ITEA Journal 2008; 29: 167–174
Copyright © 2008 by the International Test and Evaluation Association

A Framework for Understanding Experiments

Richard A. Kass, Ph.D.
GaN Corporation,
U.S. Army Operational Test Command, Fort Hood, Texas

While experimentation is an integral aspect of the capability development and acquisition process, its methods may be less familiar to testers. This article provides a framework for understanding the essence of an experiment, its central components, requirements for validity, and programmatic ways to increase experiment validity thru experiment campaigns. A follow-up article will compare experiments to tests.

Key words: Causality, experiment design, experiment campaigns, capability development, model-exercise-model, hypothesis, validity requirements.

An article on experiment techniques[1] should be an interesting read for this audience of testers. When asking test engineers and analysts whether testing and experimenting are similar activities, about half might agree they are similar. A similar question to experimenters located in Service battle labs would find far fewer considering test and experiment similar. The U.S. Department of Defense (DoD) has differentiated between testing and experimenting; tying tests to the acquisition process and experiments to the concept and capability exploration process. So is there a difference in test and experiment techniques?

The answer to this question is in two parts. Readers of this journal are familiar with the nature of testing and test design. This initial article will therefore characterize warfighting experiments and their design requirements. A follow-up article in the next issue will then compare experiments with tests.

Experiments and the capability development process

Tests are conducted on early capability modules, subsystems, prototypes, and production items to quantify the degree of design success. Experiments are also employed throughout this process. Experiments provide a scientific empirical method to identify capability gaps, explore alternative solutions, and develop and continuously update implementation techniques.

Prior to initialization of a capability development process, early experiments identify future warfighting gaps and assess relative merits of proposed doctrine, organization, training, materiel, leadership, personnel, and facilities (DOTLMPF). Analyses of alternatives (AOA) include experiments conducted with combat simulations.

Early in the acquisition process, experiments compare alternative designs and alternative competing solutions. Later, prior to testing of early prototypes, experiments assist combat developers in assessing new tactics, techniques, and procedures (TTP) required for optimizing employment of the new capability. After capability fielding, warfighting experiments can continuously examine opportunities to further enhance capability employment as environments and threats evolve.

Definition of a warfighting experiment

In its simplest formulation, to experiment is to try. In this sense, experimentation is a characteristic of human nature and has existed from earliest times. When early humans attempted different ways to chip stone into cutting edges or selected seeds to grow sturdier crops, they were experimenting.

More formally, "…to experiment is to explore the effects of manipulating a variable." (Shadish, Cook, and Campbell 2002).

This definition captures the basic themes of gaining new knowledge (explore), doing something (manipulating a variable), and causality (the effects). Based on their general definition, the author offers the following derivatives for warfighting experimentation:

Warfighting Experimentation—to explore the effects of manipulating proposed warfighting capabilities or conditions.

Experiment cause and effect and hypotheses

Identifying experiments with the investigation of causality is the key to understanding experiments and

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUN 2008		**00-00-2008 to 00-00-2008**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
A Framework for Understanding Experiments	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **U.S. Army Operational Test Command,91012 Station Avenue,Fort Hood,TX,76544-5068**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**8**	

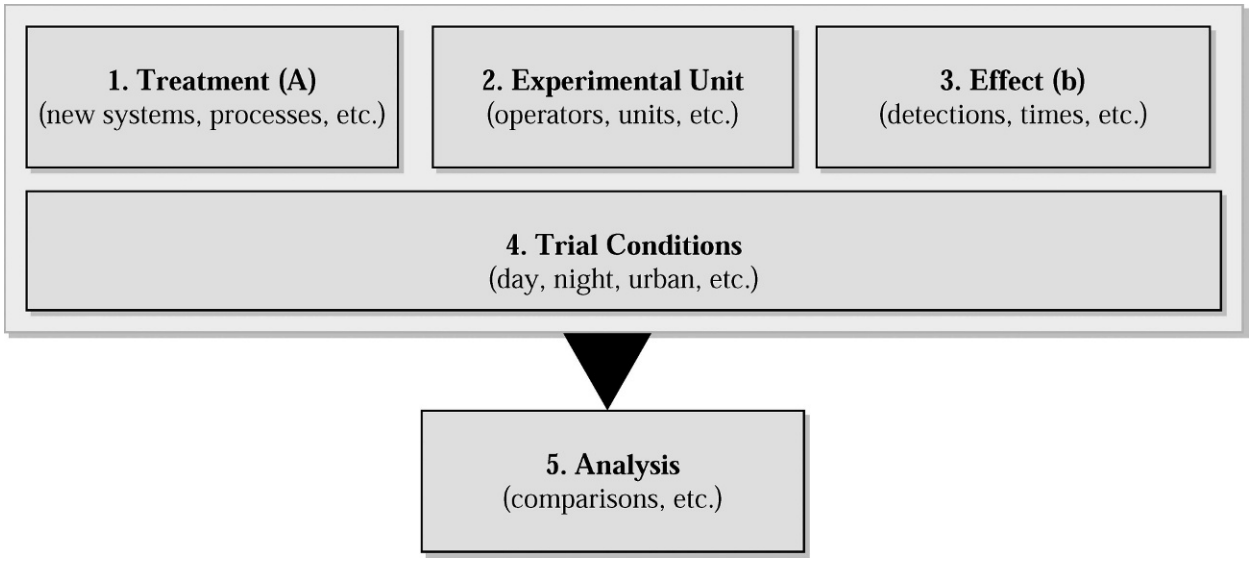

Figure 1. Five elements of an experiment

linking experiments to the transformation process. Causality is central to the transformation process. Military decision-makers need to know what to change in order to improve military effectiveness. The antecedent causes of effectiveness must be understood in order to change effectiveness. Effectiveness is improved by altering its antecedents, its causes. "Today, the key feature common to all experiments is still to deliberately vary something so as to discover what happens to something later—to discover the effects of presumed causes." (Shadish, Cook, and Campbell 2002). The notion of cause and effect is inherent in the language of experimentation and in its basic paradigm "let's do this and see what happens." All warfighting innovation questions can be translated into cause-and-effect questions expressed as: "does A cause B?" Does the proposed military capability (A) produce (cause) an increase in warfighting effectiveness (B)? This theme is fundamental to constructing the experiment hypothesis:

If a unit uses the new capability (A),
then it will increase in effectiveness (B).

Hypotheses are expectations about A causing B. The nature of experiment hypotheses prepares us to understand the five key components common to all experiments.

Five elements of an experiment

In large experiments with many moving parts it is sometimes difficult to see the forest for the trees. All experiments—large or small, field or laboratory, military or academic, applied or pure—can be described by five basic components (Cook and Campbell 1979) as depicted in *Figure 1*; and all five are related to causality.

1. The treatment, the possible cause (A), is the proposed capability, the proposed solution that is expected to influence warfighting effectiveness;
2. The experimental unit executes the possible cause and produces an effect;
3. The possible effect (B) of the treatment is the result of the trial, an increase or decrease in some aspect of warfighting effectiveness;
4. The trial is one observation of the experimental unit employing the treatment (A) or its variation (-A) to see whether effect (B) occurs and includes all of the contextual conditions under which the experiment is executed; and
5. The analysis phase compares the results from one trial to a different trial to quantify the impact of A on B.

Four requirements for a valid experiment

While defense experiment agencies have developed lists of lessons learned and best practices[2] to increase experiment rigor (validity); experiment validity is rarely formally defined. The adjective valid is defined as follows:

"Valid: well-grounded or justifiable, being at once relevant and meaningful, logically correct. [Synonyms: sound, cogent, convincing, and telling.]"—Merriam-Webster Dictionary online, 2006

When this definition is combined with the notion of cause-and-effect, a definition of a valid experiment is apparent: A valid experiment provides sufficient evidence to make a conclusion about the truth or

Hypothesis: If A, then B

Requirement		Evidence of Validity	Threat to Validity
1	Ability to use the new capability.	A occurred.	The asset did not work or was not used.
2	Ability to detect change.	B changed as A changed.	Too much noise. Cannot detect any change.
3	Ability to isolate the reason for the change.	A alone caused B.	Alternate explanations for the change are available.
4	Ability to relate results to actual operations.	Change in B due to A is expected in	The observed change may not be applicable.

Figure 2. Four requirements for a good (valid) experiment

falsity of the causal relationship between the manipulated variable and its effect.

How does one design an experiment to ensure sufficient validity? All of the good practices for designing warfighting experiments can be organized under four logically sequenced requirements[3] that must be met to achieve a valid experiment *(Figure 2)*. A simple example will illustrate these four requirements. Suppose a capability-gap analysis postulates that new sensors are required to detect time-critical targets. An experiment to examine this proposition might be a 2-day military exercise in which the current array of sensors is employed on the first day and a new sensor suite is used on day two. The primary measure of effectiveness is the percent of targets detected. The hypothesis is: "If new sensors are employed, then time-critical target detections will increase." This experiment is designed to determine whether the new sensors (A) will cause an increase in detections (B).

Ability to use the new capability

In most warfighting experiments, the majority of resources and effort are expended to bring the new experimental capability to the experiment. In the ideal experiment, the experimental capability (the new sensor) is employed by experiment players to its optimal potential and allowed to succeed or not succeed on its own merits. Unfortunately, this ideal is rarely achieved in experiments. It is almost a truism that the principal lesson learned from a majority of experiments is that the new capability, not withstanding all effort expended, was not ready for the experiment.

The experimental capability may not be ready for a number of reasons. The hardware or software does not perform as advertised. The experiment players are undertrained and not fully familiar with its functionality. Because it is new, techniques for optimum employment are not mature and by default, will be developed by the experimental unit during the initial experiment trials. If the experimental sensors (A) cannot be functionally employed during the experiment, there is no reason to expect they will detect targets (B) more often than the current array of sensors.

Ability to detect change

If the first experiment requirement is met, then transition from current to new sensors should be accompanied by a change in detections observed. If change in detections does not occur, the primary concern now is too much experimental noise. Ability to detect change is a signal-to-noise problem. Too much experimental error produces too much variability, making it difficult to detect change. Many experiment techniques are designed to reduce experiment variation: calibrating instrumentation to reduce data collection variation, limiting stimuli (targets) presentation to only one or two variations to reduce response (detections) variation, and controlling external environment variations (time of day, visibility, etc.). Sample size also affects the signal-to-noise ratio. Computation of statistical error variability decreases as the number of observations increases.

Ability to isolate the reason for change

Let us suppose the experimenter meets the first two requirements: the new sensors are effectively employed and the experiment design reduces variability and

produces an observable change (increase) in detections. The question now, is the detected change due to the intended cause (changing from old to new sensors) or due to something else. The scientific term for alternate explanations of experimental data is confounded results. In this example, an alternate explanation for any increased detections on day two is that it was due to a learning effect. The sensor operators may have been more adept at finding targets on day two because of their experience with target presentations on day one, and consequently, would have increased target detections on day two whether the sensors were changed or not. This potential learning effect dramatically changes the conclusion of the detected change.

Scientists have developed experimental techniques to eliminate alternate explanations for observed change. These include counter-balancing the presentation of stimuli to the experimental unit, use of placebos in drug research, inclusion of a control groups, and randomizing participants between treatment groups.

Ability to relate the results to actual operations

Again, let us suppose that the experiment is successful in employing the new capability, detecting change, and isolating the cause. The final question is whether experimental results are applicable to operational forces in actual military operations. Experiment design issues supporting generalization include operational realism, representativeness of surrogate systems, use of operational forces as the experimental unit, and use of operational scenarios with a realistic reactive threat.

Tradeoffs in designing experiments

A fundamental implication from these four experiment requirements is that a 100 percent valid experiment is not achievable. The four experiment requirements cannot be fully satisfied in one experiment. Satisfying one works against satisfying the other three. Thus, decisions need to be made as to which validity requirements are to be emphasized in any given experiment.

All experiments are a balance between the four validity requirements. Precision and control increase the ability to detect and isolate change but often lead to decreases in ability to relate results to actual operations. Experiments that emphasize free play and uncertainty in scenarios reflect conditions found in existent operations and satisfy external validity Requirement 4, the ability to relate results. Conversely, experiments emphasizing similar conditions with diminished free play across multiple trials serve to reduce experiment noise and confounding, thus satisfying internal validity Requirements 2 and 3, the ability to detect and isolate change.

Validity priorities differ for any given experiment. Experimenters need to minimize the loss of one validity requirement because of the priority of another. However, tradeoff is inevitable. In settings where one expects a small effect and it is important to determine the precise relationship between the experiment treatment and its effect, the priority should be internal validity. On the other hand, if one expects a large effect and it is important to determine if the effect will occur in the operational environment with typical units, then external validity is the priority.

Different warfighting experiment methods provide different strengths

Warfighting experiments can be grouped into one of four general methods: Analytic war-game, constructive, human-in-the-loop, and field experiments. The experiment requirements just discussed provide a structure for recognizing the strengths and weaknesses of these four experiment methods. Relative strengths in meeting a requirement when employing a particular method is depicted by the number of plus signs in *Figure 3*.

Analytic war-game experiments typically employ command and staff officers to plan and execute a military operation. At certain decision points, the Blue players give their course of action to a neutral, White Cell, which then allows the Red players to plan a counter move, and so on. The White Cell adjudicates each move using simulations to help determine the outcome. Typical war-game experiments might involve fighting the same campaign twice, using different capabilities each time. The strength of war-game experiments resides in the ability to detect any change in the outcome, given major differences in the strategies used. Additionally, to the extent that operational scenarios are used and actual military units are players, war-game experiments may reflect real-world possibilities. A major limitation is the inability to isolate the true cause of change because of the myriad differences found in attempting to play two different campaigns against a similar reactive threat.

Constructive simulation experiments reflect the closed-loop, force-on-force simulation employed by the modeling and simulation community. In a closed-loop simulation, no human intervention occurs in the play after designers choose the initial parameters and then start and finish the simulation. Constructive simulations allow repeated replay of the same battle under identical conditions while systematically varying parameters: insertion of a new weapon or sensor characteristic, employment of a different resource or

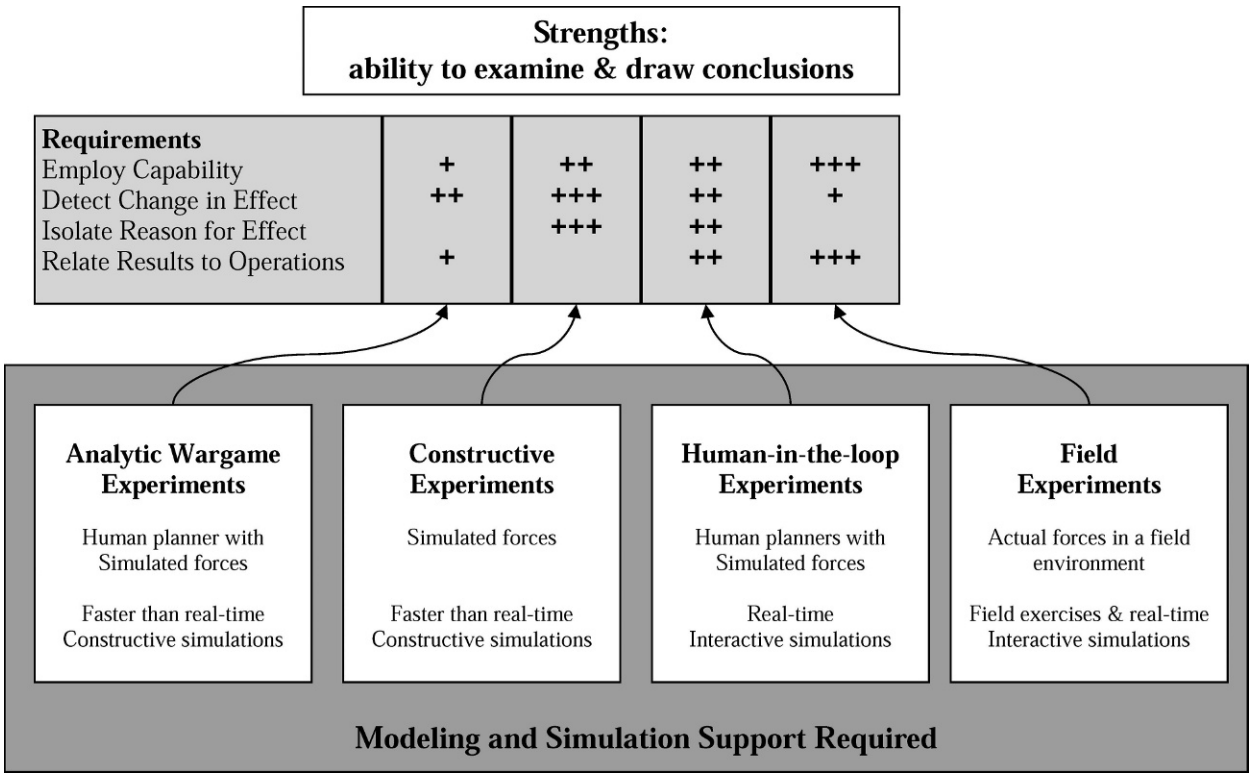

Figure 3. Different experiment venues have different strengths

tactic, or encounter of a different threat. Constructive simulation experiments with multiple runs are ideal to detect change and to isolate its cause. Because modeling complex events requires many assumptions, critics often question the applicability of constructive simulation results to operational situations.

Human-in-the-loop virtual experiments are a blend of constructive experiments and field experiments. In a command and control human-in-the-loop warfighting experiment, a military staff receives real-time, simulated sensor inputs, makes real-time decisions to manage the battlespace, and directs simulated forces against simulated threat forces. The use of actual military operators and staffs allows this type of experiment to reflect warfighting decision-making better than purely closed-loop constructive experiments. However, humans often play differently against computer opponents than against real opponents. Additionally, when humans make decisions, variability increases, and changes are more difficult to detect.

Field experiments are war-games conducted in the actual environment, with actual military units and equipment and operational prototypes. As such, the results of these experiments are highly applicable to real situations. Good field experiments, like good military exercises, are the closest thing to real military operations. A major advantage of the previous three experiment venues is their ability to examine capabilities that do not yet exist by simulating those capabilities. Field experiments, on the other hand, require working prototypes of new capabilities. Interestingly, while field experiments provide the best opportunity to examine practical representations of these new capabilities, field experiments are the most difficult environment to employ a new capability—the new capability has to function and the operators need to know how to employ it. Difficulties also reside in detecting change and isolating the true cause of any detected change because multiple trials are seldom conducted in field experiments and the trial conditions include much of the uncertainty, variability, and challenges of actual operations.

Employing a campaign of experiments to increase validity

Since a single experiment method cannot satisfy all four requirements, a comprehensive experiment campaign is required. A campaign of experiments[4] can consist of a number of successive, individual experiments to fully examine proposed solutions to complex military problems. It can also consist of a set of experiments conducted in parallel with information and findings passed back and forth. A campaign of experiments can accumulate validity across the four requirements.

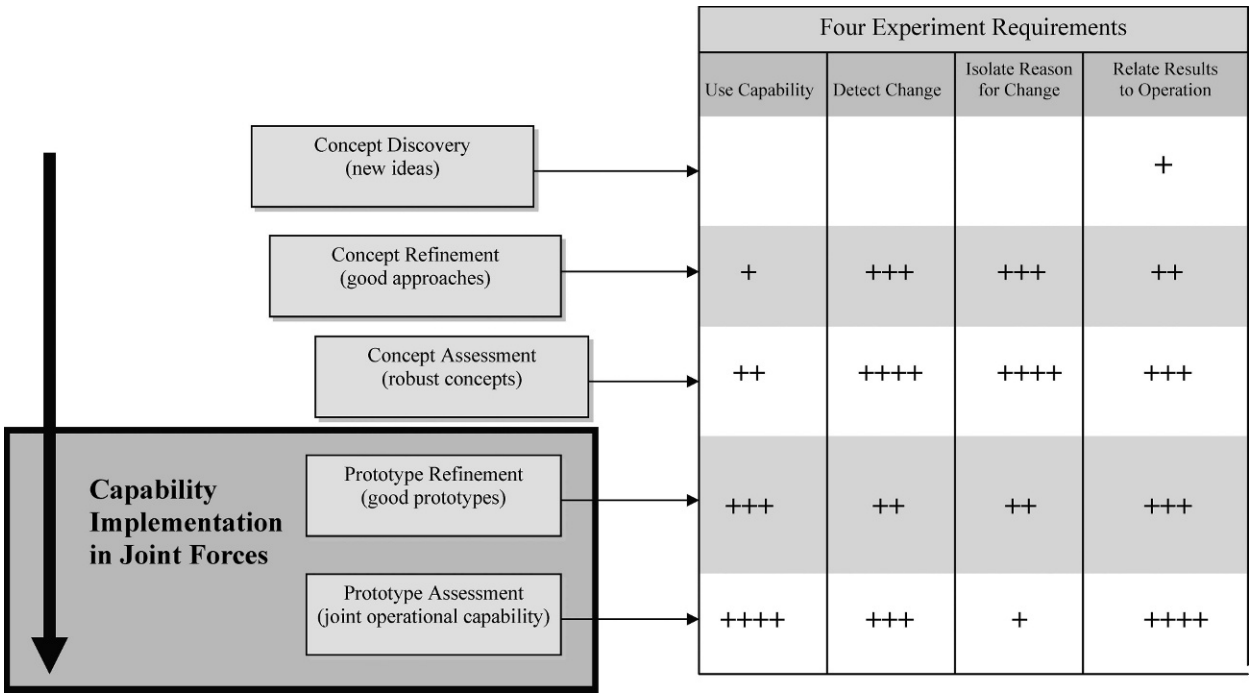

Figure 4. Experiment campaign requirements during the capability development process

Emphasizing different experiment requirements throughout the capability development process

A comprehensive capability-development program should include a campaign of individual experiments that emphasize different experiment requirements. *Figure 4* illustrates one example. The campaign starts at the top with discovery activities and proceeds to the bottom with capability implementation into the joint force. Each step in the campaign identifies possible experimentation goals. On the right of the experiment goals, the "pluses" portray the relative importance of the four validity requirements for that experimentation step. The following discussion identifies possible experiment venues that can be employed at each capability-development step to address the goals and validity requirements.

The primary consideration during concept discovery is relevance and comprehensiveness. To what extent do initial articulations of future operational environments include a comprehensive description of expected problems along with a full set of relevant proposed solutions? Relevancy, however, should not be overstressed. It is important to avoid eliminating unanticipated or unexpected proposals that subsequent experiments could investigate further.

Finding an initial set of potential capabilities that empirically show promise is most important in concept refinement. Early experiments here examine idealized capabilities (future capabilities with projected characteristics) to determine whether they lead to increased effectiveness. Initial concept refinement experiments are dependent on simulations to represent simulated capabilities in simulated environments. Accurately isolating the reason for change is less critical at this stage in order to permit "false positives." Allowing some false solutions to progress and be examined in later experiments under more realistic environments is more important than eliminating potential solutions too quickly. Concept refinement is dependent on the simulation-supported experiment such as constructive, analytic war-game, and human-in-the-loop experiments.

Quantifying operational improvements and correctly identifying the causative capabilities are paramount in providing evidence for concept assessment. Concept justification is dependent on experiments with better-defined capabilities across multiple environments. Constructive experiments can provide statistically defensible evidence of improvements across a wide range of conditions. Human-in-the-loop and field experiments with realistic surrogates can provide early evidence for capability usability and relevance. Incorporating human decision-makers into human-in-the-loop and field experiments is also essential early in the capability-development process. Human operators tend to find new ways to solve problems.

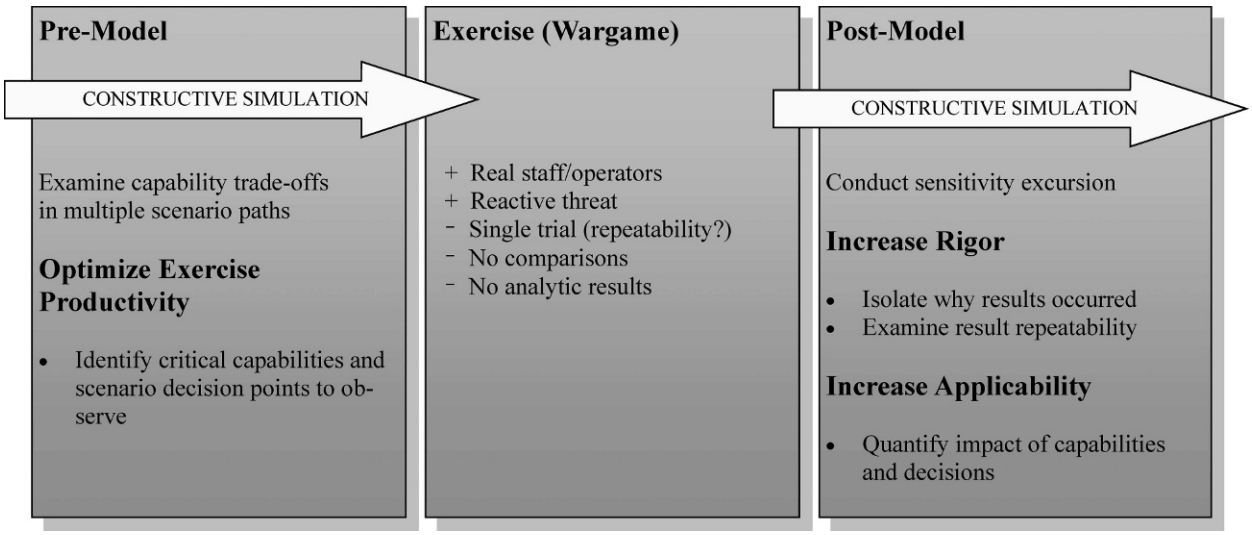

Figure 5. Model-exercise-model process

In prototype refinement, one should anticipate large effects or the implementation might not be cost effective. Accordingly, the experiment can focus on the usability of working prototypes in a realistic experiment environment. To do this, the experiment must be able to isolate the contributions of training, user characteristics, scenario, software, and operational procedures to prototype improvements in order to refine the right component. Human-in-the-loop and field experiments with realistic surrogates in realistic operational environments provide the experimental context for assessing gains in effectiveness. Human operators find unexpected ways to employ new technologies effectively.

Applicability to the warfighting operational environment is paramount in prototype assessment. If the capability is difficult to use or the desired gains are not readily apparent in the operational environment, it will be difficult to convince combatant commanders to employ it. Uncovering exact causal chains is less important while human operators are essential to ensuring that the new technology can be employed effectively. Prototype assessment experiments are often embedded within joint exercises and operations.

Emphasizing different experiment requirements via a model-exercise-model process

Another type of experiment campaign can be organized around the requirement to conduct large war-games or large field exercises to investigate the effectiveness of new capabilities. Because these large events are player resource intensive and often include multiple experimental capabilities, few opportunities exist to examine disentangled alternative capabilities or alternative situations that would allow meaningful comparisons. The model-exercise-model paradigm depicted in *Figure 5* can enhance the usefulness of war-games and exercises. This paradigm consists of conducting early constructive simulation experiments prior to the war-game or exercise and then following these events with a second set of postexercise constructive experiments.

Early constructive simulation experiments use the same Blue and Red forces anticipated to be played in the exercise. This pre-event simulation examines multiple alternative Blue force capability configurations against different Red force situations. This allows experimenters to determine the most robust Blue force configuration across the different Red force scenarios. It also helps to focus the exercise by pinpointing potential critical junctures to be observed during the follow-up exercise.

The war-game or exercise executes the best Blue force configuration identified during the pre-event simulation. The "best configuration" is the one indicated by pre-exercise simulation that the new capability dramatically improved Blue's outcome. The exercise reexamines this optimal configuration and scenario with independent and reactive Blue and Red forces. Choosing the scenario that provides the best opportunity for the new capabilities to succeed is best because large exercises include the "fog of war"—and experimental capabilities rarely perform as well in the real environment as in simulation. Therefore, it makes sense to give the new capability its best chance to succeed. If it does not succeed in a scenario designed to allow it to succeed, it most likely would not succeed in other scenarios.

Experimenters use the exercise results to calibrate the original constructive simulation for further poste-

vent simulation analysis. Calibration involves adjusting simulation inputs and parameters to better match the play of the simulation to the play of the exercise. This adds credibility to the simulation. Rerunning the pre-event alternatives in the calibrated model provides a more credible interpretation of differences now observed in the simulation. Additionally, the postevent calibrated simulation can substantiate (or not) the implications of the exercise recommendations by conducting causal analysis. Causal analysis is a series of "what if" sensitivity runs in the simulation to determine whether the exercise recommendations make a difference in the calibrated simulation outcome. Postexercise simulation runs can also examine what might have occurred if the Red or Blue forces had made different decisions during the exercise.

Summary

Can experiments fail? Yes, they can fail to provide sufficient evidence to determine whether the manipulated variable does (or does not) cause an effect. If the experimenter is unable to answer each of the four requirements in a positive manner, a meaningful conclusion is not possible concerning the impact of a proposed capability.

Designing individual warfighting experiments is an art because every experiment is a compromise. The logical approach in this article provides an understanding of the choices available to meet the four experiment validity requirements and the strengths and weaknesses inherent in typical experiment venues. Designing an individual experiment involves making cognizant tradeoffs among the four requirements to provide sufficient credible evidence bounded by explicated limitations to resolve the hypothesis.

While a single experiment will not satisfy all four requirements, a campaign of experiments can accumulate validity and overall confidence in experiment results. A comprehensive experiment program includes a series of individual experiments, each emphasizing different experiment requirements. In this campaign, no single experiment is expected to carry the entire weight of the decision. Each experiment contributes and the final results are based on accumulated confidence with each individual experiment contributing its strength to the final conclusions. The whole is greater than any part.

So, how much of this is applicable to acquisition testing? The follow-up article in the next issue will discuss the similarities and difference between tests and experiments in several areas: The planning process—especially designing valid tests and experiments—along with the execution and reporting process. The next article will focus on clearing away misperceptions of where efficiencies could be gained by sharing resources and expertise. ❑

Rick Kass has 25 years in designing, analyzing, and reporting on operational field tests and military experiments. He held multiple positions as test officer, analyst, and test director for 18 years with the U.S. Army Test and Evaluation Command (USATEC) and was chief of analysis for 7 years with the U.S. Joint Forces Command (USJFCOM) joint experimentation program. Currently, Rick works for GaN Corporation supporting the Army's Operational Test Command at Fort Hood, Texas. He has authored over 25 journal articles on methods for research, experimentation, and testing and was the primary architect establishing the permanent Warfighting Experimentation Working Group in the Military Operations Research Society (MORS). Rick is a graduate of the National War College and holds a Ph.D. in psychology from Southern Illinois University. E-mail: rick.kass@us.army.mil

Endnotes

[1]This article draws heavily from portions previously printed in my book Kass, R. A. The Logic of Warfighting Experiments published in 2006 by the Command and Control Research Program (CCRP) of the ASD/NII which has graciously granted permission to include that material in this work. *Figures 1* through *5* here are *Figures 9, 8, 20, 39*, and *40* in that work. Readers can download or order the larger document from the CCRP website at http://www.dodccrp.org.

[2]A good discussion of many best-practices is found in Alberts, D. S. and Hayes, R. E. 2002 Code of Best Practices for Experimentation. DoD CCRP publication series, D.C.: U.S. Government Printing Office.

[3]*The Logic of Warfighting Experiments* devotes a separate chapter to each of the four validity requirements.

[4]For a comprehensive examination of the value of experiment campaigns to address warfighting problems see Alberts, D. S. and Hayes, R. E. 2005 Campaigns of Experimentation. DoD CCRP publication series, D.C.: U.S. Government Printing Office.

References

Cook, T. D. and Campbell, D. T. 1979. Quasi-Experimentation: Design and Analysis Issues for Field Settings. Houghton Mifflin: Boston.

Merriam-Webster Dictionary online. http://www.m-w.com

Shadish, W. R., Cook, T. D. and Campbell, D. T. 2002. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton Mifflin: Boston. Page 507.

Shadish, W. R., Cook, T. D. and Campbell, D. T. 2002. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton Mifflin: Boston. Page 3.