Proficiency** • Initial group differences •nonrandomized assignment • Evolving group differences •drop-out differences between groups • Design group differences •change after assigning individuals to groups based on past scores • Dominator group differences •one individual can influe score • Motivational differ •initiatio •compensa •resentment	• Use randomization or matching. Report similarities and differences. • Monitor drop outs. • Use no-trea up. s, analyze data with and ers. Distribute information flow between
Effect	**16. Data Collection Accuracy differs for each Player Group** Different instrumentation, SMEs, or data collectors	• Conduct pretrial and posttrial comparability. • Rotate data collectors between groups.
Trial	**17. Player Groups operate under different Trial Conditions** Different OPFOR tactics or environmental conditions	• Use simultaneous presentation when possible. • Measure trial conditions for comparability.

Applicable to OT?

Multiple-group design validity
nhanced as **unintended differences** between treatments
are **controlled**

Four Requirements To Design Rigorous Warfighting Experiments

Internal Validity

1. **Capability Used**
2. **Detection of Change in Effect**
3. **Isolation of Reason for Change**

External Validity

4. **Relating Results to Military Operations**

4. Ability to Relate Results to Actual Operations

DEFINITION

- Given that **A was employed**
- Given that **B changed as A was applied**
- and **A alone probably caused change in B**
- Next Question: **Are these findings related to actual operations?**

Operational Validity:
Experiment effects can be expected in actual combat operations.

Threat - - Amount of change in the outcome measure (B) may not occur in actual combat

Realism in conducting experiment is key to eliminating threat

4. Ability to Relate Results to Actual Operations

Experiment Operational Realism Validation similar to M&S validation

Validation of M&S

"...determining the degree to which a **model** is an accurate representation of the real world..." (DOD VVA Recommended Practice Guide, 1996)

Techniques

Face Validation- experts provide subjective assessments

Operational Validation of Warfighting Experiments

OT …determining the degree to which an ~~**experiment**~~ is an accurate representation of the real world.

Techniques

Prototype Validation
Threat Validation
Scenario Validation
OT ~~**Exercise**~~ **Simulation Accreditation**

Experts provide subjective assessment

Predictive Validation
-comparison to training exercise results (UJTL tasks, conditions, standards)
-comparisons to actual operations

4. Ability to Relate Results to Actual Operations

	THREAT	PREVENTION
Treatment	**18. Functionality does not represent future capability** Not functionally representative	• Ensure functionality of experimental "surrogate" capability is present. •
Unit	**19. Players do not represent operational warfighters** • Level of training --undertrained or overtrained (golden crew) • Nonrepresentative players	• Use actual end us • Provide suff ent "practice time." • Use " ts
Effect	**20. Measures do not repre operational effec** • Use of approx Time w • Inadequat measure Single da or Qualitative measures only	se simulation to address complex measure based on component measure input (model-test-model). • Use multiple data collectors. • Show correlation to related quantitative measures
Trial	**21. Unrealistic scenario** • Blue operations inappropriate • Threat unrealistic • Unrealistic setting • Player familiarity with scenario	• Provide combat developer accreditation • Provide adaptive independent accredited threat •Provide appropriate political and military background • Adaptive "free play" threat enhances scenario setting and uncertainty

Applicable to OT?

Outline

Experiment References

Experiment Logic: 2-3-4-5-21

Experiment Rigor Requirements

Threats to Rigor
Good practices to counter threats

Implications of Logic

Design of individual experiments
Campaign of experiments

Understanding 4 Experiment Requirement provides insights into Experiment Design TRADEOFFS

All Experiments are tradeoffs: -can not eliminate all threats to validity
The 100% valid Experiment does not exist

Internal Validity

Is this true for OT also?

and experiment provides sufficient validity to support the pending decision

A valid experiment is a balance between - -

- Internal validity - - precision and control
- External validity - - representativeness and realism

- Example: increasing repetitions for precision, also increases scenario familiarity thus decreasing realism

Rigorous Experiment Campaigns Requi
Multiple Methods to Meet the Four Requ

…true for large Exp with multiple Initiatives …less true for OT with single SUT?

Requirements for a Good Experiment

	Analytic Wargame Experiments	Constructive Experiments	Human-in –the-Loop Experiments	Field Experiments
Employ Capability	+	+	+ +	+ + +
Detect Change in Effect	+ +	+ + +	+ +	+
Isolate Reason for Effect		+ + +	+ +	
Relate Results to Operations	+		+ +	+ + +

se con for most Rigorous Conclusions

Analytic Wargame Experiments
Human Planners with Simulated Forces
Faster than Real-Time Constructive Simulations

Constructive Experiments
Simulated Forces
Faster than Real-Time Constructive Simulations

Human-in –the-Loop Experiments
Human Planners with Simulated Forces
Real-Time Interactive Simulations

Field Experiments
Actual Forces in Field Environment
Field Exercises & Real-Time Interactive Simulations

Modeling and Simulation Support Required

Combining Simulation Experiments and Wargames
Improve Both
(combining the best of expert discussion and analytic model)

Model-Wargame-Model Paradigm

Pre-Model
Constructive Simulation Experiments

WG-Model
Wargame Simulation
Wargame
+Real staff/operators
+Reactive threat
-Single trial (repeatabi
-No Compa
-N

Post-Model
Construc
Simul

Examine capability tradeoffs in multiple wargame paths

Optimi
•Il al capabilities & scenario decision points to observe

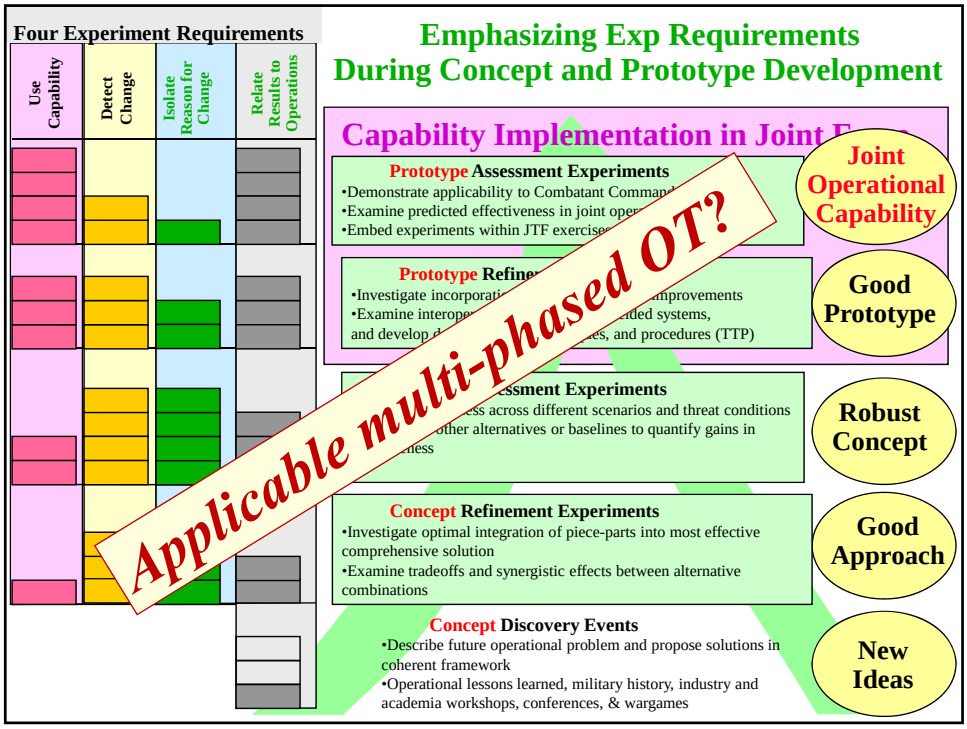
Increase Wargame Play Objectivity
•Adjudicate WG moves
•Provide COA analysis

Conduct Sensitivity Excursion

Increase Rigor
•Isolate why results occurred
•Examine result repeatability

Increase Applicability
•Quantify impact of capabilities and decisions

Applicable to Model-Test-Model Paradigm?



Summary

How to Design a Rigorous Experiment

Understand Experiment Logic: "2, 3, 4, 5 and 21"

Focus individual experiment design on – **OT**

4 OT Validity Requirements

… meeting the **4 Experiment Requirements**

… by eliminate/controlling the **21 Threats to Validity**

Embed individual experiments within an experiment campaign **OT-Phases** **OT**