Guide for Understanding and Implementing Defense Experimentation GUIDEx The Technical Cooperation Program

Book · February 2006

DOI: 10.13140/2.1.4937.6648

CITATIONS
26

READS
3,596

10 authors, including:

Paul Labbé
Defence Research and Development Canada
36 PUBLICATIONS **119** CITATIONS

SEE PROFILE

Richard A. Nunes Vaz
Defence Science and Technology Group (DST)
31 PUBLICATIONS **1,267** CITATIONS

SEE PROFILE

Guide for Understanding and Implementing Defense Experimentation

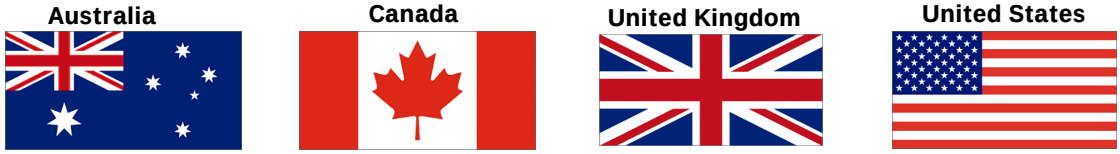
GUIDEx

The Technical Cooperation Program

The Technical Cooperation Program

Subcommittee on Non-Atomic Military Research and Development (NAMRAD)
The Technical Cooperation Program (TTCP), Joint Systems Analysis (JSA) Group,
Methods and Approaches for Warfighting Experimentation Action Group 12 (AG-12)
(TTCP JSA AG-12)

Guide for Understanding and Implementing Defense Experimentation (GUIDEx)



Version 1.1, February 2006

Any product or trademark identified in this document provides an example not a recommendation.

GUIDEx does not present the official policy of any participating nation organization. It consolidates principles and guidelines for improving the impact of science-based experimentation on defense capability development. All organizations are invited to use the guidance it provides.

**Funds for printing this document
were provided by the
Canadian Forces Experimentation Centre,
Ottawa, Canada.**

Electronic copy compiled on 10 February 2006, Ottawa

ISBN 92-95046-11-0

Guide for Understanding and Implementing Defense Experimentation (GUIDEx)

Keywords:

experiment, trial, test, hypothesis, analysis, causal, cause-and-effect, defense, warfighting, simulation, wargame, training, exercise, ethics, valid, requirement, capability, development, measure, instrument, unit, joint, force, campaign.

Contact: ttcp_docfeedback@dtic.mil

This printed version of GUIDEx includes an introduction to the experimentation flowchart and few minor changes to the September 2005 TTCP Website electronic version. The posted February electronic version of GUIDEx includes all these changes.
http://www.dtic.mil/ttcp/

Foreword

The development of allied forces has always been a difficult and complex process. However the need for force development to respond to asymmetric and unpredictable threats, the demands of coalition operations, the perceived need for information supremacy, combined with evolving transformational technologies and concepts, have caused this task to become even more difficult over the past few years. Experimentation offers a unique means to support the development and transformation of allied forces by advancing our knowledge of the complex networked systems and capabilities likely to be fielded in the near future.

"Anything we use today arrives through a process of organized experimentation; over time, improved tools, new processes, and alternative technologies all have arisen because they have been worked out in various structured ways" [Thomke 2003: p. 1]**.**

The growing importance of experimentation motivated TTCP's Joint Systems and Analysis Group (JSA) to establish Action Group 12 on ***Methods and Approaches for Warfighting Experimentation*** in 2002. The work of AG-12 over the past three years has culminated in this guide for defense experimentation. It is based on **14 Principles** to ensure that allied defense experimentation programs are genuinely able to support the evolution of the force capabilities of the future. For the benefit of readers a set of real-world Case Studies is provided to illustrate the 14 Principles in practice. They also provide further material for devising a way ahead for accelerating the acquisition of knowledge to maintain a leading advantage in military capabilities.

Although this guide has been written mainly for the practitioners and designers of defense experimentation, we hope that it will stimulate better communication among military officers, government officials and the defense scientific communities of the allied nations on all matters associated with defense experimentation. Additionally, the experimentation Principles described in this guide apply to other large enterprises and multiple agency operations, for example in homeland security.

This document is complementary to existing references in the domain of experimentation [ABCA 2004; Alberts and Hayes 2002, 2005; Dagnelie 2003; Radder 2003; Shadish, Cook and Campbell 2002] and systems assessment [NATO 2002]. It is the result of collaborative activities conducted under the TTCP umbrella that included: several workshops with members of JSA, Human Resources and Performance (HUM), and Maritime Systems (MAR) Technical Panels and Action Groups; interactions with the American, British, Canadian, and Australian (ABCA) Armies program; collaboration with the NATO Research and Technology Organisation (RTO), including an international experimentation symposium organized by the Australian Defence Science and Technology Organisation (DSTO); and the direct and indirect contributions by experts of the participating countries.

Paul Labbé
Chair, TTCP JSA AG-12

Who should Read GUIDEx?

Those who ask force capability questions and act on the answers.

GUIDEx increases the decisionmaker's confidence by promoting methods for adequate coverage of the defense problem space while providing a traceable, logical and valid path to recommendations. GUIDEx rigorously applies the scientific method for experimentation. GUIDEx shows how experimentation can deliver timely answers with a measured level of confidence, thereby contributing to sound risk management of programs and their components. It thoroughly supports defense problem solving from concepts through capability development to operations.

Those who decide how the force capability question is to be addressed and what methods are to be used.

There are three categories of issues that concern this type of decisionmaker:

1. **Fitness for purpose of the method(s) chosen.** This is fundamental. The method selected must be demonstrably and transparently capable of answering the question and capable of supporting the decision process for selecting options. The method must be able to stand up to scrutiny and peer review. The decisionmaker will also wish to be seen running as rigorous and effective a program as possible.
2. **Programmatics.** Practical programmatic issues, such as cost, timescale and internal resources, will always impose constraints. They require an optimal use of limited internal resources, including expert personnel, specialists' facilities, and the availability of military players and subject matter experts (SMEs).
3. **Wider program synergies.** The military problem in question will not be submitted to the analytical community as their sole problem in isolation. Other questions may already be dominating particular exercises or guiding programs of wargames and simulations.

Decisionmakers will typically be faced with selecting the best experimentation option to address the question and selecting the particular methods to be applied. GUIDEx provides information on whether a campaign of coordinated experimental and analytical activities is required, or if a single experiment will suffice. GUIDEx's four requirements for a good experiment and 21 threats to good warfighting experiments provide the framework required to support decisions and they can also provide the means for evaluating options and methods in terms of fitness for purpose. This should be the overarching consideration in deciding upon a program of experimentation and other methods. The decisionmaker will also be concerned with programmatics, which can result in a showstopper and lead to a re-evaluation of how the question is to be addressed but this is not the central focus of GUIDEx.

Those who design, execute, and interpret defense (warfighting) experiments.

Too often, what is learned most in defense experiments is how to do the next experiment better. The intent of this guide is to help design experiments right the first time by avoiding common pitfalls. The primary concern that arises when someone is assigned to develop and execute an experiment is how to design an experiment that is both valid and cost effective.

This guide provides insights into the best way to design experiments that have sufficient validity to address the hypothesis in terms of capability effectiveness. It provides a framework for organizing and focusing all of the experiment good practices to achieve a valid experiment. It also provides a way to examine tradeoffs when designing experiments since it is never possible to apply all of the good experiment techniques in a single experiment, *e.g.*, what is the ideal composition of the groups of subjects for the experiment. This guide focuses on the ultimate goal of an experiment such that when the experiment is complete the results will be pertinent to the capability under investigation, and that any positive results are clearly due to that capability. It also shows the best ways to ensure the experiment produces a measurable result and that the results will be applicable to the operational military environment. This guide also treats many of the modeling and simulation (M&S) and other resources available to expand the experimenter's area of application.

Furthermore GUIDEx provides practical guidance on a broad range of experiment implementation issues pertaining to human variability, ethics, international and political concerns and communication with stakeholders. Many examples and eight substantive Case Studies are provided to illuminate the points made in GUIDEx.

Those engaged in Operational Test & Evaluation (OT&E).

While GUIDEx focuses on experimentation, there is a wide applicability of the principles and techniques to those involved in OT&E. Those in this community are also concerned with planning and executing cost-effective and valid tests throughout the acquisition process. GUIDEx gives practical insights into how to design operational assessments and field tests, which integrate experimentation, modeling and simulation.

Executive Summary

This document provides critical guidance to support successful defense experimentation. It has been produced by defense experimentation expert representatives from the defense science and technology (S&T) organizations in Australia, Canada, the United Kingdom and the United States under the auspices of The Technical Cooperation Program (TTCP), Joint Systems Analysis (JSA) Group, Action Group (AG) 12 on *Methods and Approaches for Warfighting Experiments*. JSA-AG-12 worked from March 2002 until July 2005. It produced this *TTCP Guide for Understanding and Implementing Defense Experimentation* (*GUIDEx),* which describes 14 Principles leading to valid (good) experimentation that are amplified through 8 Case Studies drawn from the participating nations and coalitions. It has been prepared in three parts; Part I provides an introduction and overview to the 14 Principles, Part II contains the full explanation of the Principles, and Part III presents the Case Studies.

The reader is encouraged to apply and adapt the 14 Principles laid out in this Guide. However, many examples within the guide are based on the specific perspective and experience of different lead-nation authors with contributions from other participants. Therefore they may require supplementary effort to relate them to national perspectives.

The main thesis of GUIDEx is that, while it is true that defense experiments are not like highly abstracted and inanimate laboratory experiments, the logic of science and experimentation can be applied to defense experiments to produce credible tests of causal claims for developing effective defense capabilities. The collaboration in this forum has produced a detailed and practical guide for the conduct of experiments in the TTCP countries. While the context and examples are relevant to these countries, the Principles can be applied by any organization engaged in defense experimentation. The 14 Principles are organized into three themes:

1. **Designing Valid Experiments**,
2. **Integrated Analysis and Experimentation Campaigns**, and
3. **Considerations for Successful Experiments**.

The 14 Principles for designing valid experiments provide a solid foundation for using the scientific method to establish cause-and-effect relationships for hypothesized military capabilities.

1. The thesis of Principle 1, *defense experiments are uniquely suited to investigate the cause-and-effect relationships underlying capability development*, is that a capability change (cause) should result in a difference in military effectiveness (effect). When change is observed under controlled conditions, a conclusion about cause-and-effect is possible.
2. Principle 2, *designing effective experiments requires an understanding of the logic of experimentation*, develops the logic of defense experiments by describing the elements of the experiment hypothesis; the resolution of the conditional proposition in the hypothesis statement; the requirements for a valid experiment; and the threats to drawing valid causal inferences from

the experiment. This provides a framework for understanding the options and making the tradeoffs in designing a valid experiment. The **four requirements for a valid experiment** are the ability to 1- **use the new capability**, 2- **detect a change in effect**, 3- **isolate the reason for the change**, and 4- **relate the results to actual operations**.

3. Principle 3, *defense experiments should be designed to meet the four validity requirements*, discusses GUIDEx recommended experiment techniques to counter the threats to these requirements. All defense experiments should be designed to meet these four requirements. However, attempts to satisfy one of the requirements often work against satisfying the others.

A campaign of experiments including analytical activities will generally be required. Integrated analysis and experimentation campaigns are described by the four Principles in the second theme.

4. Principle 4, *defense experiments should be integrated into a coherent campaign of activities to maximize their utility*, describes the need for coherent sequences of experiments, combined with other methods of knowledge generation, based upon metrics derived from the characteristics of the problem. Campaigns should include a management and communication framework, as well as an analytical program.

5. Principle 5, *an iterative process of problem formulation, analysis and experimentation is critical to accumulate knowledge and validity within a campaign*, argues that such a process is required in an integrated analysis and experimentation campaign. The key aspect of the process, problem formulation, should aim to decompose capability development problems into components that can be addressed with specific analytical techniques and/or studies or with an integrated analysis and experimentation campaign. The analysis accumulates validity through the course of the campaign and can provide information to decisionmakers at any stage of the process.

6. Principle 6, *campaigns should be designed to integrate all three scientific methods of knowledge generation (studies, observations and experiments)*, advocates the integration of all three scientific methods of knowledge generation in campaigns; *rational-deductive*, in the form of studies; *empirical-inductive*, in the form of careful observation of real-world events; and *experiments (empirical-deductive)*, manipulation of events to isolate cause-and-effect. GUIDEx **focuses on experiments** and their role within capability development.

7. Principle 7, *multiple methods are necessary within a campaign in order to accumulate validity across the four requirements*, shows how understanding the four validity requirements, in Principle 3, is essential for appreciating the strengths and weaknesses of the primary methods used for defense experiments and ensuring appropriate application. The Principle describes four methods for conducting experiments; analytic wargames, constructive simulations, human-in-the-loop simulations, and live simulations (or field experiments). The best strategy is to construct integrated analysis and experimentation campaigns using multiple methods. This Principle also describes how the model-exercise-model (M-E-M) paradigm which exploits different methods is used to increase rigor when resource constraints prohibit conducting side-by-side baseline and alternative comparisons during wargames and field experiments.

The remaining 7 Principles, considerations for successful experiments, provide expert advice to support the practical implementation of defense experiments. These address issues such as human variability in experiment design, modeling and simulation (M&S) methods in experiments, the implementation of good experiment control, ethics, and advice on communications with stakeholders.

8. Principle 8, *human variability in defense experimentation requires additional experiment design considerations*, provides an insight into the effects of human variability on defense experiment observations. An understanding of these impacts is a fundamental skill required by all experimenters.

9. Principle 9, *defense experiments conducted during collective training and operational test and evaluation require additional experiment design considerations*, argues that experimenting during training exercises and operational test and evaluation (OT&E) events can provide cost-effective opportunities since considerable infrastructure is typically provided. Most nations generally do not have units and formations available to dedicate to experimentation; therefore exploiting routine training exercises and other collective events should be given serious consideration.
10. Principle 10, *appropriate exploitation of modeling and simulation (M&S) is critical to successful experimentation*, discusses how virtual simulations and analytic wargames offer an immersive, safe environment in which to analyze operational activities and conduct experiments. It is estimated that as much as 80% of defense experiments employ M&S in some fashion. However, the all-pervasiveness of simulation is not without practical problems of costs, required validity, level of effort and scarcity of expert personnel. Consequently the appropriate use of M&S is vitally important for successful experimentation.
11. Principle 11, *an effective experimentation control regime is essential to successful experimentation*, asserts that defining experiment controls is primarily a scientific activity to be undertaken during the design phase, while implementing those controls is a complex management activity which needs to be undertaken during the planning and execution phases. This Principle argues that control must be applied from start to finish, from concept development through analysis and reporting.
12. Principle 12, *a successful experiment depends upon a comprehensive data analysis and collection plan*, emphasizes the importance of adequate data analysis and collection planning. This directly affects the knowledge that can be gained from an experiment or campaign. For a causal hypothesis, controls are necessary to rule out plausible rival explanations, and these need to be considered in the data analysis and collection plan.
13. Principle 13 asserts that *defense experiment design must consider relevant ethical, environmental, political, multinational, and security issues*. Ethical issues as well as health and safety issues are especially important for any experiment involving human subjects and human data collectors.
14. Principle 14, *frequent communication with stakeholders is critical to successful experimentation*, completes GUIDEx with the advice that every integrated analysis and experimentation campaign should have a champion; otherwise it may fail to have any real impact on operational systems or future capabilities. This Principle highlights the issues associated with communication plans.

GUIDEx Case Studies provide a variety of exemplary experiments and a description in each of the relationships to the 14 Principles. They demonstrate the feasibility of conducting informative experiments on the effectiveness of defense capabilities. They also show how the validity requirements, the threats to good experiments, and the simulation methods have been applied. The inter-relationships between the Case Studies and the simulation methods are provided in Table 5 page 234. These substantive Case Studies are summarized here to help the reader appreciate the depth and breadth of this section of GUIDEx.

1. Case Study 1: *Testing Causal Hypotheses on Effective Warfighting*. This Case Study documents a series of experiments on a Common Operational Picture (COP) prototype technology using a Persian Gulf combat scenario. They clearly demonstrate that a team's use of the COP resulted in greater shared situation awareness and combat effectiveness. These experiments were consistent with Principles 1, 2 and 3 and it was found that the causal hypothesis was strongly supported by the experimental evidence. That there was a three-year hiatus between the completion of the experiments and the onset of official technology adoption and engineering,

indicates the project could have benefited from earlier and more effective communication with the decisionmaker, Principle 14.

2. Case Study 2: *UK Battlegroup Level UAV Effectiveness*. This experiment supported a major UK unmanned air vehicle (UAV) acquisition program in demonstrating the huge information gathering potential of UAVs at the tactical level, compared to existing intelligence, surveillance, target acquisition and reconnaissance (ISTAR) assets. The Case Study illustrated how one can both make the most out of scarce resources and maintain internal validity by piggybacking experimentation activities onto collective training exercises, using properly tailored design (Principle 9). This Case Study also showed how simple M&S can be used in conjunction with live action to achieve some of the benefits of both experiments using human-in-the-loop simulation and field experiments.

3. Case Study 3: *UK NITEworks ISTAR Experiment*. This experiment investigated both technological and procedural means of improving information requirements management (IRM). It showed conclusively that a collaborative working environment with appropriate working practices would have a major beneficial effect on IRM effectiveness. Principles 2 and 3 were demonstrated well in the design of this experiment, but it was shown that the addition of the M-E-M paradigm (Principle 7) would have been beneficial. It was concluded that a possible avenue to further increase external validity would be to conduct a follow-on experiment in a different venue (Principle 7) such as in a field exercise large enough for collaboration in a larger headquarters setup.

4. Case Study 4: *Pacific Littoral ISR UAV Experiment (PLIX)*. This Canadian Case Study provides insights difficult to capture without experimentation; the strong hypothesis of identifying and tracking all targets proved not to be attainable even though sensor coverage was nominally complete, pointing to integration requirements for an effective ISR architecture. This experiment is a good example of the importance of Principles 4 and 5 in that it benefited from being part of a campaign of progressively more complex experiments. The experiment could have been improved, however, by more attention to meeting the four validity requirements, Principle 3. This case shows that exclusive reliance on live experiments may have limited return on investment since other requirements could have been identified in a controlled environment, Principle 7.

5. Case Study 5: *An integrated analysis and experimentation campaign: Army 21 / Restructuring the Army 1995-99*. This Australian campaign demonstrated the importance of detailed problem definition and an iterative approach based on wargaming, field trials and analytical studies. The warfighting concept under test was found to fail under realistic environmental constraints. However the results led to an alternative concept which is the basis for current Australian Army force development. This Case Study showed the advantage of early communication with the customers to develop a commonly agreed and understood definition of the problem, Principle 14.

6. Case Study 6: *The Peregrine Series: a campaign approach to Doctrine and TTP development*. This on-going campaign of experiments and studies is contributing directly to the development of the doctrine for employment of the Australian Army's new Armed Reconnaissance Helicopters and demonstrates how experimentation can be used to inform capability development questions at unit level and below. This Case Study illustrates the advantages of a campaign, Principles 4 to 6, and demonstrates how a less controlled, exploratory experiment can be used with a number of more focused events to build validity, Principle 7.

7. Case Study 7: *Multinational Experiment Three (MNE 3)*. Despite the complexity of the MNE 3 effects-based planning (EBP) experiment, the event demonstrated the potential for EBP to make a coalition task force a more effective instrument of power and also showed the benefits for collaboration in a coalition. This experiment had strong external validity (Principle 3) through its use of an operational scenario, database, and military personnel from various nations. It also demonstrated how emphasis on external validity makes it difficult to achieve internal validity and that in designing an experiment one needs to find balance among the four validity requirements.

8. Case Study 8: *Improved Instruments Increase Program Values*. This multinational Case Study describes an integrated analysis and experimentation campaign within an AUSCANNZUKUS[1] Program that investigated the management of organic and non-organic information in a maritime environment. This example demonstrates Principle 6 by not relying exclusively on one scientific method of knowledge generation, but by exploiting all of them (experiments, studies and observations). Its success is also due to an iterative process to reach agreements between analysts and stakeholders (Principle 5), special considerations in exploiting collective training (Principle 9), exploitation of M&S (Principle 10), extensive data analysis and collection plans (Principle 12), and reporting to stakeholders (Principle 14). In addition, a critical activity initiated by Canada allowed detecting effects in operations of some of the interventions. Observations of these effects were not possible otherwise.

The Case Studies exemplify many key points in GUIDEx. They provide material to guide organizations that plan to use experimentation to accelerate the acquisition of knowledge to support capability development programs. Such organizations should consider expanding the core competencies in experimentation and contributing to the advance of methods and approaches for defense experimentation.

As previously stated, GUIDEx has been written in three parts; Part I is an overview and introduction to the 14 Principles, Part II presents the 14 Principles in detail, and Part III presents the Case Studies. Part I, with the addition of the mapping of the 21 threats to good defense experiments, has been designed to double as a standalone GUIDEx pocketbook on defense experimentation and will be known as the Slim-Ex. This is intended to be a guide for clients, people who ask the questions that lead to experiments and campaigns and for whom reports are prepared. It is also for those who decide how the question will be addressed and approve the methods that will be applied. Parts II and III, the main body of GUIDEx, is for the people who design, execute, analyze and report on experiments. These experimenters are the backbone of the community and should benefit from the full detail of the 14 Principles.

[1] Counterpart of the TTCP information exchange agreement for operational and in-development systems, related standards and problems. Member countries: Australia, Canada, New Zealand, United Kingdom and United States.

Preface and Acknowledgments

The preparation of this document would not have been possible without selected collaborative activities conducted under the TTCP umbrella that included meetings, conferences and workshops with participation from JSA, HUM and MAR (group, technical panel and action group members), interactions with ABCA, NATO RTO and Allied Command Transformation (ACT), and the direct and indirect contributions by experts of the participating countries.

The participants of TTCP JSA AG-12 with the collaboration of several experts produced[2] this document. GUIDEx authors are listed below by alphabetical order of family name.

Bowley, Dean	Defence Science & Technology Organisation (DSTO)	AU
Comeau, Paul	Canadian Forces Experimentation Center (CFEC)	CA
Edwards, Dr Roland, NL[3]	Defence Science and Technology Laboratory (Dstl)	UK
Hiniker, Dr Paul J.	Defense Information Systems Agency (DISA)	US
Howes, Dr Geoff, NL	Defence Science and Technology Laboratory (Dstl)	UK
Kass, Dr Richard A., NL	Joint Forces Command (JFCOM), Experimentation	US
Labbé, Paul, Chair	Defence Research & Development Canada (DRDC)	CA
Morris, Chris	NITEworks	UK
Nunes-Vaz, Dr Rick	Defence Science & Technology Organisation (DSTO)	AU
Vaughan, Dr Jon, NL	Defence Science & Technology Organisation (DSTO)	AU
Villeneuve, Sophie	Canadian Forces Experimentation Center (CFEC)	CA
Wahl, Mike	Joint Forces Command (JFCOM), Experimentation	US
Wheaton, Dr Kendall, NL	Canadian Forces Experimentation Center (CFEC)	CA
Wilmer, Col Mike	US Army Training and Doctrine (TRADOC)	US

National reviewers: AU, Dr Paul Gaertner; CA, Chris McMillan; UK, George Pickburn; and US, Dr Paul J. Hiniker (Completed March 2005).

Copy editor: France Crochetière.

An electronic copy of GUIDEx is available at the following URL: **http://www.dtic.mil/ttcp**

[2] Each GUIDEx contribution followed the English grammar selected by a contributor. This integrated version is in American English and its style is based on the following references [Alley 1987; Alred, Brusaw and Oliu 2003; CBE (Council of Biology Editors) Inc. 1994; University of Chicago Press 2003].

[3] UK National Leader until June 2004, then Dr Geoff Howes joined AG-12 as UK NL.

Table of Contents

Part III Introduction to, Exemplar Findings from, and Précis on GUIDEx Case Studies

List of Figures

List of Tables

Part I
Introduction and Overview of GUIDEx 14 Principles

Introduction

"The grand aim of all science is to cover the greatest number of empirical facts by logical deduction from the smallest number of hypotheses or axioms."
Albert Einstein, see the bibliographic entry under [Einstein 1950].

The Technical Cooperation Program (TTCP) Joint Systems Analysis (JSA) Group of the Non-Atomic Military Research and Development (NAMRAD) created an Action Group (AG) to improve the value which participating nations gain from defense experimentation programs and campaigns. Since the March 2002 inaugural meeting of TTCP JSA AG-12, Methods and Approaches for Warfighting Experimentation, the effort of the AG resulted in the present consolidated guidance supported by national and international Case Studies.

This TTCP JSA AG-12's report, "***Guide for Understanding and Implementing Defense Experimentation***" (GUIDEx), defines the critical components required to initiate, conduct and exploit experimentation programs and campaigns that enable a higher level of information and knowledge sharing among participating countries. GUIDEx establishes a TTCP practitioner road map in conducting defense experimentation programs and campaigns.

The experimentation practices and examples presented in GUIDEx result from the deliberation of the AG-12 participants, who have all had experience in their own countries' defense experimentation efforts. The reader is encouraged to apply and adapt the 14 Principles laid out in this Guide to improve experimentation across the TTCP nations, although they do not express national positions. Many examples within the guide are based on the specific perspective and experience of different lead-nation authors with contributions from other participants: they may require supplementary effort to relate them to national perspectives. It is anticipated that as GUIDEx is used, practitioners will develop additional good practices and examples, and this will stimulate an update to GUIDEx in the future[4].

Scope

GUIDEx is about the use of the experimental method in the defense domain. A number of terms are used by the TTCP nations to describe such activities, including "warfighting experimentation," "defense experimentation" and "military experimentation." GUIDEx has settled on a single term, "**defense experimentation**" in order to present its ideas in a consistent manner. Consequently, "defense experimentation" is defined here as "the application of the experimental method to the solution of complex defense capability development problems, potentially across the full spectrum of conflict types, such as warfighting, peace-enforcement, humanitarian relief and peace-keeping." Most

[4] From a practical viewpoint, a five-year cycle provides the stability required for this guidance to be effective and should be supported by discussions posted on the TTCP Portal.

of the examples available to this guide, however, have been based on warfighting scenarios, simply because of the legacy of the primary focus of defense experimentation to date. In addition, the major focus of GUIDEx is experiments based upon field events and human-in-the-loop simulations, but the Principles of GUIDEx are also applicable to experiments based on analytic wargames and constructive simulations.

The thesis of GUIDEx is that, while it is true that defense experiments are not like highly abstracted and inanimate laboratory experiments, the logic of science and experimentation can be allied to defense experiments to produce credible tests of causal claims for developing effective defense capabilities.

To better achieve these broad objectives in developing effective defense capabilities, GUIDEx presents the idea of ***Integrated Analysis and Experimentation Campaigns***, in which experiments are combined with other analytical techniques; both to tackle larger problems that would not be possible with single experiments, and to exploit the strengths of different techniques. Because the focus of this document is on the critical components required to initiate, conduct and exploit defense experimentation programs and campaigns, references are provided for readers who need detailed information on case studies (*e.g.*, see GUIDEx Case Studies), experimental techniques [ABCA 2004; Alker 1971; Campbell and Stanley 1963; Cook and Campbell 1979; Dagnelie 2003; Rosenbaum 2002], statistical analysis of experimental data [McClave and Dietrich II 1991; Shorack and Wellner 1986; Snedecor and Cochran 1989], and methods for tracing causality in complex situations [Pearl 2001; Shadish *et al.* 2002]

Outline of Report

This report has three Parts besides the front matters and annexes. Part I—introduces readers to the GUIDEx 14 Principles used to structure the rich material of the science of defense experimentation and has been published to be used as a stand alone document (pocketbook). Furthermore, this Part provides an experimentation-planning flowchart that shows what needs to be done in one page. Part II—presents the science of the 14 Principles in full with a précis at the beginning of each. Part III—provides selected Case Studies to illustrate the value to organizations of using these Principles. Annexed material includes; 1- a list of acronyms, initialisms and abbreviations, 2- a lexicon to develop a common language for defense experimentation, 3- bibliographic references, and 4- a subject index.

GUIDEx provides the reader with a perspective of the role and importance of experimentation in a defense capability development process, and this introduction provides the required snapshot of the material covered for the reader to better select the particular material needed at a time.

Introduction

The core of this document is organized along the following 14 Principles for effective experimentation. They are grouped under three dominant topics or themes in the overview as it follows:

Designing Valid Experiments

1. Defense experiments are uniquely suited to investigate the cause-and-effect relationships underlying capability development.
2. Designing effective experiments requires an understanding of the logic of experimentation.
3. Defense experiments should be designed to meet the four validity requirements.

Integrated Analysis and Experimentation Campaigns

4. Defense experiments should be integrated into a coherent campaign of activities to maximize their utility.
5. An iterative process of problem formulation, analysis and experimentation is critical to accumulate knowledge and validity within a campaign.
6. Campaigns should be designed to integrate all three scientific methods of knowledge generation (studies, observations and experiments).
7. Multiple methods are necessary within a campaign in order to accumulate validity across the four requirements.

Considerations for Successful Experimentation

8. Human variability in defense experimentation requires additional experiment design considerations.
9. Defense experiments conducted during collective training and operational test and evaluation require additional experiment design considerations.
10. Appropriate exploitation of modeling and simulation is critical to successful experimentation.
11. An effective experimentation control regime is essential to successful experimentation.
12. A successful experiment depends upon a comprehensive data analysis and collection plan.
13. Defense experiment design must consider relevant ethical, environmental, political, multinational, and security issues.
14. Frequent communication with stakeholders is critical to successful experimentation.

Increasingly, nations such as the United States, Great Britain, Canada, Australia, New Zealand and indeed NATO itself are relying on experimentation to assist in the development of their future military forces. For example, the United States Department of Defense stresses the importance of experimentation as the process that will determine how best to optimize the effectiveness of its joint force to achieve its vision of the future [US Joint Staff 2000]. Is this confidence in the ability of experimentation to support the military transformation process appropriate? Certainly, experimentation has proven itself in the science and technology by producing dramatic advances. Can the methods of experimentation, which have so expeditiously and radically developed science and technology, be applied to the military transformation process to achieve similar advances in military effectiveness?

Introduction

The thesis of this guide is that robust experimentation methods from the sciences can be adapted and applied to military experimentation and will provide the basis for advancements in military effectiveness in the transformation process. The authors have structured the relevant experimentation material under 14 Principles, which ensure that defense experimentation programs positively impact coalition organizations' ability to evolve force capabilities of the future. Also, they have provided an experimentation-planning flowchart that in one page shows what needs to be done, together with a set of Case Studies that demonstrate the value of the principles in practice.

GUIDEx is not meant to duplicate information already available in other documents and textbooks on experimentation such as those referenced here, [ABCA 2004; Alberts and Hayes 2002, 2005; Dagnelie 2003; Radder 2003; Shadish *et al.* 2002] or on command and control (C2) assessment [NATO 2002], but organizes and expands this detailed information under 14 Principles to guide successful defense experimentation.

Overview

Experiments and Science

In about 400 B.C., philosophers Socrates and Plato investigated the meaning of knowledge and methods to obtain it using a *rational-deductive* process, or pure logic (logic), without reference to the real world. Aristotle was a transitional figure who advocated observation and classification, bridging to later scientists like Ptolemy and Copernicus who developed *empirical-inductive* methods that focused on precise observations and explanation of the stars. These early scientists were not experimenters. It is only when later scientists began to investigate earthly objects rather than the heavens, that they uncovered a new paradigm for increasing knowledge.

In the early 1600s, Francis Bacon introduced the term experiment and Galileo moved from astronomical observations to conducting earthly experiments by rolling balls down an inclined plane to describe bodies in motion. The realization that manipulating objects would yield knowledge spawned a new research paradigm, one unimagined in the previous 2000 years of exploring the out-of-reach heavens. The basis of this new science paradigm called experimentation (the *empirical-deductive* approach) was a simple question [Feynman 1999]: "If I do this, what will happen?" The key to understanding experimentation, and the characteristic that separates experimentation from all other research methods, is manipulating something to see what happens. The scientific aspect of experimentation is the manipulation of objects under controlled conditions while taking precise measurements. In its simplest form [Shadish *et al.* 2002: p. 507], an experiment can be defined as a process "*to explore the effects of manipulating a variable*."

Designing Valid Experiments

Principle 1. Defense experiments are uniquely suited to investigate the cause-and-effect relationships underlying capability development.

Principle 2. Designing effective experiments requires an understanding of the logic of experimentation.

Principle 3. Defense experiments should be designed to meet the four validity requirements.

Improved capabilities cause improved future warfighting effectiveness. Experimentation is the unique scientific method used to establish the cause-and-effect relationship of hypothesized capabilities. If experimenters design the five experiment components to meet the four experiment validity requirements, defined later, the defense experiment will provide the scientific evidence to proceed. Defense experiments are essential to develop empirical- and concept-based capabilities that yield implementable prototypes. The use of a "**develop–experiment–refine**" approach ensures that a rigorous methodology relates new capabilities to warfighting effectiveness. The development and delivery of defense concepts and capabilities is thus supported through experimentation.

Experiment Hypotheses

To understand cause-and-effect relationships between capabilities and increased warfighting effectiveness is to understand experiment hypotheses. Any national or coalition capability problem may be stated as: **Does A cause B?** An experimental capability or concept—a new way of doing business—is examined in experimentation to determine if the proposed capability **A** causes the anticipated military effect **B**. The *experiment hypothesis* states the causal relationship between the proposed solution and the problem.

It is an "*If…then…"* statement, with the proposed cause—innovative concept—identified by the *if* clause, and the possible outcome—the problem resolution—identified by the *then* clause.

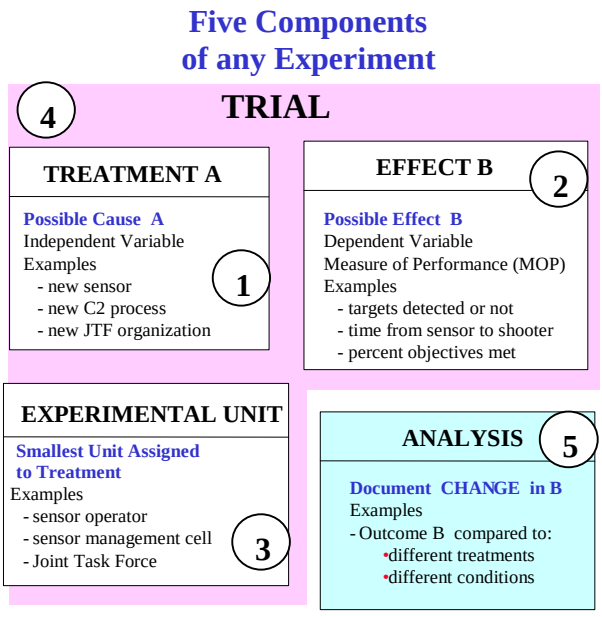
Hypothesis

If… **"proposed change"**
Then… **"improved warfighting capability"**

Components of an Experiment

All experiments—large or small, field or laboratory, military or academic, applied or pure—consist of five components[5] [Shadish *et al.* 2002: p. 2]:

1. The *treatment*, the possible cause **A**, is a capability or condition that may influence warfighting effectiveness.
2. The *effect* **B** of the treatment is the result of the trial, a potential increase or decrease in some measure of warfighting effectiveness.
3. The *experimental unit*[6] executes the possible cause and produces an effect.
4. The *trial* is one observation of the experimental unit under treatment **A** or under the alternative **~A** to see if effect **B** occurred, and includes all of the contextual conditions of the experiment.
5. The *analysis* phase of the experiment compares the results of one trial to those of another.



These five components are useful in understanding all defense experiments including large field experiments. Some field experiments are grand exercises with multiple

[5] For application of these concepts to test and evaluation, see [Kass 1997].
[6] An experimental unit includes all operators with their gear, procedures, and concept of operations. In experimentation, the apparatus includes the experimental unit and necessary conditions for effecting changes and observing effects.

experimental initiatives (possible causes), sometimes as many as 20 to 30 different initiatives in one experiment. To be useful, each individual experimental initiative should be configurable as a unique mini-experiment with its own subset of the five components. Each initiative is a particular treatment with its own experimental unit (operators in one area of the task force), its own set of outcome measures, and its own set of trial conditions. However, in practice it is very difficult to maintain independence among these many experiments within the large exercise, which makes it difficult to isolate specific causal influences.

What Is a Good Experiment?

A good, or valid, experiment provides information to ascertain whether **A** caused **B** [Shadish *et al.* 2002: p. 3]. **Four logically sequenced requirements** are necessary to achieve a valid experiment.[7] A simple experiment example will illustrate these four requirements. A proposed concept postulates that new sensor capabilities are required to detect future targets. An experiment to examine this proposition might employ *current sensors* on the first day of a two-day experiment and a *new sensor capability* on the second day. The primary measure of effectiveness is the number of targets detected. The experiment hypothesis could be: "If new sensors are employed, then target detections will increase."

1 Ability to use the new capability A

Developing and generating the new experimental capability for the experiment is often a major resource commitment. In an ideal experiment, operators employ the experimental capability, in this case the new sensors, to its optimal potential; thereby allowing the new capability to succeed or not succeed on its own merits. Unfortunately, this ideal is rarely achieved. A lesson repeatedly learned from defense experiments is that new experimental capabilities are frequently not fully realized in the experiment.

A number of things can go wrong with an experimental surrogate. For example, the hardware or software does not work as advertised or anticipated. The experiment players may be undertrained and not fully familiar with its functionality. Because the experimental treatment represents a new capability, the trial scenario and potential outcomes may not be sensitive to the new capability's enhanced performance.

A valid experiment design ensures that the new capability works under relevant conditions prior to execution, that the operators are adequately trained to employ it appropriately, and that the scenario is sufficiently sensitive to determine the capability's effectiveness. Experimenters continually monitor these aspects during experiment execution. If the experimental sensors **A** do not function during the experiment, the new capability will most likely not affect the military unit's ability to detect targets **B**, which is the next experiment validity requirement.

[7] Many detailed good practices developed by experiment agencies through experience (and described in recent books such as [Alberts and Hayes 2002, 2005]) can be organized under these four requirements and the 14 Principles.

2 Ability to detect a change in the effect B

When the player unit correctly employs a new capability, does it result in any noticeable difference in the effect **B** during the experiment trial? Ideally, a change in the number of detections accompanies a transition from old to new sensors. If this is not the case, this may be because there is too much experimental noise[8]—the ability to detect change is a signal-to-noise ratio problem. Too much experimental error produces too much variability, hampering detection of a change. Reduction of experiment variation, through data collection calibration, limited stimuli presentations, and a controlled external environment, mitigates experiment-induced error. In addition, since the computation of variability in statistics decreases as the number of repetitions increases, a larger sample size increases the signal-to-noise ratio making it easier to detect change.

Analysts measure change in effectiveness by comparing the results of one experiment trial to those of another. Typically, different experiment trials represent different levels of applications of the same capability, alternative competing capabilities, or the same capability under different conditions. A change in military effectiveness may also be detected by comparing the results of an experiment trial to a pre-existing baseline, a task standard, or a desired process.

3 Ability to isolate the reason for change in the effect B

If an experimenter employed a useable capability that produced a noticeable increase in the number of target detections, was the observed change in detections due to the intended cause—changing from old sensors to new—or due to something else? In the sensor-experiment example, an alternative explanation for the increase in detections on the second day could be that of a learning effect. That is, the sensor operators may have been more adept at finding targets because of their experience with target presentations on Day One and, consequently, would have increased target detections on Day Two, whether or not different sensors were employed. An increase in operator experience coincidental with a change in sensors would dramatically alter the interpretation of the detected change in effectiveness. An experiment outcome with alternative explanations is a *confounded* result. Scientists have developed experimentation techniques to eliminate alternative explanations of the cause of change: counterbalancing the presentation of stimuli to the experimental unit, the use of placebos, the use of a control group, random assignment of participants to treatment groups, and elimination or control of external influences.

4 Ability to relate the results to actual operations

If the player unit ably employed the capability, and if an experimenter detected change and correctly isolated its cause, are the experiment results applicable to the operational forces in actual military operations? The ability to apply, or *generalize*, results beyond the experiment context pertains to experiment realism and robustness. Experiment design issues that support operational realism revolve around the representation of surrogate systems, the use of operational forces as the experimental unit, and the use

[8] Experimental noise interferes with the observation of the desired variable at a required degree of precision.

of operational scenarios with a realistic reactive threat. To ensure the operational robustness, the experiment should examine multiple levels of threat capabilities under various operational conditions.

Experiments during Capability Development and Prototyping

Nations employ a variety of processes to support development of improved empirical- and concept-based capabilities and are, increasingly, employing defense experimentation to support the delivery of this improved warfighting effectiveness. These capability development and prototyping processes are not the same across the different nations (in some nations these processes are referred to as **concept development and experimentation**, **CD&E**). However, in most cases they follow similar **develop–experiment–refine** stages. For the purposes of GUIDEx, therefore, a generic description of these stages is presented with the hope that the ideals embodied can be mapped onto each nation's own way of doing business.

Stage	**Aim**
Discovery	To clarify future warfighting problems and to seek potential solutions.
Refinement	To examine and refine the extent to which proposed capabilities or concepts solve military problems.
Assessment	To ensure that solutions from *refinement* are robust; that they are applicable to a wide range of potential operational requirements in an uncertain future.
Prototype Refinement	To transition capability surrogates into potential operational capabilities by developing complete prototype packages for front line commands.
Prototype Validation	To provide the final demonstrated evidence that the prototype capability can operate within theater and will improve operational effectiveness.

Experiments are required throughout a capability development and prototyping process. They provide an empirical method to explore new capabilities, to refine concepts, and to validate new prototypes for implementation. For example, during *refinement*, experiments quantify the extent to which proposed capabilities solve military problems. Experiments also examine capability redundancies and tradeoffs and reveal capability gaps. Prior *discovery* stage activities only speculate whether proposed further capabilities would solve identified gaps in military effectiveness, whereas experimentation during *refinement* empirically substantiates and quantifies the extent proposed capabilities increase effectiveness in specific case examples. In some instances, experimentation may suggest prototypes for early implementation, or identify areas needing future investigation. Experiments during *assessment*, on the other hand, investigate the robustness of the solution developed during *refinement* for possible future military operations. These experiments examine different future contingencies, different multinational environments, and different threat scenarios to ensure that the *refinement* stage solution is robust; that it is applicable to a wide range of potential operational requirements in an uncertain future.

Prototypes derived from the earlier stages are often not ready for immediate operational use. Experiments during *prototype refinement* can transition concept prototypes into potential operational capabilities by developing complete prototype packages for front line commands. These experiments develop the detailed tactics, techniques, procedures (TTPs), and organizational structures for the prototype as well as developing the tasks, conditions, and standards to facilitate training. They can also examine the latest hardware and software solutions and their interoperability with existing fielded systems. Experiments during *prototype validation* provide the final demonstrated evidence to the combatant commander that the prototype capability can operate within theater and will improve operations. Often these experiments are embedded within exercises or other training events and are used to validate the predicted gains in effectiveness of the force.

Integrated Analysis and Experimentation Campaigns

Principle 4. Defense experiments should be integrated into a coherent campaign of activities to maximize their utility.

Principle 5. An iterative process of problem formulation, analysis and experimentation is critical to accumulate knowledge and validity within a campaign.

Principle 6. Campaigns should be designed to integrate all three scientific methods of knowledge generation (studies, observations and experiments).

Principle 7. Multiple methods are necessary within a campaign in order to accumulate validity across the four requirements.

Experimentation is a necessary tool in addressing large capability development problems, but this should be embedded in an integrated campaign of experiments, studies and analytical activities. Such Integrated Analysis and Experimentation Campaigns would typically also have an integrated analytical and management process, and use a variety of techniques to ensure that weaknesses in one technique can be mitigated by others.

Campaigns use a mix of defense experiments and parallel studies to understand the problem's context, the associated warfighting concept and the capabilities required. The product of the campaign is advice to decisionmakers on the utility, versatility and maturity of the concept and the capabilities required to implement the concept. Campaigns can address issues at all levels from joint and combined operations to platforms and components.

An integrated campaign using a variety of techniques ensures that weaknesses in one technique can be mitigated by others. Where results (*e.g.*, inferences) correlate between activities, it increases confidence and where they diverge, it provides guidance for further investigation. It is only when all activities are brought together in a coherent manner and the insights synthesized, that the overall problem under investigation is advanced as a whole.

Such campaigns can address force development issues at any level, for example: technological (*e.g.*, systems of systems), tactical, operational, as well as strategic. Instances of activities at each of these levels in Australia, for example, are as follows:

- at the **technological** level: helicopter operations within a combined arms team, surface and sub-surface platforms for maritime operations, and the JSF within the air control system;
- at the **tactical** level: amphibious and airmobile task groups;
- at the **operational** level: the capability balance required to achieve the Future Warfighting Concept; and finally,
- at the **strategic** level: the Effects Based Operations concept is being developed in conjunction with many government agencies.

Why use a Campaign

An integrated analysis and experimentation campaign will be required for a variety of reasons. There may be resource or political reasons why a campaign is preferred to a single activity, or more likely it will be necessary because without a coordinated campaign, the problem or issue under investigation simply cannot be satisfactorily resolved. A campaign allows the problem to be approached in a coordinated, manageable manner with a variety of analytical techniques and allows a degree of iteration and synthesis between activities that help ensure that the overall problem is sufficiently addressed. The problem may initially be ill-defined and a campaign of activities will allow assessment and adjustment as the problem is refined. Some of the analytical reasons for using a campaign approach are described in the following sub-sections.

- **Problem Characteristics.** Military capability development problems are generally complex and coercive. The socio-technical nature of the system and the interaction between the components and the environment characterize the system as *complex*. The importance of an opposing force, itself a socio-technical system, means the system is *coercive*. Many problems that might be explored through defense experimentation are simply too complex to be dealt with in a single activity.
- **Increased Confidence.** An integrated campaign of experiments and other activities allows a gradual build-up of the knowledge surrounding the problem or issue under investigation, leading to a more refined and robust concept. This increases confidence that the findings are valid and creates a systematic body of knowledge to inform and investigate capability development.
- **Synthesis of Military and Analytical Skills.** A campaign, by integrating different techniques, provides improved opportunity for analytical and military skills to be applied to the problem.

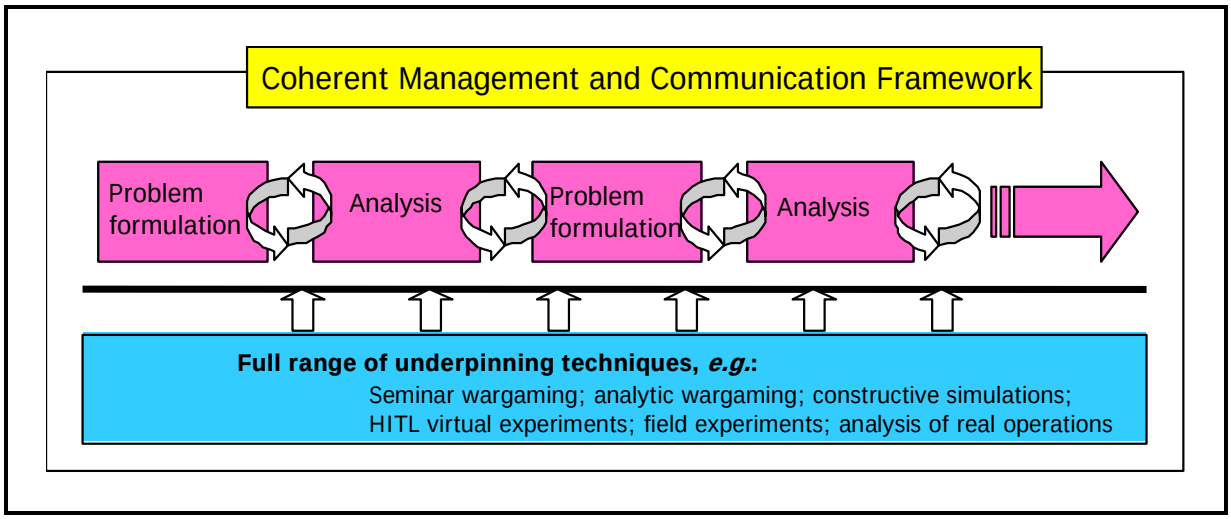
- **Problem Formulation.** When the strategic environment is uncertain and unprecedented, and the impact of technology unknown, the experience base is usually too narrow to conduct the problem formulation confidently. Within the campaign we must therefore build a **synthetic experience base** and the process of scientific inquiry is used to increase our confidence in the problem formulation.

Iterating Methods and Experiments

The initial stage of any campaign is problem formulation. Effective problem formulation is fundamental to the success of all analyses, but particularly at the campaign level because the problems are normally ill-defined, complex and adversarial, involving many dimensions and a rich context. Problem formulation involves decomposition of the military and analytical aspects of the problem into appropriate dimensions. Decomposition cannot normally be achieved without detailed analysis using a matrix of tools such as seminars and defense experiments supported by analytical studies and operational experience. Detailed analysis also assists in the reconstruction of the problem segments and interpretation of results.

In dealing with fuzzy or uncertain interactions, the problem formulation process needs to explore and understand the significance of each interaction before making (or seeking from customers) assumptions about it. This involves keeping an open mind, during the early stages of problem formulation, about where the boundaries lie and their dimensional nature. This is difficult because it makes the process of modeling the problem more complicated. A call for hard specification too early in that process must be avoided. In the end, of course, the problem must be formulated in order to solve it, but formulation should be an output from the first full iteration, not an early input to it.

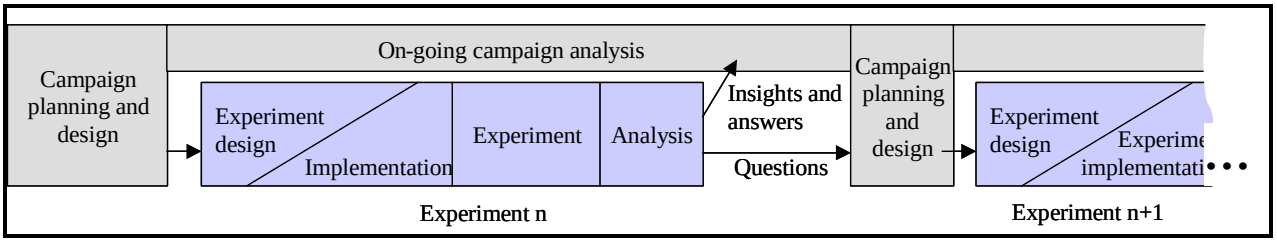
As shown in the following illustration, the problem is being formulated and refined throughout the entire campaign in an iterative cycle that never really completes until the campaign itself completes. The process of problem formulation and analysis is undergoing constant review to reshape the direction of the campaign and to ensure that the real issue or concept is being addressed.



Wargames, and in particular seminar wargames, have an important role in problem formulation. In wargaming it is possible to balance the physical and psychological aspects of the problem by using warfighters as the players while adjudicating their actions using models or rulesets. Most importantly, wargaming introduces an adversary early in the problem formulation process, providing a stressful environment in which to explore the concept and develop the hypotheses for subsequent analysis. Although human-in-the-loop simulations and live simulations also introduce a human adversary, they are frequently too expensive and unwieldy for the problem formulation phase.

Integration of Scientific Methods

The aim of a campaign is to integrate a range of methods: experiments (observations with manipulation—empirical-deductive); observational studies (observations without manipulation—empirical-inductive) and analytical studies (rational-deductive) into a coherent package that addresses a complex capability development problem. The phases of campaign design are the same as for any evaluation, that is, problem formulation and analysis. The complexity arises because after the completion of each activity the problem formulation is reassessed and adjusted and subsequent activities may be redesigned. As a result a campaign plan is a flexible instrument, with a supporting risk-management framework and an iterative approach to constantly review and reshape the remainder of the campaign to ensure that the overall goals are achieved.



In all likelihood, seminars, workshops, historical analysis, and the like, will also be required as part of the campaign to support and help inform the experimenters who will ultimately address the overall question. The campaign plan process must take these other activities into account within its design phase. The ultimate aim is to synthesize the outputs from all activities into coherent advice to the decisionmakers.

Different Methods Offer Different Strengths

All experiments must strike a balance among the four experiment validity requirements. Attempts to satisfy one work against satisfying the other three. Consequently, 100 percent-valid experiments are unachievable. Precision and control increase the ability to detect change and to isolate its cause, but decrease the ability to apply the results to imprecise, real-world situations. Experiments designed to identify change emphasize strict control of trial conditions and feature multiple repetitions of similar events; experiments designed to relate results emphasize free-play, uncertainty, and reactive threats. Each individual experiment design must consider requirement tradeoffs in order to minimize the loss of one requirement due to the priority of another.

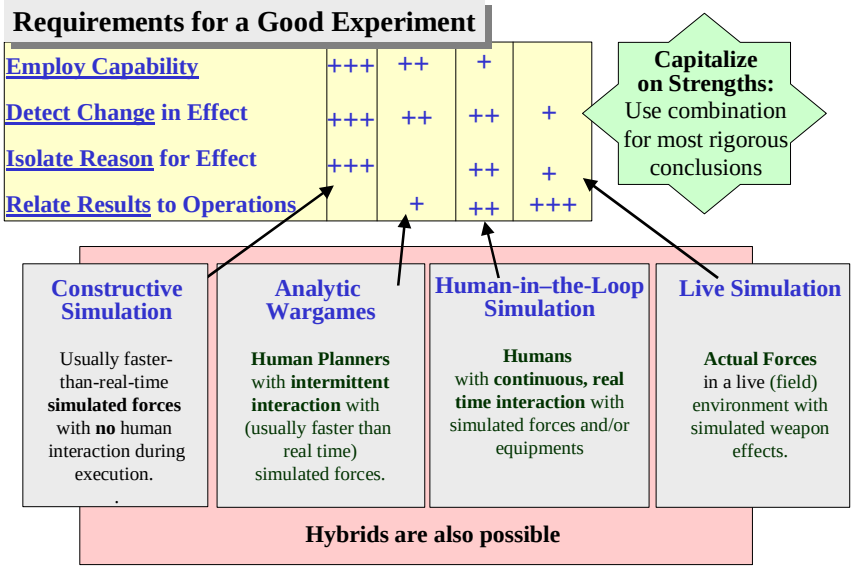
Most defense experiments use some form of simulation, which can be grouped into one of four general methods: *constructive simulation, analytic wargames, human-in the-loop simulation, and live (field) simulation*. Each of these four methods has its own strengths and weaknesses with respect to the four experiment validity requirements discussed previously. Since one particular method cannot satisfy all four requirements, an integrated analysis and experiment campaign requires multiple methods.

Constructive simulations are those in which no human intervention occurs in the play after designers choose the initial parameters and then start and finish the simulation. Constructive simulations are a mainstay of military analytical agencies. They allow repeated replay of the same battle under identical conditions, while systematically varying parameters—the insertion of a new weapon or sensor characteristic, the employment of a different resource or tactic, or the encounter of a different threat. Experiments using constructive simulations with multiple runs are ideal to detect change and to isolate its cause. Because modeling complex events requires many assumptions, including those of variable human behavior, critics often question the applicability of constructive simulation results to operational situations.

Analytic wargames 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 a simulation to help determine the outcome. A typical analytic wargame might involve fighting the same campaign twice, using different capabilities each time. The strength of such wargames for experimentation 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, analytic wargames 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.

Rigorous Experimentation Requires Multiple Methods to Meet the Four Validity Requirements



Human-in-the-loop simulations represent a broad category of real-time simulations with which humans can interact. In a human-in-the-loop defense experiment, military subjects receive real-time inputs from the simulation; make real-time decisions, and direct simulated forces or platforms against simulated threat forces. The use of actual military operators and staffs allows this type of experiment to reflect warfighting decisionmaking better than experiments using purely constructive simulation. However, when humans make decisions, variability increases, and changes are more difficult to detect and consequently to attribute to the cause.

Live simulation is conducted in the actual environment, with actual military units and equipment and with operational prototypes. Usually only weapon effects are actually simulated. As such, the results of experiments in these environments, often referred to as field experiments, are highly applicable to real situations. Good field experiments, like good military exercises, are the closest thing to real military operations. A dominant consideration however, is the difficulty in isolating the true cause of any detected change since field experiments include much of the uncertainty, variability, and challenges of actual operations; but they are seldom replicated due to costs.

Different Methods during Capability Development and Prototyping

As potential capabilities advance through capability development and prototyping stages, the following considerations are useful in selecting which of the four experiment validity requirements to emphasize. For example, finding an initial set of potential capabilities that empirically show promise is most important in the *refinement* stage. Experiments in this early stage examine idealized capabilities (future capabilities with

projected characteristics) to determine if they lead to increased effectiveness, and are dependent on the simulation-supported experiment, using techniques such as constructive simulation, analytic wargames and human-in-the-loop simulation. Accurately isolating the reason for change is not critical at that stage, as the purpose is only to apply a coarse filter to the set of idealized capabilities. However, during the *assessment* stage, quantifying operational improvements and correctly identifying the responsible capabilities is paramount in providing evidence for concept acceptance. This is also dependent on experiments with better-defined capabilities across multiple realistic environments. Experiments conducted using constructive simulation can provide statistical defensible evidence of improvements across a wide range of conditions. Human-in-the-loop and field experiments with realistic prototypes in realistic operational environment can provide early evidence for capability usability and relevance. Early incorporation of the human decisionmaker in this way is essential, as the human operators tend to find new ways to solve problems.

In *prototype refinement* experiments, one should anticipate large effects, otherwise its implementation might not be cost effective. Accordingly, the experiment can focus on the usability of working prototypes in a realistic experiment environment. Isolating the real cause of change is still critical when improving prototypes. The experiment must be able to isolate the contributions of training, user characteristics, scenario, software, and operational procedures. As previously described, human-in-the-loop and field experiments provide the opportunity for human decisionmakers to influence development. In *prototype validation*, human decisionmakers ensure that the new technology can be employed effectively. Prototype validation experiments are often embedded within joint exercises and operations.

Employing Multiple Methods to Increase Rigor

Since experiments using the four main simulation methods emphasize the four validity requirements differently, an integrated analysis and experimentation campaign must capitalize on the strengths of each method to accumulate validity. For example, the model–exercise–model paradigm integrates the strengths of, on the one hand, the constructive simulation (*i.e.*, “model”) and, on the other, any of the methods that involve human interaction (*i.e.*, “exercise” in a generic sense). This technique is especially useful when resource constraints prohibit conducting side-by-side baseline and alternative comparisons during wargames and field experiments.

In the model-exercise-model paradigm, the early experiments using constructive simulation examine multiple, alternative, Blue-force capability configurations and baselines. Analysis of this pre-exercise simulation allows experimenters to determine the most beneficial Blue-force configuration for different Red-force scenarios. An analytic wargame, human-in-the-loop or field experiment can then be designed and conducted, which provides independent and reactive Blue- and Red-force decisionmakers and operators. One can then re-examine this optimal configuration and scenario.

Experimenters use the results of the exercise to calibrate the original constructive simulation for further post-event simulation analysis. Calibration involves the adjustment of the simulation inputs and parameters to match the simulation results to those of the experiment, thus adding credibility to the simulation. Correspondingly, rerunning the pre-exercise alternatives in the calibrated model provides a more credible interpretation of any new differences observed in the simulation. Additionally, the post-exercise calibrated simulation improves analysts' ability to understand fully the implications of the experiment results by conducting "what if" sensitivity simulation runs. Experimenters examine what might have occurred if the Red or Blue forces had made different decisions during the experiment.

The model–exercise–model method increases overall experiment validity by combining the contrasting strengths of the following methods:

1. experiments using constructive simulation, which is strong in detecting differences among alternative treatments, and
2. experiments using either human-in-the-loop simulation, analytic wargame, or field experiments, which are stronger in incorporating human decisions that better reflect the actual operating environment.

This paradigm also helps to optimize operational resources by focusing the exercise event on the most critical scenario for useful results, and by maximizing the understanding of the event results through post-event sensitivity analysis.

Considerations for Successful Experimentation

Principle 8. Human variability in defense experimentation requires additional experiment design considerations.

Principle 9. Defense experiments conducted during collective training and operational test and evaluation require additional experiment design considerations.

Principle 10. Appropriate exploitation of modeling and simulation is critical to successful experimentation.

Principle 11. An effective experimentation control regime is essential to successful experimentation.

Principle 12. A successful experiment depends upon a comprehensive data analysis and collection plan.

Principle 13. Defense experiment design must consider relevant ethical, environmental, political, multinational, and security issues.

Principle 14. Frequent communication with stakeholders is critical to successful experimentation.

This guide identifies a number of considerations that are intended to support the practical implementation of experiments. These considerations relate to the need to recognize and accommodate the human element in experiment design, and they also provide advice on how to make the best use of operational test and evaluation events or training exercises. They also give guidance on some issues relating to modeling and simulation, on the implementation of good experiment control and highlight national regulations, security rules and practices that may need special consideration; and finally, there are also some practical steps that can be taken to achieve good communications.

Human Variability

The implications arising from using human subjects in defense experimentation are often overlooked. Most, if not all defense experiments examine impacts on socio-technical systems but experiment designs rarely cater sufficiently for the human element. Because humans are unique, highly variable and adaptable in their response to an experimental challenge, they are more than likely to introduce a large experimental variability. In addition, humans will have different experiential baselines in terms of, for example training and aptitude and, unlike technology, will become tired and possibly demotivated. They may also learn during experiments. The experiment design and the data analysis and collection plan must recognize and accommodate human variability, which will be much larger than would be predicted if the socio-technical system were treated solely as technology. What is sometimes overlooked is that this variability provides important information on why a socio-technical system responds to a challenge in a particular way. Indeed there is an argument that human variability should not be minimized, as this would lose important information. High variability may indicate a fault in the system under examination, or in the experiment design. An understanding of the impact of human variability on experiment design and outcome is a fundamental skill required by all experimenters.

Regardless of the experimenter's ability to control human variability, it is important, if possible, to measure it. This is done mainly to see if detected effects can be explained in terms of human variability rather than the experimental treatments. For example, where a single group is the subject for all the treatments, then learning by that group during and between the treatments may have a confounding effect on the whole experiment. It may be possible to measure learning effects within each treatment, and thus estimate any confounding effect of learning between treatments. Of course, this may increase the complexity of the experiment design as the data analysis will then also need to control for human variability measures and assess their impact upon the main variables.

Although objective measures of variables are favored by experimenters, subjective measures are important for ascertaining the mental processes underlying observed behaviors. This information may be important, especially if a subject adapts to using a capability in a way not considered by the experimenter. Asking subjects why they have changed their behavior can enhance understanding of maladaptive ways of using of a new capability. Consideration needs to be given to the timing of subjective interviews, particularly whether they should take place soon after the action occurs or at the end of the experiment. The former may be obtrusive to the subjects and may impact the results, with the latter being affected by factors such as memory decay and motivation.

Exploiting Operational Test and Evaluation and Collective Training Events

Opportunities to conduct experimentation may be found in training exercises and in Operational Test and Evaluation (OT&E) events. Operational assessments, in particular, provide an opportunity for conducting experimentation with substantial technological and expert staff support. The drive to conduct experimentation activities during training exercises and OT&E events is almost entirely due to the difficulty of acquiring the resources (equipment, estate, human) to undertake defense experiments of any significant size. Arguably, the equipment programs that require most support from experimentation are those intended to enhance *collective* rather than team or individual effectiveness, and thus collective groups of personnel (which may comprise command teams with higher and lower controllers) are required to undertake that experimentation. It is a simple fact of life in the early 21st Century that most nations generally do not have units and formations available to dedicate to experimentation, except for the most limited-scale activities. Therefore exploiting routine training exercises and other collective events should be given serious consideration.

Exploiting collective training (exercises) has a range of benefits as well as disadvantages and a variety of factors must be taken into account in both planning and execution. The principal one is that training always has primacy and the experimenter has little control over events, thus the skill is in understanding the constraints that the exercise opportunity will present and knowing how to work within them. Exploiting training exercises for the purposes of experimentation is most achievable during the prototype validation phase of a capability development program when functional prototypes exist.

The potential to include experimentation within OT&E programs is very high. This is so in part because many of the components of OT&E events are the same as their counterparts in experiments. They are well supported by the technical/engineering community and valued by the operational community as a component of the operational readiness process. The operational community will therefore generally be engaged in OT&E events and the potential to include experiments in these events as well can be very good. An important benefit to experimenters is the OT&E infrastructure, which includes engineering/technical staffs and facilities; planning support; test support during execution and evaluation support for the after-action review or report (AAR). The benefit from the use of OT&E staffs and facilities is realized because of the strong overlap between the two processes. An important benefit to the OT&E community is that the prototypes from experiments may soon be operational systems. In such circumstances, there is a significant advantage to be obtained by the inclusion of OT&E staffs in experimentation on these systems.

Although training exercises and OT&E events do not allow execution of elaborate experiment designs because it would impede training and impact operational readiness, scientific methodology and the four experiment validity requirements can be applied to such embedded experiments. Experimentation in these situations naturally provides the

strongest venue for meeting the fourth experiment validity requirement, *i.e.*, the ability to relate results to actual operations. While operational necessity restricts the ability to meet the first three experiment validity requirements in training exercises, and to a lesser extent in OT&E events, the experimenter can ameliorate the limitations to some degree. With respect to the first experiment validity requirement, *i.e.*, the ability to use the new capability, prototype testing prior to the training exercise enhances the usability of the experimental capability and should ensure that it will function correctly during the exercise trials. This is less of an issue for OT&E, as this activity is generally for validating the performance of new operational systems and the testing is implicit. Additionally, to address the second experiment validity requirement in training exercises, *i.e.*, the ability to detect a change in the effect, establishing a pre-exercise definition of expected performance and comparing the prototype's actual performance during the exercise to its expected performance provides the necessary ability to detect change. For OT&E, the performance of new operational systems is typically documented in manuals and validated computer models may exist. Therefore, the baseline system performance should be well established and the potential for detecting change should be good.

While the ability to isolate the reason for the observed change effect, *i.e.*, the third experiment validity requirement, is the most problematic in experimentation embedded in training exercises, experimenters can nevertheless achieve some level of satisfaction here as well. When examining different capabilities during a single exercise, the experimenter should conduct different prototype trials at different times so the effects of one prototype do not influence the effects of the other. It is prudent to have an experienced exercise "observer-controller" view the prototype trial to assess the extent that any observed results were the results of the experimental capability instead of unintended causes. Additionally, showing that the rigorous experiment data accumulated during the concept development phase of the prototype is still relevant to the exercise conditions also supports GUIDEx third experiment validity requirement. Experimentation embedded in OT&E events also creates considerable challenges for meeting the third experiment validity requirement. The best approach in this case is through comprehensive, detailed data collection, which is typically the case in OT&E events anyway.

Finally, for both the use of training exercises and OT&E events, a Model-Exercise-Model paradigm that was successfully calibrated to the event results would allow follow-on sensitivity analysis to demonstrate that inclusion and exclusion of the experimental capability accounted for decisive simulation differences.

Training Exercises	OT&E Events
Benefits	
♦ Availability of experimental subjects in large numbers ♦ High level of engagement of experimental subjects ♦ Use of training infrastructure ♦ Moderate sample sizes, for repeated exercise series ♦ Ability to use repeated exercises as a control group, or baseline ♦ They rate highly in terms of relating any detected change to real operations.	♦ Availability of operational staff and platforms ♦ High level of engagement of technical community ♦ Use of OT&E infrastructure ♦ Moderate sample sizes, for repeated test series ♦ Ability to use repeated tests as a control group, or baseline ♦ Strong potential for relating any detected change to real operations.
Constraints	
♦ Exercises are designed to stimulate various training points that may not satisfy an experiment design ♦ Training has primacy—can a genuine experiment design be fitted around training? ♦ Scenarios and settings designed for training purposes ♦ Limited opportunities to make intrusive changes to the exercise or collected data intrusively ♦ Can results be published without breaching the anonymity of the training audience? ♦ Interventions by Exercise Control for training reasons, *e.g.*, the training force is winning too easily ♦ Exploitation of an exercise too early in a unit's training cycle can yield poor results, *e.g.*, the collective skills may be too low	♦ OT&E events are designed to quantify aspects of equipment performance or to determine if a standard is being met that may not satisfy an experiment design ♦ OT&E has priority and the experiment may not interfere with test objectives ♦ Scenarios and settings designed for OT&E purposes ♦ Limited opportunities to make intrusive changes to the test or collected data intrusively ♦ Can results be published without breaching the anonymity of the test audience?

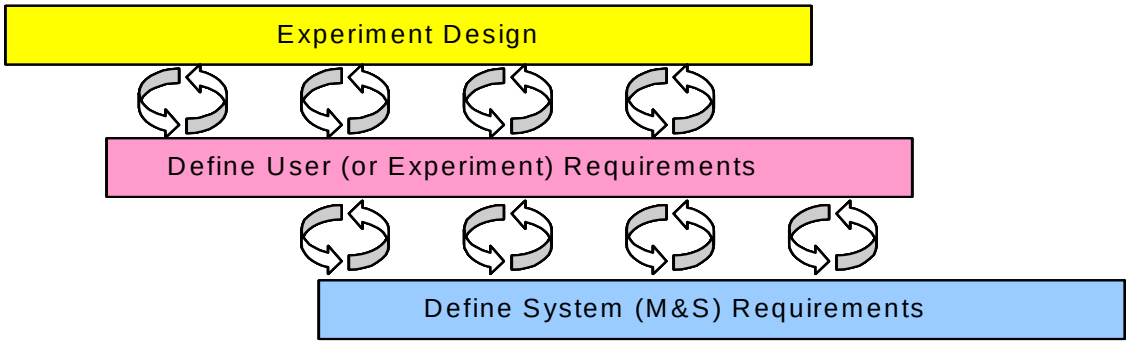
Modeling and simulation Considerations

This guide presents modeling and simulation (M&S) as intrinsic to conducting most defense experiments. There is now a wide range of M&S techniques available and this makes the innovative use of M&S cost effective for many defense experimentation applications. However, there are some significant issues associated with selecting both the types of M&S to be used and the specific elements of the experiment federation.

A balanced view of fidelity and validity. For many years, as rapidly increasing computing power led to many new modeling possibilities, there was a generally held view that greater fidelity, or accuracy, was always better. Indeed, many took the term "validity" to be almost synonymous with fidelity and detail. The modern view is that validity actually means "fit for purpose," with the *purpose* being to execute the desired

experiment design. This means that we should consider the main measure of merit for M&S to be *adequacy*, not *fidelity*. The experiment design should effectively define what level of fidelity is adequate. Furthermore, the main point of modeling is to rationalize the complexity of real life by simplifying it. In "The Lanchester[9] Legacy" [Bowen and McNaught 1996: Vol. III, Ch. 9], the authors wrote: "It has long been understood by Operational Analysts that, in dealing with complicated situations, simple models that provide useful insights are often to be preferred to models that get so close to the real world that the mysteries they intend to unravel are repeated in the model and remain mysteries." We can therefore imply an axiom that M&S should be as simple as possible while remaining adequate for the task in hand.

M&S definition. It is a key principle that the definition of the M&S to be used in an experiment should be derived from the experiment design, and not the other way around. However, rarely will practitioners have the luxury of completing their experiment design and then moving through a user requirements and subsequently system requirements definition process in sequence. Usually a concurrent process is necessary, with the processes beginning in the order given above. A spiral development process can then take place. There are several well-established processes for achieving this, *e.g.*, the US Federation Development and Execution Process (FEDEP) and the European Synthetic Environment Development and Exploitation Process (SEDEP).



Modeling the process to be examined by the experiment. Experiments and observational studies (where a concept is subjected to objective observation, but without manipulation) are intrinsically connected to the idea of hypotheses. The hypothesis is simply a plausible proposition about either causal or associative relationships. Thus in a general sense there is always implicitly a model of the process being experimented with by virtue of there being one or more hypotheses. However, it is possible, and in most cases desirable, to model the process in advance in a much more tangible way, regardless of whether a strict **model-exercise-model** paradigm is being followed. In particular, architectural frameworks such as Zachman [Zachman 1987] and DoDAF[10] represent an excellent and increasingly popular means to describe military problems and potential candidate solutions in a variety of different ways. When

[9] F.W.Lanchester was one of the pioneers of military operational research.
[10] DoD Architecture Framework, see [DoDAF Working Group 2004]

a model-exercise-model paradigm is being followed, process models based on these frameworks can often be preferable to complex constructive combat simulations.

Experiment Control

Experimentation is intrinsically a controlled activity, although the degree of possible and required control varies from case to case. The experiment design should be explicit in describing which variables must be controlled in order to prevent rival explanations for the findings, and which variables can be allowed to remain uncontrolled though usually recorded. It should also describe the control regimes to be put in place to ensure that this occurs in practice. The identification of intervening variables and learning effects must be well understood. However, simply outlining the required measures in the experiment design document is not sufficient. The experiment director and his team must actively seek to impose the required controls throughout the planning and execution phases of the experiment.

Experiment Design. The experiment design process is a logical journey from the questions to be answered, or hypotheses to be tested, to the detailed definition of the experiment. Thus the experiment design is the cornerstone of the control regime throughout the life of the experiment, since it sets out in broad terms what needs to be done. Success in designing experiments is rooted in early stakeholder engagement to establish objectives and intent. An integrated analysis and experimentation campaign goes a long way toward providing the framework for detailed stakeholder guidance. Furthermore, nothing allows for the control of variables during experiment design more than early, firm decisionmaking. The longer decisions on scenario, participation, funding, technical environment, and study issues are allowed to linger, the more options the experiment designers must keep open and the harder it is to control the variables that can affect the outcome of the experiment.

Experiment Planning. The planning of major defense experiments requires a management team, which takes the decisions required to settle high-level issues, has oversight on the activities of the various teams, and ensures that the experiment planning and organization develops toward the objectives in a timely manner. A series of reviews throughout the planning period is usually necessary to ensure that the process of preparing for the experiment is remaining on track. For larger experiments, *e.g.*, joint or coalition ones, it is common to employ conferences for this purpose, organized and run by the management team; typically three or four might be used.

Experiment Execution. The experiment management team usually transforms into the control staff during execution. The controller's role is to ensure that the experiment is progressing according to schedule or to be on top of the situation if it is not. The controller observes the players and collects their input daily and works closely with the analysts in monitoring the progress of the experiment. The controller provides feedback to the experiment director and implements changes as required to ensure the event achieves the experiment objectives. In doing so, the controller must deal with military

judgment (observations from the players) and scientific objectivity (input from the analysts).

Experiment Analysis. The analysis or assessment team for an experiment should ideally be derived at least partly from the experiment design team, and they should work closely with the team responsible for the concept under experiment and the team responsible for providing the experiment's technical environment. Initially, they should review the concept and approach planned to conduct the experiment and prepare an analysis plan to meet the needs of the experiment design. During the course of an experiment, analysts compare observations and results and begin to integrate their views of what is being learned from the experiment. As sufficient data is collected, analysts begin to form preliminary insights. However, the temptation to announce some startling finding (especially one that it is believed the experiment sponsor will like) should be resisted at all costs, because it is quite likely that when the analysis is complete, that finding will at best need to be modified, and at worst, changed altogether. Thus, first impressions should generally be conservative; this is an important control consideration.

Data Analysis and Collection

Data collection is designed to support the experiment analysis objectives that in turn rely on a conceptual model underlying the experiment. The data analysis offers the opportunity to revisit the underlying conceptual model identified for the experiment and determines cause-and-effect relationships. A data analysis and collection plan is an essential part of an experiment.

A significant part of the experiment consists of gathering data and information. Interpreting the information into findings and combining them with already known information to obtain new insights tends to be challenging. Once it is determined what needs to be measured, a decision is required to identify the data necessary and to analyze it using appropriate (usually statistical) analysis techniques. The plan ensures appropriate and valid data are generated and that the key issues of the experiment are addressed. When determining analytical techniques to use, an estimate for the number of observations must be considered, depending on the expected variability in the dependent variables and the number of them. It is essential to prioritize and ensure there are sufficient observations for all objectives, measures of performance, and measures of effectiveness requiring analysis. There exist various types of collection mechanisms used in experiments.

Questionnaires (also referred to as surveys) are often used in data collection. They can be used to gather numerous types of information. The participants' background can be obtained through this means. This can be done before the start of the experiment. The participants can also be questioned about aspects of the experiment such as their perceptions about the systems and processes tested, their view on others participating, strengths and weaknesses of the systems and processes as well as recommended improvements.

With information systems becoming more crucial, *Automated Collection Systems* to collect data are now more important. It is important to determine what clock each system that is used to collect data is synchronized to in order to facilitate analysis.

Observers have an important part in the experiment by capturing interactions between participants. For instance they take notes about what is going on, crucial events taking place, notable behaviors and other such activities. Observers can also be used to provide a chronological narrative of the events that occurred. This provides documentation about what happened during the experiment and can be used to explain why certain results occurred.

Ethics, Security and National Issues

This guide describes a number of different aspects of defense experimentation. However, in addition, distinctive national regulations, security rules and practices should not be underestimated and proper consideration must be given to them in planning experiments.

Environmental considerations. Wherever there is live activity, there will be some level of environmental impact. In particular, great care must be taken regarding proximity to historical or cultural sites. As well as legal and multinational environment issues, environmental constraints generally will have an impact on the scope of any live experiment or exercise. It is essential that results be interpreted in the light of all environmentally imposed artificialities. The test and training communities have been working with environmental issues for years and there is no reason for the experimentation community to deviate from the various protocols that already exist.

Security considerations. Even within single-nation experiments, security issues can give rise to real practical problems. In particular, the rise of secure digital C4I and sensitive ISTAR sources (which are often themselves at the centre of the experiment purpose) has resulted in security considerations becoming much more prominent in the design and execution of defense experiments than hitherto. As a general rule, the lower the security classification of these elements, the lower the cost and risk of the experiment and thus experiments should be run at the lowest classification level possible. This is not to say, of course, that undue efforts should be made to make everything unclassified or artificially low in classification. As previously discussed, all experiments are compromises, and the experimenter needs to decide where the benefits of (for example) higher classification and therefore higher fidelity representations of equipments or scenarios outweigh the benefits of using lower classification analogues.

Ethics considerations. Any experiment which involves human subjects and human data collectors could potentially pose ethical issues. By recruiting subjects to undertake an experiment, or by exposing the data collector to a potentially hazardous military environment the experimenter is expecting them to operate outside their normal working practices. Although ethics is a complex field, its fundamental concerns in

professional contexts can be defined. Research that lacks integrity is considered to be ethically unacceptable, as it not only misrepresents what it claims to be but also misuses resources. In addition, there is an obligation for defense experiments to comply with relevant national Health and Safety legislation and to provide working conditions that would ensure, as far as reasonably practicable, a healthy and safe working environment for experimenters and subjects alike.

Communication with Stakeholders

The final product of any defense experiment must be the evidence that the right question has been addressed and the evidence required for its findings to be exploited effectively. This will also provide the experimenter with the necessary foundation for advising on the applicability and feasibility of advancing an evaluated concept, or elements of a concept, toward eventual realization as actual operational capabilities. Good and continuous communication is central to achieving such a successful outcome; and yet it is still possible to find an experiment, or integrated analysis and experimentation campaign, which does not have a rational plan for communicating with stakeholders.[11] A communications plan must consider how the different stages in running an experiment may require different approaches to good communication; stages such as determining the right set of questions and issues to be addressed, maintaining the confidence of key stakeholders that the potential changes to their priorities are being considered, ensuring all stakeholders have appropriate access during the experiment and making sure that they understand the output.

Determining the right set of question and issues. A key prerequisite to a single experiment or campaign is the identification of the origins of the question to be addressed and identification and commitment of key stakeholders. One difficulty is that the obvious stakeholder is often not the person that originally posed the question. Therefore an initial step must be to chase down the origins of the question, and from that define the key stakeholders that need to be influenced. However, the question may arise from many sources and it may not always be possible to directly engage or even identify the original source. For example the question may have arisen from a strategic plan which states that “there is a need to enhance interoperability with our allies to a level which will allow us to undertake concurrent medium scale operations.” This will reflect a political imperative, and whoever is responsible for the strategic plan may have appointed intermediaries whose task is to implement this directive. In this case, these are all key stakeholders, and it is essential to determine their relationships and how they work together. Intermediaries will have formed their own understanding of the question being posed and defined a campaign to implement the directive.

Communicating in the run up to the experiment. Although this will be a particularly busy period, it is essential that regular dialogue be maintained with the

[11] Stakeholders are defined as persons who have a vested interest in the product from the experiment or campaign.

stakeholder community prior to the experiment. By maintaining this regular dialogue, changes in priorities can be quickly identified and accommodated.

Communicating during the experiment. In most cases the major interaction with stakeholders occurs during the visitor day. Visitors should be encouraged to view the entire experimentation process from the pre-brief to the post exercise wash up, and invited to observe and interact with the subjects in a way that does not interfere with the experiment. Additional attendance outside the specific visitor day of stakeholders with a direct involvement in the campaign implementation improves communication in that they are then briefed at regular intervals.

Communicating after the experiment. A well-written report will contain a one-page abstract, an executive summary and a full report. The traditional approach to dissemination of results has been to produce a paper that is sent to key stakeholders, with or without a presentation. While this has obvious merits the general experience is that this approach tends to produce "shelf-ware."[12] It should be remembered that these are busy people who will wish to gain quick appreciation of the key issues and findings, in order to exploit the information. A far better approach is to continue the dialogue with the key stakeholders to determine how the work has been received, to assist in interpreting results and, more importantly, to advise on how it should be exploited. Where the experiment is part of a wider campaign supporting concept or capability development, the experimenter may also have the opportunity to advise on the consequences for the over-arching concept of the particular experiment findings.

[12] A UK term, which means that the report is produced but never read in full.

GUIDEx Experiment and Campaign Planning Flowchart

In order to help practitioners in applying the GUIDEx principles to address their specific problems, the following flowchart was developed. This is by no means a prescriptive recipe for perfect experimentation, but an attempt to lay out the chronological sequence for experiment and campaign related activities and to show the iterations and linkages between various stages of the experimentation process. Indeed, GUIDEx encourages that the specific application of Principles to a given problem should be tailored according to the scale and nature of the issue under investigation. There is no single "best" way to undertake experimentation, rather the skill of the practitioner is to use a degree of artistic license in applying the science advocated within GUIDEx in order to maximize what can be achieved for a given problem under real-world constraints of resources, time, expectation and understanding.

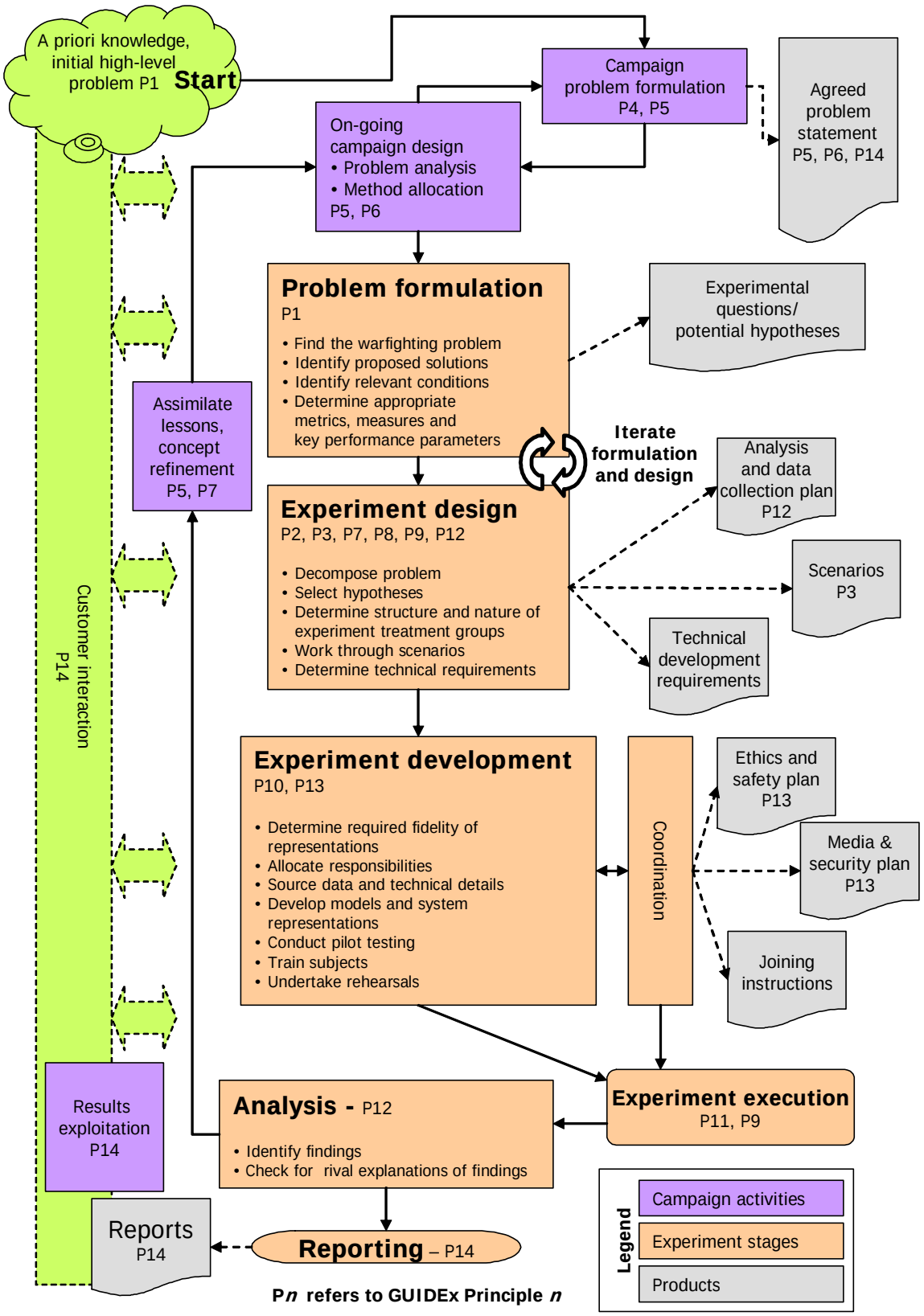
The color code of the flowchart separates the integrated analysis and experimentation campaign activities (in purple) from the specific individual experiment stages (in orange). The products of the experimentation process are indicated by the grey areas, while the customer or stakeholder interactions are shown in green. The flowchart itself begins from the green cloud at the top-left hand corner, representing the initial problem, as posed by the customer.

The campaign of integrated analysis and experimentation then commences with a number of iterations around the campaign problem formulation and campaign design loop in order to develop with the customer an agreed campaign-level problem statement. During this process the campaign designer begins to identify the analytical methods and experiments that might be used to answer the problem. Once a required experiment is identified, the more detailed process of experiment problem formulation can begin. Again, the flowchart suggests that the problem formulation should iterate and overlap with the experiment design in order to ascertain the problem scope suitability for experimentation imposed by real-world considerations. A number of potential experimental questions may require some initial design work to be undertaken before an acceptable, workable and useful problem defined can then be submitted to a complete experiment design and development. The lesson is "be prepared for exploratory activities or false starts before one can move forward with a good concept for detailed design."

The flowchart outlines some of the products needed for successful experimentation, such as analysis and data collection plans, technical development requirements, ethics and safety plans and finally joining instructions for the participants. The practitioner's role at this stage is to manage the competing demands of technical development, customer and player expectation, legislative requirements, rehearsal and training requirements while still maintaining overall control of the scientific and analytical rigor. Finally the experiment itself is executed and the process of analysis and reporting can begin.

In general as the individual experiment is being planned, designed and undertaken, the campaign analysis continues and once the results from the experiment emerge from the collected data, the campaign itself may evolve to take account of the knowledge gained. Lessons must be assimilated. If necessary, further experimentation or analytical activities can be undertaken and the cycle repeats. Throughout this entire process, the interaction with the customer is key to ensuring that the answers generated do indeed answer the questions posed.

Overview


A priori knowledge, initial high-level problem P1
Start
Campaign problem formulation P4, P5
Agreed problem statement P5, P6, P14
On-going campaign design
• Problem analysis
• Method allocation
P5, P6
Problem formulation
P1
• Find the warfighting problem
• Identify proposed solutions
• Identify relevant conditions
• Determine appropriate metrics, measures and key performance parameters
Experimental questions/ potential hypotheses
Assimilate lessons, concept refinement P5, P7
Iterate formulation and design
Analysis and data collection plan P12
Experiment design
P2, P3, P7, P8, P9, P12
• Decompose problem
• Select hypotheses
• Determine structure and nature of experiment treatment groups
• Work through scenarios
• Determine technical requirements
Scenarios P3
Technical development requirements
Customer interaction P14
Experiment development
P10, P13
• Determine required fidelity of representations
• Allocate responsibilities
• Source data and technical details
• Develop models and system representations
• Conduct pilot testing
• Train subjects
• Undertake rehearsals
Coordination
Ethics and safety plan P13
Media & security plan P13
Joining instructions
Results exploitation P14
Analysis - P12
• Identify findings
• Check for rival explanations of findings
Experiment execution
P11, P9
Reports P14
Reporting – P14
Pn refers to GUIDEx Principle n
Legend
Campaign activities
Experiment stages
Products

GUIDEx Case Studies

The following is a high-level overview of the results of the eight Case Studies offered by GUIDEx.

1. *Testing Causal Hypotheses on Effective Warfighting*: This was a series of experiments for a common operational picture (COP) experimental treatment condition using a Persian Gulf air/sea scenario where all parties—higher echelon and lower echelon—had both the national intelligence supported big picture and the local tactical picture. This combination was experimentally proven to be superior technology for such operations, resulting in greater shared situation awareness and better bottom line combat effectiveness.
2. *UK Battlegroup Level UAV Effectiveness*: This experiment supported a major UK UAV acquisition program in demonstrating the huge information gathering potential of UAVs at the tactical level, compared to existing ISTAR assets. However, equally importantly, it showed that if integration into the supported HQs is not achieved effectively, then the resulting information overload can have a hugely detrimental effect on mission success.
3. *UK NITEworks ISTAR Experiment*: The UK, like other nations, is presently investing heavily in ISTAR sensors and systems. However, it is widely recognized that effective information requirements management (IRM) is vital to the efficient use of those systems. This experiment investigated both technological and procedural means of improving IRM. It showed conclusively that a collaborative working environment with appropriate working practices would have a major beneficial effect on IRM effectiveness. This assisted the development of ISTAR management priorities in the UK.
4. *Pacific Littoral ISR UAV Experiment (PLIX)*: This Case Study provides insights difficult to capture without experimentation; the strong hypothesis of identifying and tracking all targets proved not to be attainable even though sensor coverage was nominally complete, pointing to integration requirements for an effective ISR architecture.
5. *An Integrated Analysis and Experimentation Campaign: Army 21 / Restructuring the Army 1995-99*: This campaign demonstrated the importance of detailed problem definition and an iterative approach based on wargaming, field trials and analytical studies. The warfighting concept under test was found to fail under realistic environmental constraints. However, the results led to an alternative concept which is the basis for current Australian Army force development.
6. *The Peregrine Series: a Campaign Approach to Doctrine and TTP Development*: This on-going campaign of experiments and studies is contributing directly to the development of the doctrine for employment of the Australian Army's new Armed Reconnaissance Helicopters and demonstrates how experimentation can be used to inform capability development questions at unit level and below.
7. *Multinational Experiment Three (MNE 3)*: Despite the complexity of the MNE 3 effects-based planning (EBP) experiment and the findings that the concept and supporting tools require further development, the event demonstrated the potential for EBP to make a coalition task force a more effective instrument of power. It also showed the benefits for collaboration to produce the best ideas from a collective thought process in a coalition, which included a civilian interagency component.
8. *Improved Instruments Increase Campaign Values*: While improved experimentation instruments provided the opportunity to generalize some results, they also increased the validity of campaign's results and knowledge generation synthesized for future information management systems.

Part II
GUIDEx 14 Principles

Principle 1.

Defense experiments are uniquely suited to investigate the cause-and-effect relationships underlying capability development

Principle 1 promotes defense experiments to a prominent role in supporting capability development decisions by showing what experimentation and science bring to military transformation. Defense experimentation is uniquely suited to supporting capability development decisions at all levels from force to system. The notion of cause-and-effect is an essential attribute of capability development in that a capability change (the cause) should result in a difference in military effectiveness (the effect). Correspondingly, the principal paradigm of experimentation is changing something and observing what happens. When change occurs under controlled conditions, a conclusion about cause-and-effect is possible. Experimentation is the preferred scientific method for establishing causality; empirically determining what potential effects will result from proposed changes.

Principle 1. Defense experiments are uniquely suited to investigate the cause-and-effect relationships underlying capability development

The reader is encouraged to apply and adapt this Principle. However, the examples herein are based on the specific perspective and experience of different lead-nation authors with contributions from other participants. Therefore they may require supplementary effort to relate them to national perspectives.

Principle 1 asserts the importance of experimentation to the capability development process. This section examines the basic scientific experimentation process and how it can provide the empirical foundation for transforming military forces. Understanding the scientific method and the role of experimentation in science will provide a better understanding of how experimentation can support capability development in the transformation process.

1.1 What Experimentation Brings to Military Transformation

Increasingly, the United States and other nations such as Great Britain, Canada, Australia, New Zealand and indeed NATO itself employ experimentation to assist in developing their future military forces. The United States Department of Defense stresses the importance of experimentation as the process that will determine how best to optimize the effectiveness of its joint force to achieve its vision of the future [US Joint Staff 2000]. An experimentation strategy is also the cornerstone of the U.S. 2001 Quadrennial Defense Review (QDR) transformation strategy [US Department of Defense 2001].

Admiral William A. Owens, United State Navy (USN) (Retired), who served as Vice Chairman of the Joint Chiefs of Staff reiterates that experimentation holds the greatest promise as a method for designing a more effective joint force.

"*Joint experimentation—unconstrained in scope and devoted to defining military structures, organizations, and operational approaches that offer the best promise from new technology—joins [with] joint standing forces as the most efficient, effective, and expeditious means of designing the future in parallel with improving the ability to fight jointly*" [Owens 2002]**.**

Most evident of the recognized importance of experimentation is that the DoD designated US Joint Forces Command (USJFCOM) as the DoD executive agent for joint experimentation in 1998.

Is this confidence in the ability of experimentation to support the military transformation process appropriate? Certainly, experimentation has proven itself in the sciences and technology by producing dramatic advances. Can the methods of

experimentation that have so expeditiously developed science and technology be applied to the military transformation process to achieve similar advances in military effectiveness? The implicit thesis of this guide is that robust experimentation methods from the sciences can be adapted and applied to military experimentation and will provide the basis for advancements in military effectiveness in the transformation process.

Why is experimentation so uniquely suited to the military transformation process? The U.S. Secretary of Defense has written that transforming the U.S. military is essential in order to "defend our nation against the unknown, the uncertain, the unseen, and the unexpected" [Rumsfeld 2002]. Military transformation can be described quite badly "as innovation on a grand scale, sufficient to bring about a discontinuous leap in military effectiveness..." [Krepinevich 2001]. General Richard B. Myers, Chairman of the Joint Chiefs of Staff (CJCS), on the other hand, cautions that "revolutionary changes...should not be the sole focus of our transformational activities" [Myers 2003: p. 6]. Whether transformational change occurs dramatically or incrementally, the key question is how does one decide what to change in order to transform the military?

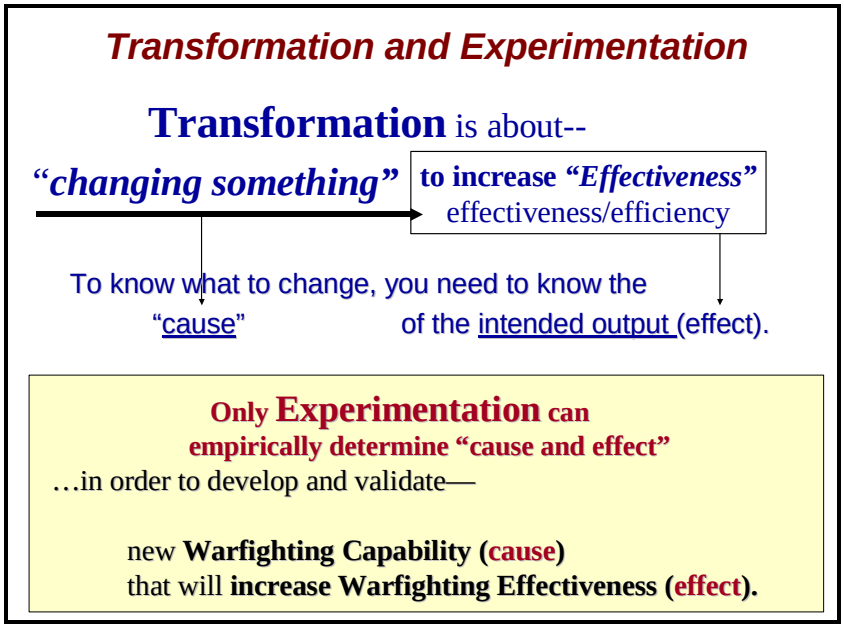


Figure 1 Transformation paradigm

Two essential attributes embedded in the idea of military transformation (Figure 1) are the idea of change and the idea of cause-and-effect. If something in the military is innovated (changed), then it will result in (cause) a change in military effectiveness. Correspondingly, the principal paradigm of experimentation is manipulating (changing) something and observing what happens [Feynman 1999]. When this manipulation is conducted under controlled conditions, conclusion about cause-and-effect can be made. Defense experimentation is uniquely suited to supporting decisions about change to effect transformation. Other techniques are available for causal analysis including: surveys; path analysis; cross-lag panel analysis and case histories [Shadish *et al.* 2002: Ch. 12, 13]. However, experimentation is the preferred mode in science.

While this document focuses on the requirements of experimentation to better support military transformation, one should not discount the role of military experts and operational lessons learned. Military experts represent a critical bank of knowledge in this process. However, sometimes experts do not agree what the best approach should be; and sometimes the "obvious answer" may not be the best. For example, most experts initially agreed that aircraft carriers should primarily be used for long-range surveillance to support battleship tactics. Defense experiments with free-play Blue and Red forces can examine many employment alternatives before going to war and let the experiment data show which alternative is most effective.

Operational lessons learned are critical to identifying how particular warfighting capabilities were organized, equipped, and employed. Lessons learned will also identify the results of military engagements, the number of targets engaged, ordnance expanded, casualties, and so on. The problem, however, is that a lesson learned can only speculate on which capabilities accounted for which effects. In a complex military operation, it is seldom clear exactly why some aspects went well and some did not. This is problematic for transformation when one is interested in deciding which capabilities need to be continued and which deficiencies need to be remedied. Defense experimentation, by its nature, is designed to take this information and systematically sort through the cause-and-effect relationships, thereby lending science to the lessons learned transformation process.

1.2 Science and Defense Experiments

Defense experiments have two characteristics that separate them from other types of experiments. First they examine the determinants of military effectiveness as opposed to experiments in physics, chemistry, and agriculture, which focus on determinants of physical phenomena. Second, defense experiments examine military operations involving humans and their equipment engaged in combat operations. Other types of basic and applied defense experiments conducted in military research laboratories develop new military technologies. These basic defense experiments closely resemble the laboratory experiments one finds in industry and academia. However, we must keep in mind that the focus of GUIDEx is defense experiments.

There is ambivalence toward discussing science and defense experiments. Those responsible for funding defense experiments, and who depend on their results for making decisions, desire that the same scientific rigor responsible for technological and medical advances can be applied to transformation decisions. Many practitioners of defense experimentation, however, are not convinced that "laboratory science" can be applied to the complexity and chaos of military operations. This section discusses science and the scientific method and the basis of experimentation, and provides the foundation for understanding the logic of defense experiments presented in Principle 2.

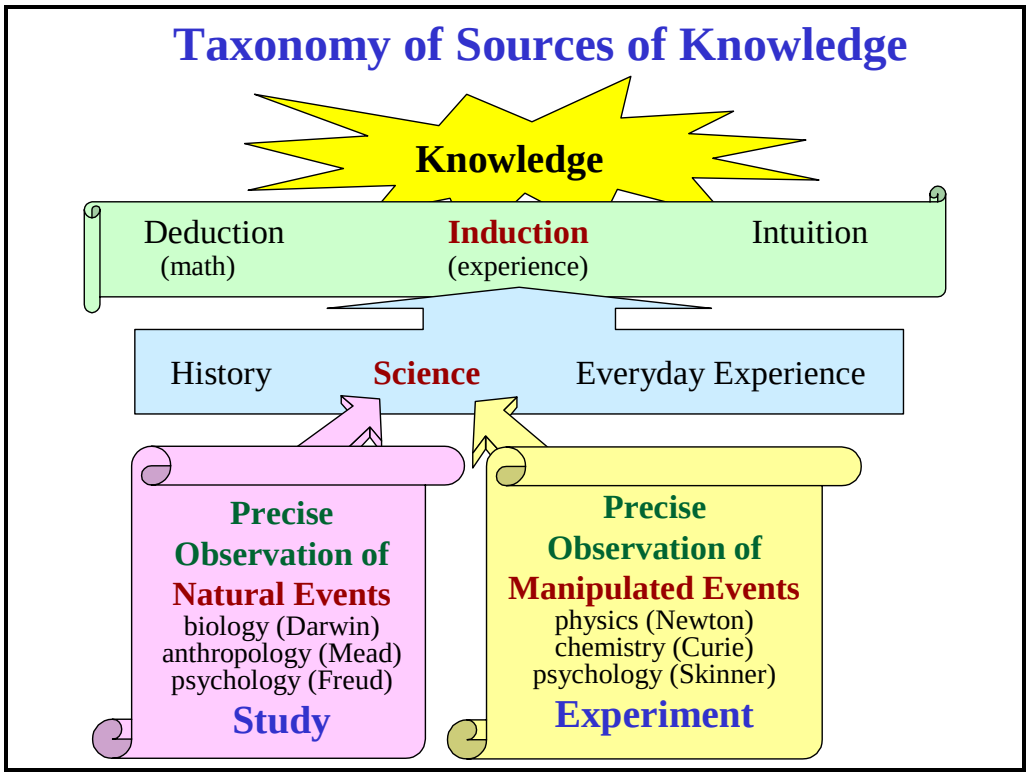


Figure 2 A typical taxonomy of the sources of knowledge

1.2.1 Science and Experimentation

In about 400 BC, Socrates, Plato, and Aristotle investigated the meaning of knowledge and the means to obtain it. Their method was primarily a *rational-deductive* process. Later *empirical-inductive* methods developed by scientists such as Ptolemy and Copernicus focused on precise observations and explanations of the stars. They were not experimenters. When scientists turned from investigating the heavens to investigating earthly objects, they uncovered a new paradigm for increasing knowledge. Because they could *manipulate* those earthly objects, new and exciting answers to questions about objects within their reach were obtainable (Figure 2).

In the early 1600s, Francis Bacon introduced the term experiment, while Galileo conducted experiments by rolling balls down an inclined plane to describe bodies in motion. The realization that manipulating objects would yield new knowledge spawned a new research paradigm, one unimagined in the previous 2000 years of exploring the out-of-reach heavens. The basis of the new science paradigm called experimentation was a research question [Feynman 1999]: "If I do this, what will happen?"

The key to understanding experimentation, and the characteristic that separates experimentation from other research methods, is manipulating something to see what happens. The scientific aspect of experimentation is the manipulation of objects under controlled conditions while taking precise measurements. In its simplest form, an experiment can be defined as a process "*to explore the effects of manipulating a variable*" [Shadish *et al.* 2002: p. 507].

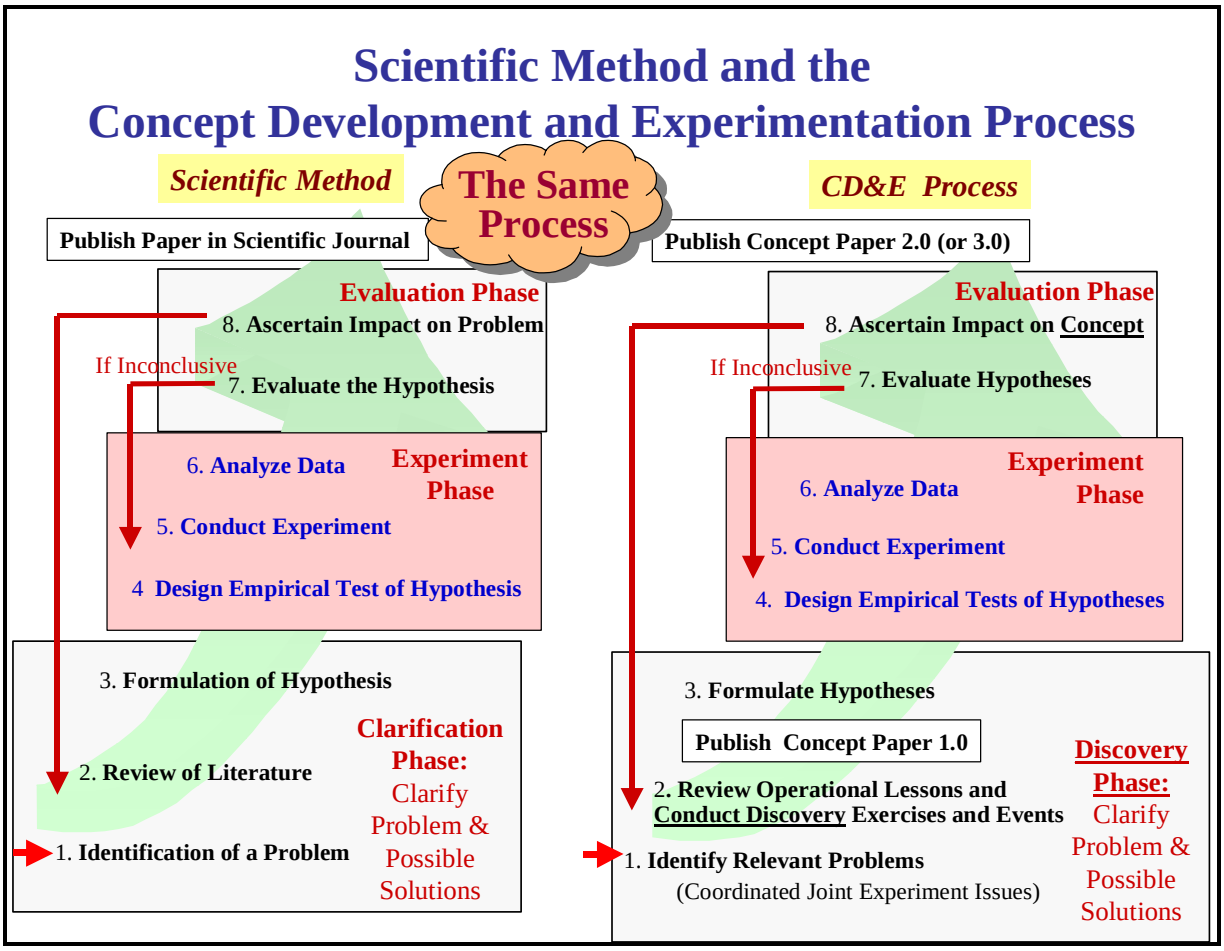


Figure 3 An interpretation of the scientific method

The scientific method for experimentation has evolved during the last 400 years. Figure 3 shows how one nation uses a concept development and experimentation (CD&E) process that progresses through the eight steps of the scientific method. The process begins with discovery to clarify future warfighting problems and seek potential solutions. Current operational lessons learned, the commanding staff, defense planning guidance, combatant commands, and other sources all help to identify and clarify the initial operational problems. Similarly, military experts, history, industry, and academia are important for developing the initial set of potential future solutions.

An initial concept paper summarizes the operational problems discovered and their proposed (hypothesized) solutions. This concept paper provides the basis for defense experimentation. If the experiment results are inconclusive, such that one cannot determine if the original concept was either supported or not supported, then a better experiment is in order. Clear results, on the other hand, whether positive or negative, provide an empirical basis to refine and improve the concept.

1.2.1.1 What is an Experiment?

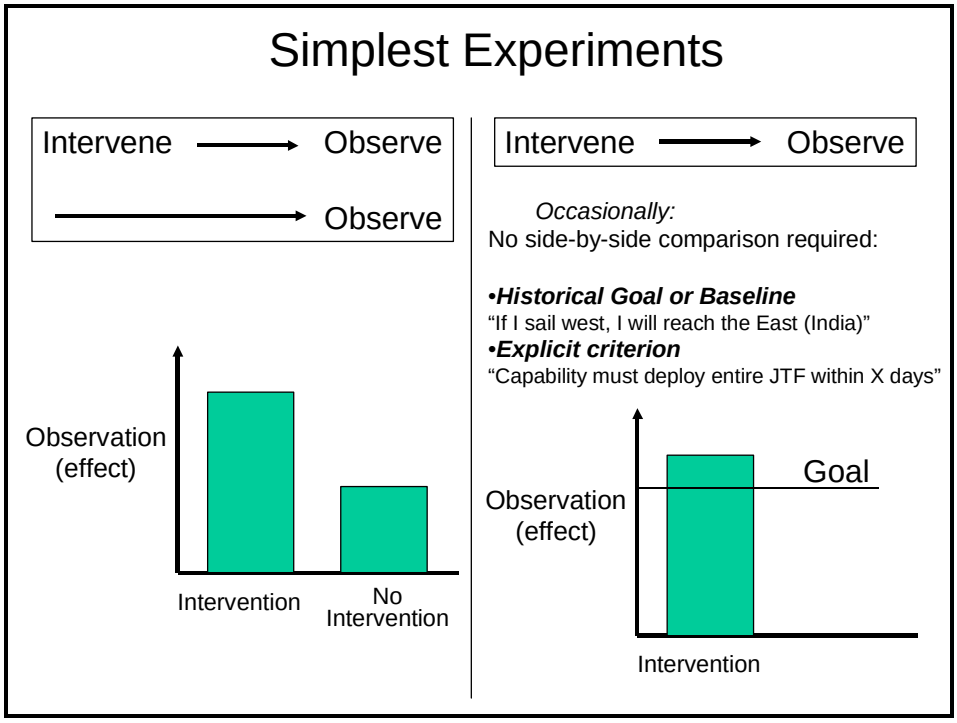


Figure 4 Illustration of simple experiments

As indicated above, all experiments include the notion of doing something or manipulating something. The simplest kind of experiment is displayed in the left-hand side of Figure 4. It compares an intervention (manipulation) to a non-intervention. It is a side-by-side comparison you might see your son or daughter propose for their science fair. Plant two seeds in a box full of soil. For one seed, intervene by adding fertilizer and for the other, no fertilizer. Water both regularly and record the height of both growing plants at some future date.

A simple defense experiment might be to start with two units and **intervene** by giving one of the units a new capability, a new piece of equipment, and then observe both units as they execute a military exercise. At the completion of the exercise compare the two units on some measure of military effectiveness, perhaps the time to complete the exercise.

Occasionally, one can design an experiment with no side-by-side comparison (right-hand side of Figure 4). This occurs when there is a well-established threshold, so that the results of an intervention can be compared to this threshold instead of comparison to an alternative condition. Sometimes thresholds are available from historical knowledge. For example, prior to 1947 no one had flown faster than the speed of sound. Experimental aircraft were flown to achieve this threshold rather than to beat some other aircraft. Thresholds are also available in the military acquisition arena where a system must meet a specific threshold, say fire so many rounds per minute, before the system will be funded. Experiments designed to compare a manipulation to a threshold are often called tests.

1.2.1.2 Definition of Defense Experiment

Over 35 different definitions of "experiment" are available when conducting a web dictionary search.[13] Two common themes permeate these definitions: the notion of "**doing something**" and the notion of "**new knowledge**." Shadish, Cook, and Campbell provide a third theme in their 2002 monumental book ***Experimental and Quasi-Experimental Designs for Generalized Causal Inferences*** [Shadish *et al.* 2002]. They state that the purpose of an experiment is to ascertain the truth or falsity of a causal inference. Identifying experiments with the investigation of causality is a very useful construct for understanding defense experiments. Causality is central to the transformation process. Military decisionmakers need to know what to change in order to improve military effectiveness. This is to say that the antecedents of effectiveness, the causes of effectiveness, must be understood in order to change effectiveness. Effectiveness can only be improved by altering its antecedents, its causes.

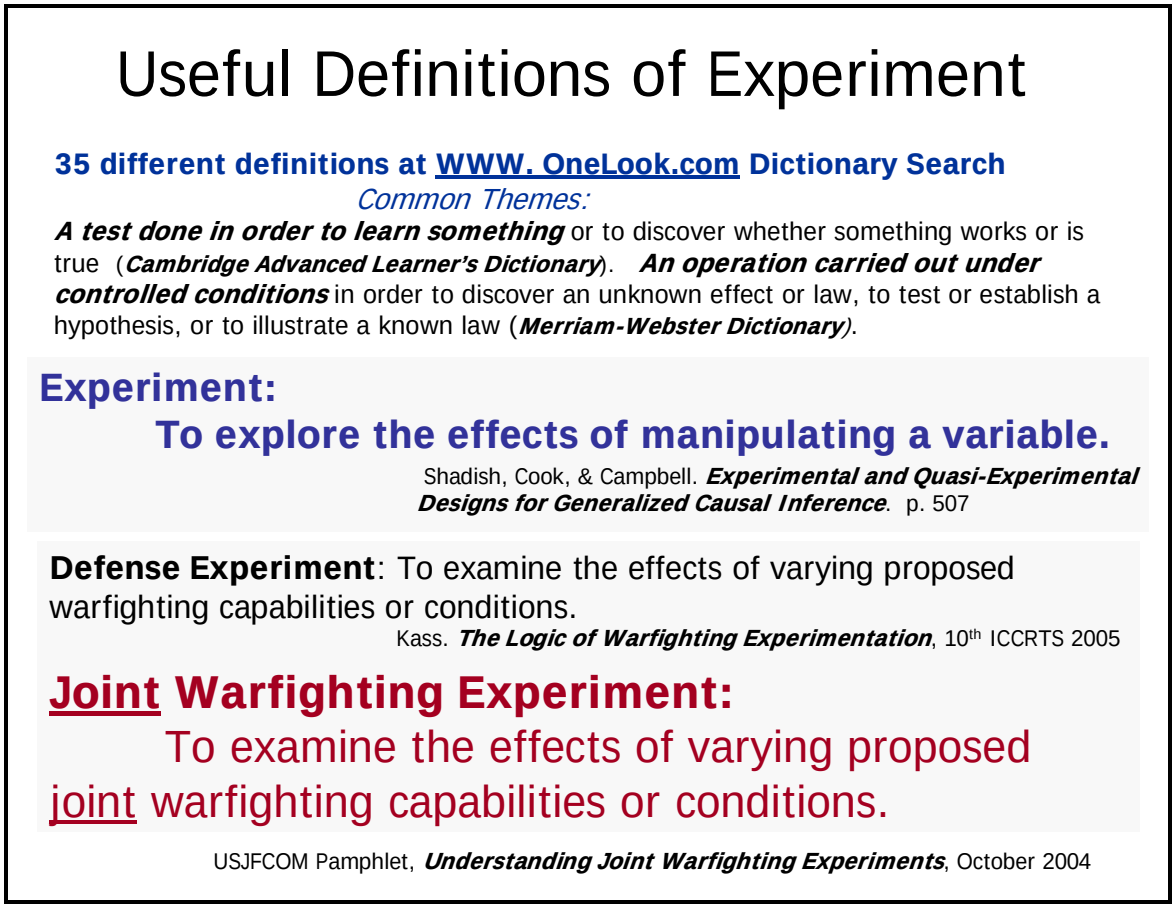


Figure 5 Some useful definitions of experiment

Using the three themes of doing something, gaining knowledge, and cause-and-effect, more formal definitions of defense experiments can be offered as in Figure 5.

[13] http://www.onelook.com/

1.2.2 Cause-and-Effect in Defense Experiment

The notion of cause-and-effect is inherent in the very language of an experiment and in the basic experiment paradigm; **let's do this, and see what happens.** All warfighting innovation questions can be translated into a cause-and-effect question expressed as: Does **A** cause **B**? A proposed military capability **A**, a new way of warfighting, is assessed by determining the extent by which the new capability **A** produces (causes) an increase in effectiveness **B**. The idea of cause-and-effect is central to constructing the experiment hypothesis: If the unit uses the new capability **A**, then it will increase its effectiveness **B**. The hypothesis provides an expectation concerning the causal proposition to be observed in the experiment. The definition of an experiment trial follows naturally. A trial is one presentation of the capability **A** to see if effect **B** occurred; and the non-presentation of the capability **~A** to see if effect **B** does not occur.

Many writers on defense experimentation miss the fundamental aspect of cause-and-effect as the underlying logic of experimentation. The idea of **cause-and-effect** will permeate much of the ensuing discussion on defense experimentation. It is worthwhile to pause here and explain why. Experimentation is all about cause-and-effect.

> **"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 *et al.* 2002: p. 3].

Try to imagine an attempt to design a defense experiment devoid of any interest in cause-and-effect. In this instance the experimenters might have some new technology and want to give it to a unit to see what they do with it, to see if it helps them to do anything. The experimenters believe that they should not control anything because they do not want to preclude any possibilities. They may also contend that they do not even have sufficient information to formulate a hypothesis: *If this capability, what?*

After some consideration, however, they realize that contradictions abound in this "non-causality" approach. If they want to know if the new technology helped the unit do anything different or better, they have to enter the world of cause-and-effect, *i.e.*, "Did the technology (cause) produce a change (effect) in the unit's performance?" Moreover, the initial hypothesis could simply be stated as "If the unit employs this new technology, they will develop procedures to do some military task X better." Skeptics might counter that the experimenter does not know that the new technology will, in fact, produce a change for the better; and thus, it is premature to specify a hypothesis with a positive outcome. The response to this objection is that **the role of hypotheses is not to state what is known, or what we are certain of, but rather to state an educated guess on what we are looking for.** There is always the chance that the experimental data may prove the hypothesis wrong.

It is also difficult to think of reporting the results of a defense experiment that did not involve cause-and-effect. One would have to report something like the following: "*In this experiment the unit used the new capability and the unit accomplished task X; but*

we do not know how the unit accomplished task X. The unit may have accomplished task X even if they did not have the new capability." If these were actual reported results, the worth of the experiment would be questioned. And yet, the centrality of cause-and-effect has been heretofore overlooked. Indeed, the ability of experiments to resolve causal inferences is what makes them uniquely suited to address the underlying issue of transformation—what future capabilities are required to cause an increase in military effectiveness in future warfare?

> **"How do we know if cause-and-effect are related?**
> **In a classic analysis formulated by the 19th-century philosopher John Stuart Mill, a causal relationship exists if**
> ***(1) the cause preceded the effect,***
> ***(2) the cause is related to the effect, and***
> ***(3) we can find no plausible alternative explanation for the effect other than the cause.***
>
> **These three characteristics mirror what happens in experiments in which**
> ***(1) we manipulate the presumed cause, observe an outcome afterwards,***
> ***(2) we see whether variation in the cause is related to variation in the effect, and***
> ***(3) we use various methods during the experiment to reduce the plausibility of other explanations for the effect...*"** [Shadish *et al.* 2002: p. 6].

In Case Study 1 we have an example of a series of controlled experiments which were designed to test the hypothesis that if a warfighting team shares a common operational picture of the battlespace (cause, **A**), then they will kill more of the enemy in combat while sustaining fewer Blue losses (effect, **B**). Since these experiments were conducted with military units operating under controlled conditions **C**, the investigators were able to infer with confidence that the observed superior combat performance was, in fact, due to the warfighting teams' use of the prototype capability under investigation.

1.3 How Experiments Support the Capability Development Process

Some countries use a concept development and prototyping process, and experiments are required throughout such a process. Experiments provide an empirical method for exploring, refining new capabilities, and validating prototypes for force capability implementation (Figure 6 and Figure 7).

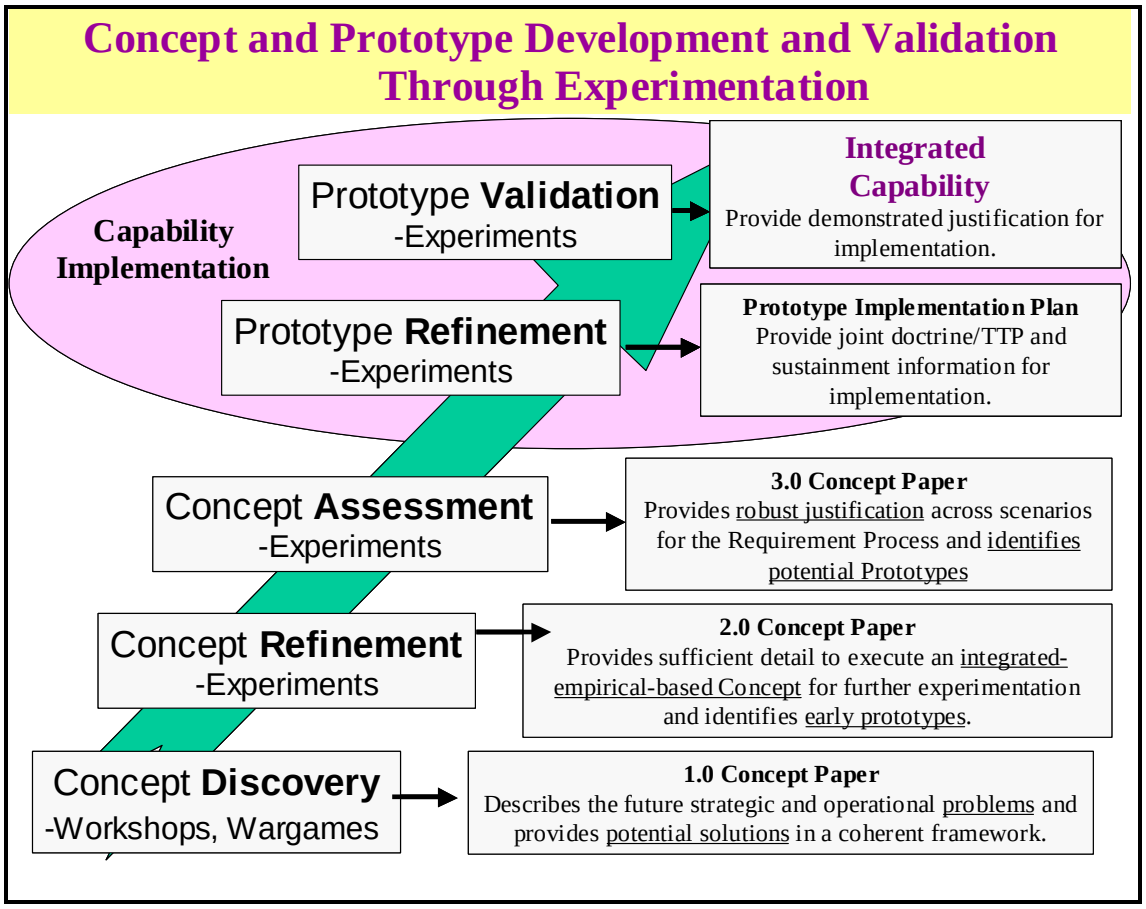


Figure 6 Concept and prototype development and validation through experimentation

Warfighting Doctrine – describes how the force fights today. "Fundamental principles that guide the employment of forces of two or more Services in coordinated actions towards a common objective." (JP 1-02)
Warfighting Concept – describes how the force will employ future capabilities to fight in the future.
Warfighting Capability – A combination of means (process, organization, or system) and ways designed to achieve a desired effect. It represents the potential to perform a task under conditions and to standards necessary to enact the force commander's plan.
Warfighting Prototype – An initial working model of a capability designed to support operational concepts or operational requirements and may consist of people, processes, and technology.
Warfighting Experiment – to examine the effects of varying proposed warfighting capabilities or conditions.

Figure 7 Related definitions

During the concept discovery phase, military and industrial experts review current operational lessons and apply the lessons of military history in order to clarify the future environment. Working through conferences and seminar-type wargames, these experts also identify potential future capabilities that may provide solutions to future uncertainties. An initial concept paper provides a summary of the future operational problem and proposed capability solutions within a coherent framework.

During concept refinement, experiments empirically quantify the extent that proposed capabilities solve military problems. Experiments also examine capability redundancies, tradeoffs, and reveal capability gaps. Discovery phase activities only speculate whether proposed future capabilities would solve identified gaps in military effectiveness, whereas experimentation empirically demonstrates causality [Shadish *et al.* 2002: p. 3]. Concept Paper version 2.0 summarizes the refinement phase with a description of an integrated, optimized set of capabilities for the identified problem. In some instances, robust experimentation may suggest early prototypes for proposed implementation.

Experiments during concept assessment investigate the robustness of the solution developed during refinement over a full spectrum of possible future military operations. These experiments examine and adjust the optimized concept under many different future conditions, environments, and scenarios to ensure the refinement phase did not over optimize the preliminary solution. Results from this phase provide the robust justification to generate prototypes for eventual implementation.

Prototypes identified from robust capabilities are often not immediately ready for fielding. Experiments during prototype refinement turn capability surrogates into implementable capabilities by developing complete prototype implementation packages and strategies for the prototype's intended operational environment. These experiments examine the latest hardware solutions and software updates, interoperability with existing fielded systems, and develop detailed tactics, techniques, and procedures to facilitate prototype training and implementation.

Experiments during prototype validation provide the final demonstrated evidence to the combatant commander that the prototype can operate within theater and improve operations. Often these experiments are embedded within exercises or training events and are used to validate the predicted gains in effectiveness of the operational force.

Principle 2.

Designing effective experiments requires an understanding of the logic of experimentation

Principle 2 develops the logic of defense experiments along a mnemonic of 2, 3, 4, 5, and 21. The logic illustrates there are two (2) parts in an experiment hypothesis (if and then sides); three (3) decisions to resolving the conditional proposition contained in the hypothesis statement; four (4) requirements for a valid experiment (ability to use the capability, ability to detect a change, ability to isolate the reason for change, and ability to relate results to an operational environment); five (5) components to every experiment (the treatment, the experimental unit, the effect, the trial conditions, and analysis); and twenty-one (21) general threats associated across the five experiment components that make it difficult to meet the four experiment validity requirements.

This logic has a threefold purpose. It illustrates that there is a coherency in the piece-parts of an experiment and the art of designing valid experiments. Second, this logic provides a framework for organizing and understanding the interrelationships among different lists of existing experiment “best practices” or “good techniques.” And finally, it provides a rationale for tradeoff considerations among competing best practices to assist in designing individual experiments and campaigns.

Principle 2. Designing effective experiments requires an understanding of the logic of experimentation

The reader is encouraged to apply and adapt this Principle. However, the examples herein are based on the specific perspective and experience of different lead-nation authors with contributions from other participants. Therefore they may require supplementary effort to relate them to national perspectives.

2.1 The Logic of Defense Experiments: "2, 3, 4, 5, 21"

It has always been difficult to translate "design of experiments" textbooks into useful prescriptions for defense experiments. Principally because designing defense experiments involves more compromises due to time and resource constraints than in most businesses. These restrictions prohibiting the design of textbook experiments has led some to admonish that defense experiments operate by a different set of principles than scientific experiments. Often this translates to a more relaxed set of principles, prompting a *laissez-faire* approach to designing defense experiments. When faced with constraints, however, the key is not to abandon the basic principles but to apply the principles in a rational manner to accomplish experiment goals. The key to the rational application of experimentation principles is an understanding of their logic. Case Study 3 showed that though it is difficult, it is feasible to follow such logic. The logic of experimentation can be expressed by a mnemonic in the numbers 2, 3, 4, 5 and 21. The following discussion focuses on how each of these numbers represents a sequential state in the logic of experiments. Subsequent sections will show how this logic is applied to design more effective individual defense experiments (Principle 3) and more effective **integrated analysis and experimentation campaigns** (Principle 7).

2.1.1 Two (2) Parts in Experiment Hypothesis

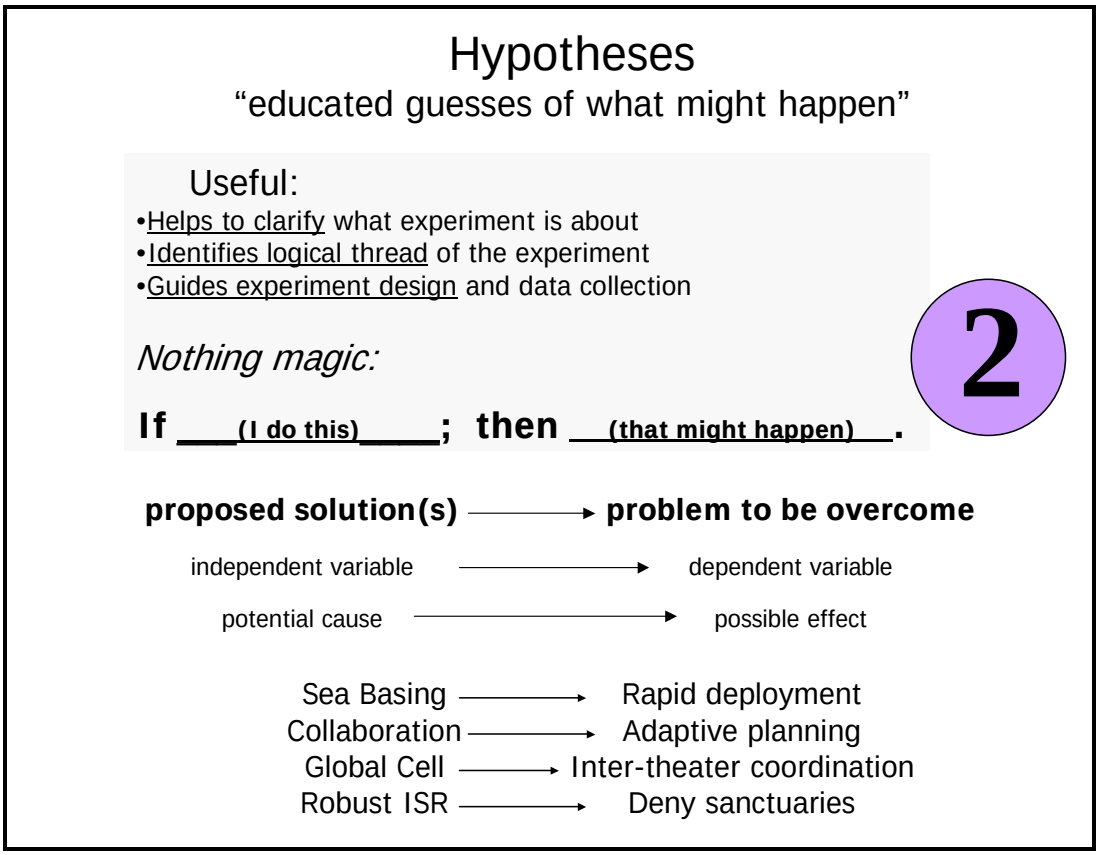


Figure 8 Two-sided hypotheses

The number "2" represents the two components of the hypothesis in Figure 8—the left-hand side and the right-hand side, the "if" side and the "then" side. There are two basic ways one can approach the experiment hypothesis. In most cases one has an operational problem that needs a solution. These operational problems are in the form of a requirement; such as the requirement to deploy forces more rapidly or the requirement to deny the enemy the use of sanctuaries where they can rest and restore. In this instance, the "then" side of the hypothesis comes first and concept developers are in search of possible solutions. When they think they have found one or more solutions, they are ready to express the "if" side of the hypotheses followed by the "then" side expressing the potential resolution of the requirement: *If New Solution X is used, then Operational Problem Y might be solved.*

A second approach to hypothesis development is to begin with the left-hand side. This occurs, for example, when a new technology is available and experiments are conducted to determine if the new technology has military application or utility. In this case, the new technology is the "proposed solution" and it is in search of a military problem to be solved or military tasks that can be enhanced. Often the technology sponsor offers ideas for possible applications. The hypothesis could be formulated as follows: *If New Technology X is employed, then Operational Tasks Y and Z will be enhanced.*

2.1.2 Levels of Hypotheses

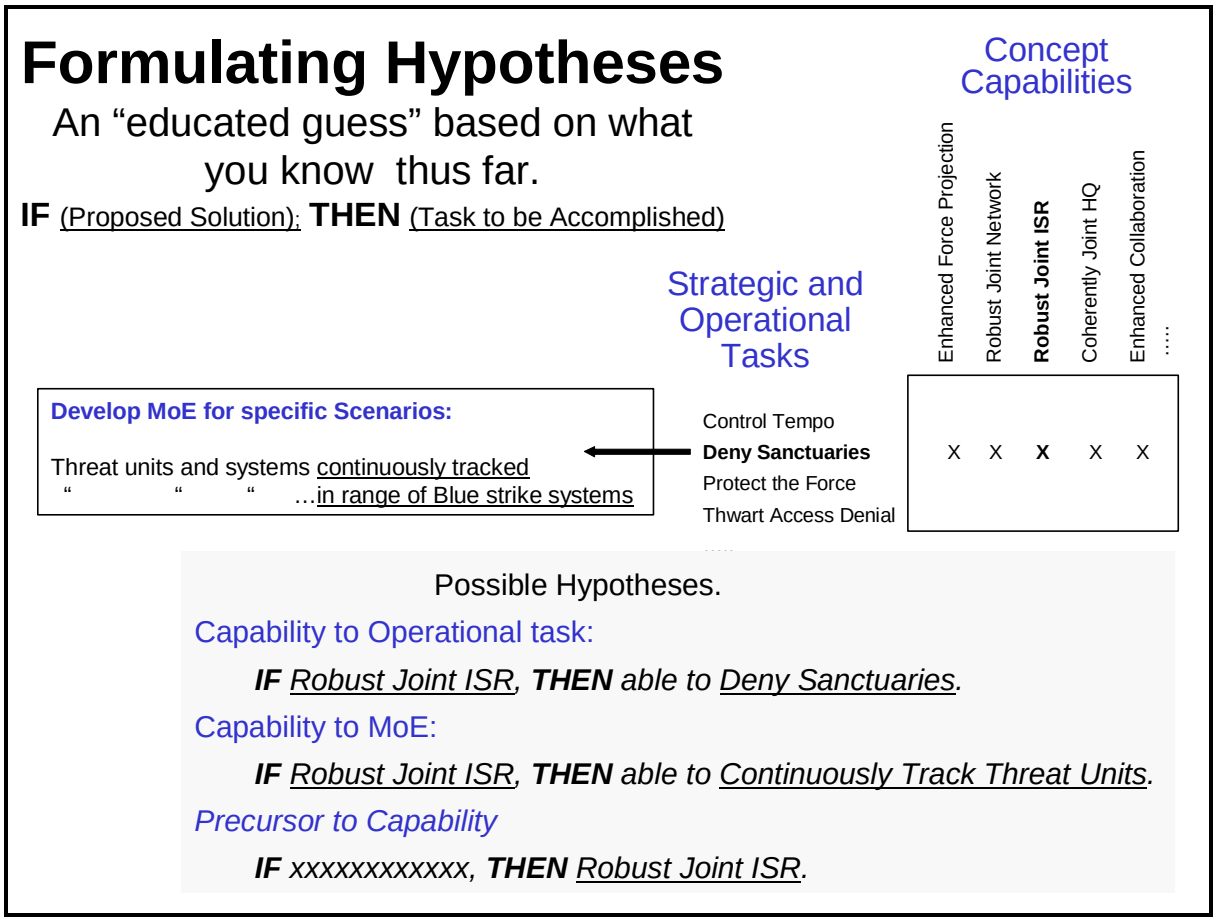


Figure 9 Formulating hypotheses

Figure 9 above combines the approaches to hypothesis development and depicts their relationship between capabilities and tasks. The problem to be overcome can be characterized as the strategic, operational, or even tactical task to be accomplished. The potential solution can be characterized as the potential capability or concept capability. Once new capabilities are mapped to appropriate tasks, the capability is expressed as the "if" portion of the hypothesis and the task to be accomplished or enhanced is expressed as the "then." This high-level "capabilities hypothesis" needs to be translated into a number of "experimental level" hypotheses (capability to MoE). This is accomplished in Figure 10 below by developing measures of effectiveness (MoE) for each operational task. From these experimental hypotheses, the experiment analyst can develop statistical hypotheses and conduct statistical analysis of the data to determine if the results support the hypotheses to some level of confidence. More will be said about the value of statistical analysis in Principle 6.

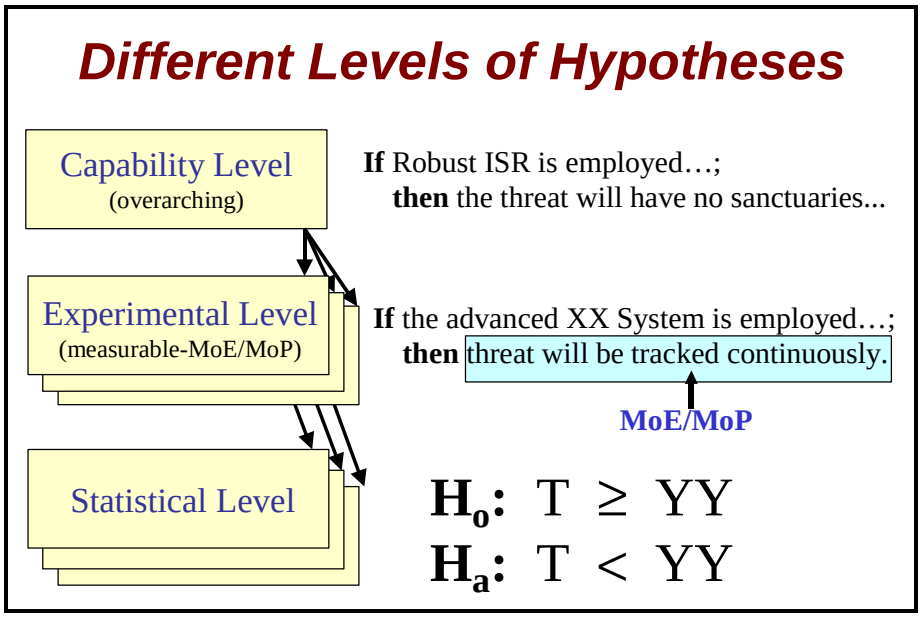


Figure 10 Some levels of hypotheses

2.1.3 Experiment Hypotheses and Null Hypotheses

The notion of the **null hypothesis H_0** is well established in statistical analysis and classical experimentation. The original requirement for null hypotheses provided a means to quantify the probability that a particular sample of data could be said to be derived from a particular hypothetical "parent" distribution. The technique goes something like this in a concrete example. The experimenter has three different riflemen employing a new weapon in an experiment. Prior to the experiment, these riflemen are considered to represent a hypothetical population of riflemen who use the current weapon and historical data indicate that riflemen with the current weapon score an average 250 points on the rifle range. Since this is an average, sometimes shooters with the current weapon scored higher and sometimes lower. During the experiment, the operators achieved an average score of 275 with the new weapon.

The question is, does this post-experiment sample represent just a variation from the original "current system" population; or does it represent a different "improved" population? To answer this question, the experimenter constructs a hypothetical population of current-weapon shooters based on a historical average of 250. This is the **null-hypothesis H_0** population that represents the situation if the experiment treatment does not work. The **alternative-hypothesis H_a** population represents a more speculative, and currently non-existent, population that on the average is better than the null-hypothesis population. It represents what a new population of riflemen will look like if the new weapon is better. Identification of statistical parameters of the null hypotheses is a pre-condition to determining analytically if the results of experiment sample are still similar to the null-hypotheses population; or if the results are sufficiently extreme (higher than the historical average) to "reject" that idea and declare, by default, that it is more likely the sample now represents the new, alternative population.

By convention, the capability- and experiment-level hypotheses illustrated in Figure 10 are worded to reflect what statistics consider the alternative hypothesis. This is the best way to communicate the purpose of the experiment. Often the null hypothesis, the status quo, is unstated at the capability- and experiment-level hypothesis because it is obvious, or at least implied; *e.g.*, if the experiment capability does not work, the "threat will continue to have sanctuaries" and "the threat will not be continuously tracked."

2.1.4 Experiment Hypotheses in Training Exercises

The use of operational tasks for the "then" portion of hypotheses is quite useful when defense experiments are conducted in conjunction with military training exercises. Many opportunities exist to explore new technologies and processes during training exercises. The hypothesis associated with this type of experiment is a natural summary of what is proposed to result from the insertion of something different. Military training exercises are built around a series of tasks, conditions, and standards. In the joint training arena these are documented in the Universal Joint Task List (UJTL). The task specifies what needs to be accomplished, conditions provide the context, and the standards provide the measures of effectiveness. The associated capabilities-level hypothesis would look like this "*If the JTF staff employs new capability XXX, then Task YYY will be enhanced*." The corresponding experimental hypothesis might be *"If the JTF staff employs system XXX, then Task YYY will be accomplished in less time (MoE)."*

2.1.5 Concerns about Hypotheses

A number of concerns have surfaced in recent years about the use of hypotheses in defense experiments. These concerns take one or more of the following forms:

1. There is not enough information to formulate hypotheses in early defense experiments.
2. Hypotheses are too constrictive in early defense experiments and are thus detrimental to serendipitous discovery.
3. Defense hypotheses tend to be too general and thus not very useful to the experimenter.
4. Defense hypotheses are not justified because hypotheses are supposed to be derived from theory and there is no military theory.
5. Hypotheses require cause-and-effect analysis (described below) and warfighting data are not sufficient for determining cause-and-effect.
6. Hypotheses are not appropriate for messy field experiments; they are only useful in "controlled" experiments.

In general, these concerns arise for two reasons. First, hypotheses are thought to be formal deductions derived from a scientific theory. This is a very narrow view of hypotheses. Few, even science experiments, are derived from formal scientific theories. Hypotheses are "educated guesses" or formulations of expectations to guide the experiment design. Second, because hypotheses are "guesses," it is mistakenly believed that wrong hypotheses will too narrowly focus the experimenter and preclude seeing spontaneous, serendipitous results. All experimenters are trained to watch for the unanticipated results. If we understand hypotheses as educated guesses, we understand they are only a starting point. Without hypotheses, there is no expectation;

and without expectation there can be no unanticipated findings. The first key to serendipity is to be sensitive to the possibility of "finding unanticipated findings" by realizing that hypotheses are only educated guesses and could be wrong. The second key is to enhance the possibility of defense experiments developing unanticipated events by allowing both Blue and Red Forces great latitude in using and attempting to counter the new experimental technology. Unanticipated findings can be used to define new hypotheses for a subsequent experiment or a distinct unplanned analysis of the current experiment. They are the byproduct of good experiment design and quality of data analysis and collection. Basically, new hypotheses are educated guesses or induction from careful observations that depend on the quality of data collected from an experiment.

2.1.6 Three (3) Logical Steps to Resolve Hypotheses

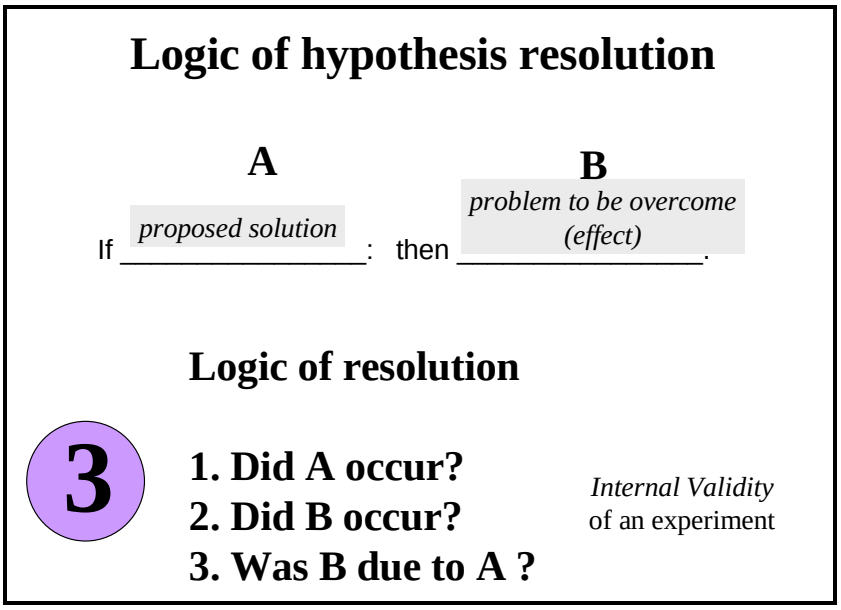


Figure 11 Three logical steps to resolve hypotheses

There are three considerations (Figure 11) in resolving the conditional proposition contained in the hypothesis statement.

1. The first logical question is whether the proposed solution, the left-hand side of the hypothesis, was adequately represented in the experiment. This is not always easy to do given that new proposed solutions often involve surrogate software, hardware, and new procedures that are to be implemented temporarily for the first time.
2. The second question is whether the experimenter was able to observe the right-hand side of the hypotheses. That is, did the experiment produce evidence in an objective manner that the problem to be solved was, in fact, solved?
3. Given that the proposed solution was adequately represented and given that progress was observed in solving the problem, the third logical question concerns whether the observed problem resolution was due to the proposed solution. This third component of a hypothesis is the toughest challenge in defense experiments where so many alternative explanations of positive results exist; for example, the players with the proposed solution were better trained or more motivated.

2.1.7 Four (4) Requirements for a Good Experiment

What is a good experiment? How does one tell a good experiment from a bad experiment? The scientific term for a good experiment is **valid** experiment. Four logically sequenced requirements must be met to achieve a valid experiment. It should come as no surprise that the first three requirements reflect the three considerations of hypothesis resolution just discussed. This further reflects the centrality of hypotheses to defense experiments. The fourth requirement reflects the relevance of the defense experiment to operations outside the experiment environment. The four requirements[14] represent a logical, progressive sequence within themselves. If each successive requirement is not met in sequence, there is no need to proceed to the next one.

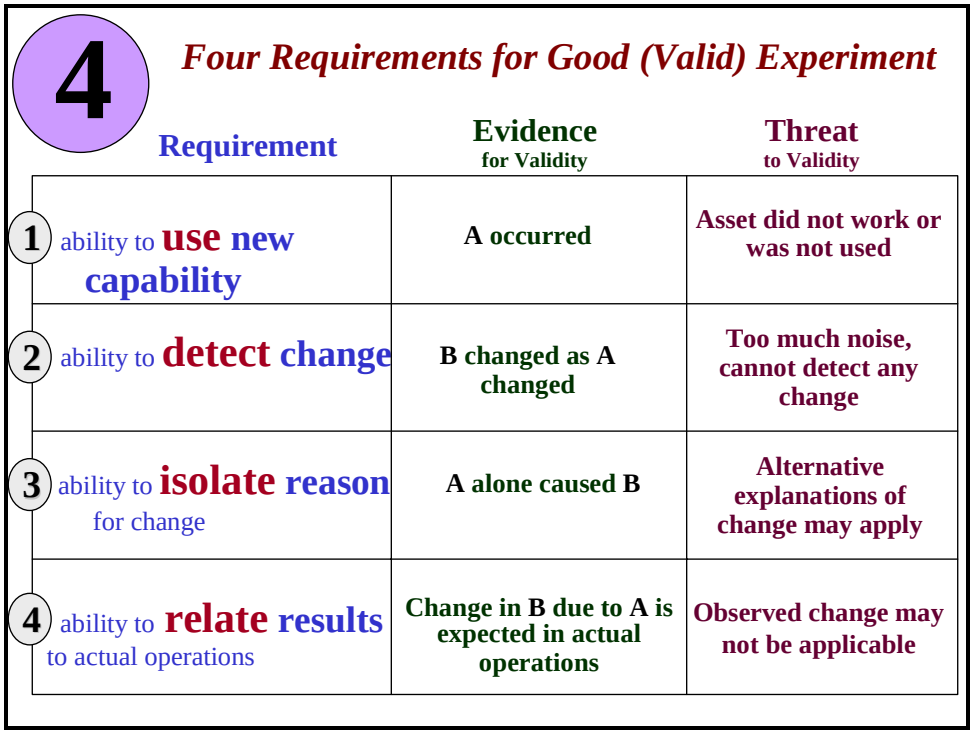


Figure 12 Four requirements for good (valid) experiment

The Figure 12 simple example illustrates these four requirements. Suppose a proposed concept postulates that new sensors will be required to detect time-critical targets. One experiment to examine this proposition might be a two-day military exercise where the old array of sensors is employed on Day One 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 if the new sensors **A** will cause an increase in detections **B**.

[14] The four validity requirements presented here were adapted from [Campbell and Stanley 1963: Ch. 1, 2]. Their four requirements were combined into three (Requirements 2, 3, and 4) and Requirement 1, ability to use the capability, was added. This requirement was implicit in Cook and Campbell [Cook and Campbell 1979] but tied solely to external validity. Making "employment of the capability (treatment)" a separate category under internal validity reinforces the logic of defense experimentation presented here.

2.1.7.1 Ability to Use the New Capability

In most defense 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 the experiment players to its optimal potential and allowed to succeed or not succeed on its own merits. Unfortunately this ideal is rarely achieved in defense experiments. It is almost a truism that the principal lesson learned from the majority of experiments is that the new capability, notwithstanding all of the expended effort, was not ready for the experiment. There are a number of things that go wrong with experimental surrogate capabilities. The hardware or software does not perform as advertised or as anticipated. The experiment players are frequently undertrained and not fully familiar with its functionality. Because it is new, the techniques for optimum employment are not mature and will, by default, be developed by the experimental unit during the experiment trial. These threats and others to meeting the first experiment validity requirement will be discussed further in Section 3.1. If the experimental sensors **A** could not be functionally employed during the experiment, there is no reason to expect that they will affect the ability to detect targets **B** any greater than the current array of sensors, which is the next experiment validity requirement.

2.1.7.2 Ability to Detect Change

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

To detect ***change***, experiments require two or more trials: before and after treatment, various treatment levels, alternative competing treatments, or the same treatment under different conditions. Change detection also requires a high signal-to-noise ratio so that difference between one trial and the next trial will be noticed above the experiment noise level. The threats to the ability to detect change, and further details on attenuating these threats, are the topics of Section 3.2.

2.1.7.3 Ability to Isolate the Reason for Change

Let's suppose the experimenter met the first two requirements: the new array of sensors was effectively employed and the experimental design reduced variability and produced an observable change (increase) in the percent of detections. The question

now is whether the detected change was due to the intended cause, changing from old sensors to new, or due to something else. The scientific term for alternative explanations of experimental data is ***confounded*** results. In this example an alternative explanation for the increase in detections on Day Two is that it was due to a learning effect. The sensor operators may have been more adept at finding targets as a result 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 would dramatically change the conclusion of the detected change.

Scientists have developed experimental techniques to eliminate alternative explanations of the cause of change. These include counter-balancing the presentation of stimuli to the experimental unit, the use of placebos in drug research, use of a control group, randomizing participants between treatment groups, and elimination or control of external influences. These techniques will be discussed more fully in Section 3.3.

2.1.7.4 Ability to Relate the Results to Actual Operations

Again, let's suppose that the experiment was successful in employing the new capability, detecting change, and isolating the cause. Now the question is whether the experimental results are applicable to the operational forces in actual military operations. Experimental design issues supporting operational realism revolve around the representation of surrogate systems, the use of operational forces as the experimental unit, and the use of operational scenarios with a realistic reactive threat. More details on enhancing operational realism in order to extend experimental results to real operations are provided in Section 3.4.

2.1.8 The Four Experiment Validity Requirements in Perspective

These four requirements for a good experiment are applicable to all experiments, whether conducted in a prestigious science lab, as a high school science project, or as defense experiments. In this context, a "good" experiment is synonymous with the scientific notion of a "valid" experiment. A valid experiment can be defined[15] as an experiment that provides sound evidence for ascertaining the truth or falsity of the causal proposition formulated in the experiment hypothesis. The first three experiment validity requirements represent the internal validity of the experiment; the ability to determine if a causal relationship exists between two variables. The fourth requirement represents the external validity of an experiment, the ability to generalize the cause-and-effect relationship found in the experiment environment to the operational military environment.

Familiarity with the four requirements for a good experiment is useful for understanding why a defense experiment can fail. One often hears the expression "There is no such thing as failure in experimentation, because we always learn something from every experiment." This statement can be misinterpreted. The statement should mean that when the results of a valid experiment indicate that a new experimental capability did

[15] Definition of validity and internal and external validity based on [Campbell and Stanley 1963: p. 37].

not live up to its expectations, as indicated by the hypothesis, this is not a failure for the experimentation process. An experiment that produces clear evidence for or against the hypothesis is a success.

Unfortunately, experiments can fail. They can fail to provide the information necessary to resolve the hypothesis, *"Did the new capability cause a change?"* If the experiment provided definitive data that the proposed new capability did not live up to expectations, the experiment was successful. If, on the other hand, one still does not know if the proposed capability is useful or not at the completion of the experiment, then the experiment failed. The experiment was poorly designed. In this situation, little was learned about the utility of the proposed capability. All that was learned was that the experiment was poorly designed.

Understanding the four requirements for a good experiment will go a long way toward avoiding failed experiments. The purpose of this section is to present the rationale and examples of good scientific experimentation practices that can be applied to military experimentation. A good experiment is one that increases knowledge. A poorly constructed experiment is one that casts doubts on any of its findings, thus failing to increase our knowledge about the hypothesis. The only knowledge gained in a poor experiment is a better understanding of how to conduct a more valid experiment to meet the four experiment validity requirements.

2.1.9 Five (5) Components of an Experiment

All experiments—large or small, field or laboratory, military or academic, applied or pure—consist of five components [Cook and Campbell 1979]:

1. The treatment, the possible cause **A**, is the proposed capability, the proposed solution that is expected to influence warfighting effectiveness.
2. The possible effect **B** of the treatment is the result of the trial, an increase or decrease in some aspect of warfighting effectiveness.
3. The experimental unit executes the possible cause and produces an effect.
4. The trial is one observation of the experimental unit under treatment **A** or under the alternative **~A** to the new capability to see if effect **B** occurred or not and includes all of the contextual conditions under which the experiment is executed.
5. The analysis phase of the experiment compares the results from one trial to a different trial.

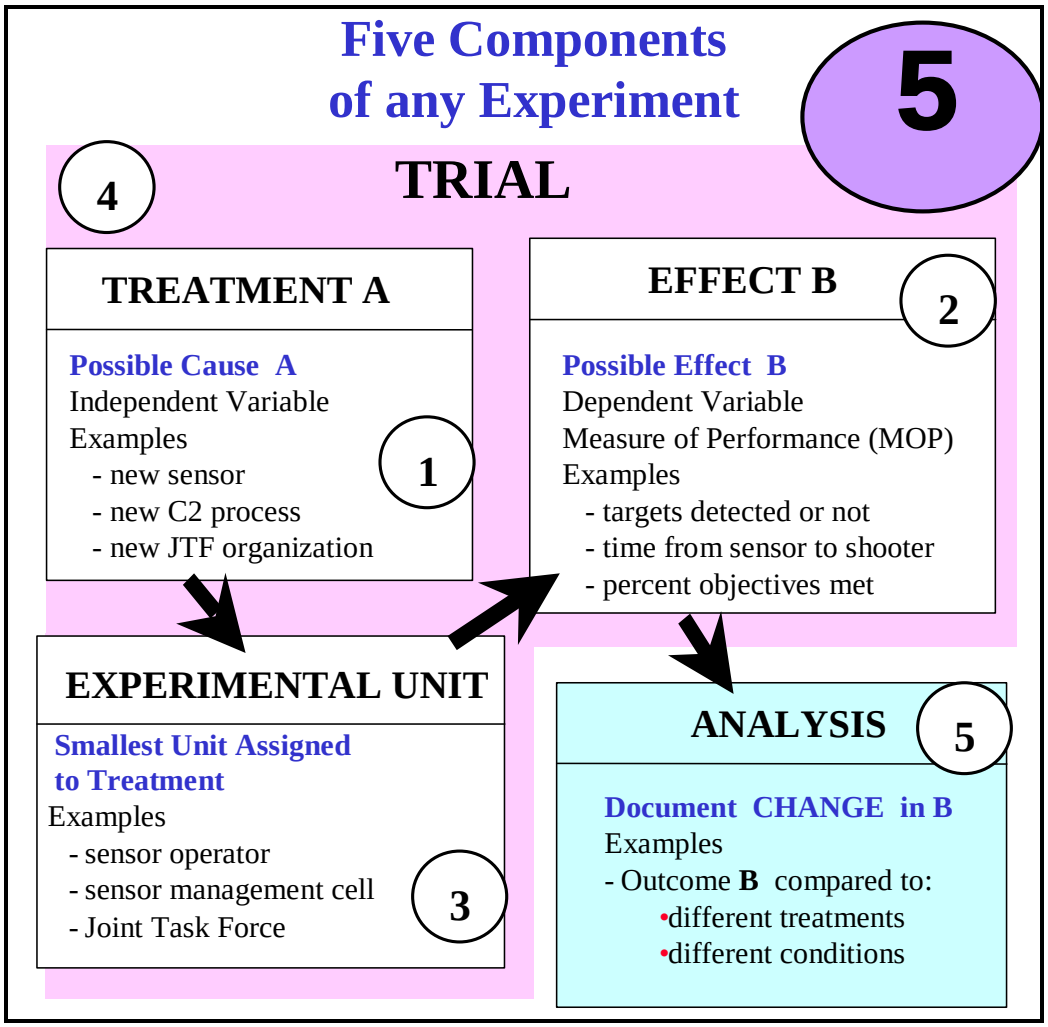


Figure 13 Logical links among four of the five experiment components

There is a strong bond between the first two experiment components and the experiment hypothesis. The experiment treatment **A** represents the left-hand side of the hypothesis as the proposed solution; and the experiment effect **B** represents the right-hand side as the problem to be overcome. Consequently, it is difficult to see how one could think of an experiment without a hypothesis. The arrows of Figure 13 show how one can proceed using the hypothesis as the thread to gain the knowledge sought.

Some field experiments are grand exercises with multiple experimental initiatives (possible causes), sometimes as many as 30 to 50 different initiatives in one experiment. The five components are useful in understanding these large field experiments. These field exercises may be viewed as multiple small experiments inside the overarching experiment. Each individual experimental initiative is configurable as a unique subset of the five components. Each initiative is a separate treatment with its own experimental unit (operators in one area of a command post), its own set of outcome measures, and its own set of trial conditions which may or may not impact the other initiatives in the grand experiment. Moreover, each initiative with its five components will probably have a different number of trials.

2.1.10 Twenty-one (21) Threats to a Good Experiment

How does one design a good experiment? As we have learned, a good experiment among scientists is termed a valid experiment. However, it is too often the case in agencies conducting defense experiments that "Experiment validity is kind of like art—I can't explain it, but I know it when I see it." Questions about experiment validity are often answered by sending first-time experiment designers to the most experienced analyst to receive a list of do's and don'ts and lessons learned. These “good practices” are seldom written and when lessons learned and good practices are written, they tend to be a “laundry list” with little organization or rationale related to the idea of experiment validity. The list of good practices often refers to the importance of sample size, realistic threats, representative units and operators, and so on. Many practical lists exist admonishing what should be done to design good defense experiments. In general, there is much overlap and agreement among various “codes of best practices” for defense experimentation. Good practices in experimentation are collectively known as experiment methodology.

There is a more heuristic method to approach prescriptions for designing valid experiments. The logic of experimentation has identified the four requirements for a good experiment. Building on the work of Cook and Campbell,[16] one can identify the things that can go wrong in an experiment. Cook and Campbell call these threats to validity—identified problem areas that can cause one to not meet any one of the four experiment validity requirements. Experiment good practices then become ways to eliminate, control, or ameliorate the threats to validity. Cook and Campbell’s threats to validity can be distilled down to 21 threats to defense experiments. These threats can be arrayed within a two-dimensional matrix to better understand the actions the experimenter can take to counter these threats. In Figure 14 the 21 threats to validity are arrayed with respect to the four experiment validity requirements and the five experiment components.

All good experiment practices are counters, or antidotes, to the 21 threats to experiment validity. A good experiment plan should show how each of the 21 threats has been accounted for and countered. To help the experimenter, the multitude of good experiment design practices developed over the years to counter each of the 21 threats will be presented and discussed in Principle 3 and summarized in Section 3.5.

[16] While [Cook and Campbell 1979] identified 33 threats to validity, GUIDEx combined and distilled these down to 21 potential threats to defense experiments. Moreover, GUIDEx rearranged their threats into a two-dimensional matrix to better systematically illustrate how the threats to experiment validity can be understood and treated with respect to each of the four requirements and the five experiment components. Additionally, many names of their threats to validity have been changed to reflect military experimentation terminology. For example, *learning effects* is substituted for Cook and Campbell's *maturation*.

4 Exp Requirements
5 Exp Components

21 Threats to a Good Defense Experiment

	1 Ability to Use Capability	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 Results to Operations
1 Treatment	**1 Capability not workable:** Do the hardware and software work?	**5 Capability variability:** Are systems (hardware and software) and use in like trials the same?	**11 Capability changes over time:** Are there system (hardware or software) or process changes during the test?	N/A	**18 Nonrepresentative capability:** Is the experimental surrogate functionally representative?
2 Players	**2 Player non-use:** Do players have the training and TTP to use the capability?	**6 Player variability:** Do individual operators/units in like trials have similar characteristics?	**12 Player changes over time:** Will the player unit change over time?	**15 Player differences:** Are there differences between groups unrelated to the treatment?	**19 Nonrepresentative players:** Is the player unit similar to the intended operational unit?
3 Effect	**3 No potential effect in output:** Is the output sensitive to capability use?	**7 Data collection variability:** Is there a large error variability in the data collection process?	**13 Data collection changes over time:** Are there changes in instrumentation or manual data collection during the experiment?	**16 Data collection differences:** Are there potential data collection differences between treatment groups?	**20 Nonrepresentative measures:** Do the performance measures reflect the desired operational outcome?
4 Trial	**4 Capability not exercised:** Do the scenario and Master Scenario Event List (MSEL) call for capability use?	**8 Trial conditions variability:** Are there uncontrolled or unmonitored changes in trial conditions for like trials? Look for intervening variables not recorded.	**14 Trial conditions change over time:** Are there changes in the trial conditions (such as weather, light, start conditions, and threat) during the experiment?	**17 Trial condition differences:** Are the trial conditions similar for each treatment group?	**21 Nonrepresentative scenario:** Are the Blue, Green, and Red conditions realistic?
5 Analysis	N/A	**9 Low statistical power:** Is the analysis efficient and the sample sufficient? **10 Violation of statistical assumptions:** Are the correct analysis techniques used and error rate reduced?	• **The purpose of an experiment is to verify that A causes B.** • **A valid experiment allows the conclusion, A causes B, to be based on evidence and sound reasoning…** **- by reducing or eliminating the 21 known threats to validity.**		

Figure 14 Mapping of the 21 threats to good defense experiments

The two-dimensional framework of Figure 14 for organizing good experiment practices provides a substantial advantage over the traditional "laundry list" of good practices. It can be used for managing your ammunitions or counter-measures of Principle 3 against threats falling into the 21 threat categories presented.

The framework associates different good practices with each of the four experiment validity requirements. This facilitates understanding why particular good practices are important and the impact on experiment validity if the threat is not properly attended to. As will be discussed later, it is impossible to implement all of the good practices in any particular experiment. Thus, an understanding of the impact of unimplemented good practices is critical to designing the "best available" experiment. Furthermore, associating good practices with the different experiment components allows the experiment designer to see the interaction of good practices across all aspects of the experiment.

2.2 Summary

Understanding the "2, 3, 4, 5, 21" logic of defense experimentation (exemplified by Figure 8 to Figure 15) allows one to see the "big picture." It provides a rationale and road map for sorting through the myriad of details encountered when designing defense experiments. It also provides a straightforward explanation of what defense experiments are all about without requiring a PhD in experiment design. Finally, the logic and resulting two-dimensional framework provides a coherent rationale for organizing experiment lessons learned and good practices as preventable threats to validity to increase the scientific rigor of defense experiments.

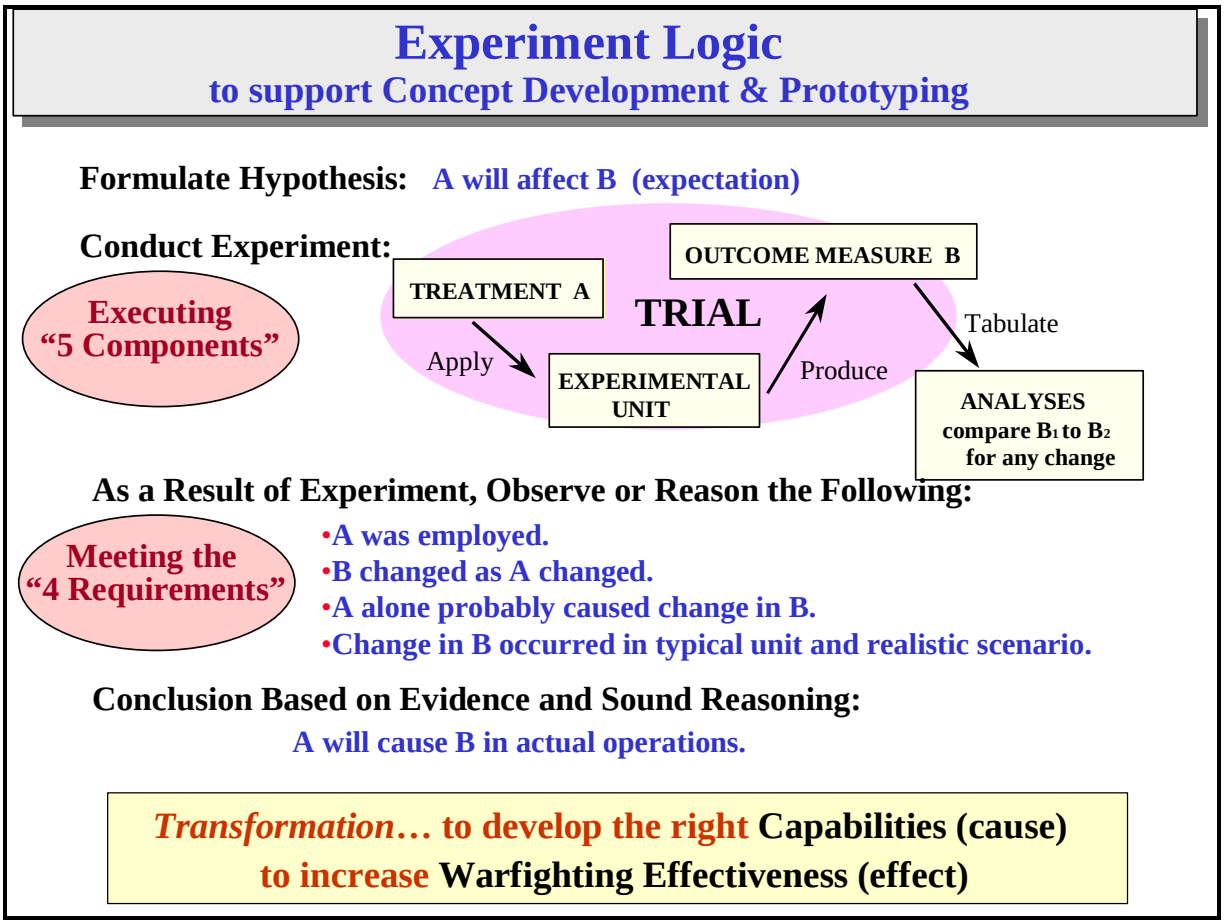


Figure 15 Experiment logic to support joint concept development and prototyping

Defense experiments are essential to developing empirically based concepts and capabilities. New capabilities include the doctrine, organization, training, materiel, leadership, personnel, and facilities that will enable or cause future warfighting effectiveness. Experimentation is the unique scientific method for establishing whether hypothesized concepts are causally related to effects. If the five experiment components are designed to meet the four experiment validity requirements, the defense experiment will provide the concept developer with the basis to proceed. Application of these scientific principles ensures that the new warfighting concept will be empirically related to warfighting effectiveness, thus providing the foundation for transforming military forces.

Principle 3.

Defense experiments should be designed to meet the four validity requirements

Principle 3 discusses GUIDEx recommended experiment techniques to counter the 21 threats to the four experiment validity requirements.

1. *Ability to use the new capability.* Developing and getting the new experimental capability to the experiment is often a major resource commitment. In the ideal experiment, the experimental capability is employed by the experiment players to its optimal potential and allowed to succeed or not succeed on its own merits. Unfortunately, this ideal is rarely achieved.
2. *Ability to detect a change in the effect.* If the unit is able to employ the new capability, the next logical question is whether any noticeable difference is observed during the experiment trial. In the ideal situation, a change in the experiment measure of effectiveness accompanies a transition from the old capability to the new capability. If this does not occur, the concern is too much experimental noise. The ability to detect change is a critical requirement of all experiments.
3. *Ability to isolate the reason for change in the effect.* If the experimenter had a good design where the capability was useable and produced a change in the effect; the question now is whether the detected change was due to the postulated cause, changing from old capability to new, or due to something else. When alternative explanations of experiment results are available, the results are confounded. Scientists have developed experimental techniques to eliminate alternative explanations of the cause of change.
4. *Ability to relate the results to actual operations.* If the unit could employ the capability and the experimenter was successful in detecting change and isolating the cause of change, the question is whether the experimental results are applicable to the operational forces in actual military operations. Experimental results are only useful to the extent they say something about the real world. Generalizability is the scientific term for the ability to apply results outside the experiment context. Ability to relate results pertains to experiment realism and robustness.

All defense experiments are designed to meet these four requirements. However, a 100-percent valid experiment is not achievable. Attempts to satisfy one of the requirements work against satisfying the other three. Precision and control increase the ability to detect change and isolate the cause, but decrease the ability to apply the results to real-world situations because they are exceedingly complex and difficult to track. Experiments designed to detect and identify change emphasize strict control of trial conditions and feature multiple repetitions of similar events. On the other hand, experiments designed to relate results emphasize free-play, uncertainty, and a reactive threat. Consequently, designing experiments is a matter of making well-informed tradeoffs in order to achieve sufficient validity to support the purpose of the experiment.

Principle 3. Defense experiments should be designed to meet the four validity requirements

The reader is encouraged to apply and adapt this Principle. However, the examples herein are based on the specific perspective and experience of different lead-nation authors with contributions from other participants. Therefore they may require supplementary effort to relate them to national perspectives.

Designing defense experiments to meet each of the four experiment validity requirements is an art. This section will discuss the threats to validity associated with each of the four experiment validity requirements (Figure 12). A thorough understanding of the 21 threats and the associated good experiment practices is critical to understand when to apply the good practices and what tradeoffs are required. Tradeoffs are required when designing defense experiments because different good experiment practices often work against one another. For example, one good practice is to have **multiple similar trials,** called replications, to increase statistical rigor. However, constructing similar trials where the Red-players operate the same way in successive trials works against the good practice of ensuring **independent Red-player actions during each trial** to increase realism. A thorough discussion of the tradeoffs among the four requirements will be discussed in Principle 7, and also to some degree in Principles 4 to 6.

The following discussion of the four experiment validity requirements, the threats to validity, and the experiment techniques to address these threats is adapted from [Shadish *et al.* 2002]. Their work serves as the foundation for the following discussion, although several changes to their presentation are introduced here. Much of their original terminology has been translated into military terminology, for example their "maturation effects" is translated as "learning effects" and all examples of good experiment practices are in regards to military experiments. Additionally, the following discussion combines two of their original four requirements (construct validity and external validity) into a single external validity Requirement 4, the ability to relate results. In defense experimentation most effects of interest are straightforward (detections, engagements, *etc*.) and there is far less emphasis on constructs. And finally, the following discussion of Requirement 1, ability to use the capability, is not considered as one of their original four validity requirements. They discuss it as a "special problem" of experiment treatment implementation. It is elevated here as Requirement 1 because it is consistent with the logic of experimentation (the left-hand side of the hypothesis) and because it is such a prevalent problem in defense experiments. Notwithstanding these adaptations to Shadish, Cook, and Campbell's validity framework, the following discussion would not have been possible without their book which culminates 40 years of investigating experiment validity in non-laboratory settings.

3.1 Experiment Validity Requirement 1: Ability to Employ the New Capability

Perhaps some of the most frustrating and, unfortunately, most consistent "lessons learned" from defense experiments are the following:

1. The proposed capability did not work as well as promised.
2. The players did not know how to use the new capability properly.
3. The experiment scenario was not sufficiently sensitive to the new capability. The trial results occurred because of some dominant factor unrelated to the use or non-use of the new capability.
4. The experiment trial did not give the players the opportunity to use the new capability.

These experiment lessons are most frustrating since, in most cases, the majority of pre-experiment resources and effort is expended toward developing and getting the new experimental capability to the experiment. Ensuring that the experimental capabilities can make a difference in the experiment outcome is the first logical step in designing a valid defense experiment. In Case Study 2 as an example, failure to recognize this early in a trial limited the value of the experiment.

The first four threats (Figure 16) to experiment validity, discussed below, indicate the things that can go wrong when attempting to employ a new experimental capability in an experiment.

Threats to the Ability to Use the Capability

THREAT	PREVENTION
Treatment **1. Capability not workable** • Do the HW & SW work?	• Ensure functionality of experimental capability is present.
Unit **2. Player non-use** • Do the players have the training and TTP to use the capability?	• Ensure player is organized, equipped, and trained for capability use. • Provide sufficient doctrine and SOPs for capability use. • Provide sufficient pre-experiment "practice time".
Effect **3. No potential effect in output** • Is the output sensitive to capability use?	• Conduct pilot-test impact on experiment outcome. • Verify model input-output logic.
Trial **4. Capability not exercised** • Do the scenario and MSEL call for capability use?	• Pilot-test the scenario and MSEL. • Prepare White cell specific scenario injects and monitor for use.

Figure 16 Threats related to the ability to use the capability

3.1.1 Threats to Experiment Validity Requirement 1

3.1.1.1 Threat 1. New Capability Does Not Function

The most frequent threat to Requirement 1 is that the experimental hardware or software does not work as advertised. It is well known that the experiment players will attempt to make just about anything work but they cannot overcome primary deficiencies in basic system functionality. One of the major corollaries to this threat in the command, control, and communications area is interoperability. Systems that interoperated in the designer's facility almost surely will not when brought to the experiment. Good experiment practices to alleviate this threat are obvious but challenging nonetheless. The experiment director needs to schedule frequent demonstrations of the new capability's functionality and interoperability prior to the experiment. These demonstrations should include pilot tests in the environment of the experiment with all of the other systems where possible.

3.1.1.2 Threat 2. Experiment Players Cannot Use the New Capability to its Full Extent

The second most prevalent threat to Requirement 1 is that the experiment players are frequently undertrained and not fully familiar with the new capability's functionality. This frequently occurs because the new system is not available for training until the last minute. And on those rare occasions when the system is available, it is not fully functional (Threat 1). Thus, a five-day pre-experiment training period turns into four days of lectures about the system's functionality with hands-on practice with an incomplete system on the last day of scheduled training. Even when the system and its functionality are available, new equipment training tends to focus on operator skills rather than employment skills because the tactics, techniques, and procedures (TTPs) for optimum employment are non-existent or immature. Too often the TTPs are developed by the experimental unit during the early experiment trials. Similarly, for new and complex staff-support systems the standard operating procedures (SOPs) are not developed. So while the operators may be trained on their operational role with the new processes, the procedures for receiving inputs and providing and incorporating the outputs of a new process will falter.

Once again the good practices are obvious, especially in the military where training is an integral aspect of the everyday mission. The key is to anticipate the problems identified above and provide sufficient "practice time" for players to be able to operate and optimally employ the system. This means that not only does the new functionality and interoperability need to be available and thoroughly tested prior to experimental unit training, but also that the TTPs and SOPs have to be developed concurrently with the new capability development; no easy task when developing operational procedures for a capability that does not yet exist in its final form.

3.1.1.3 Threat 3. New Capability Cannot Impact Experiment Outcome

While the previous two threats are generally acknowledged and the associated good practices are well established, Threat 3 often falls below the horizon. Threat 3 identifies the need to ask oneself: "If this system is used to its fullest extent, will it make a

noticeable difference in the experiment?" Is the experiment environment sensitive to its potential impact? Several good practices ameliorate this threat.

Pilot tests, full-dress rehearsals[17], prior to the start of experiment trials not only provide a check on Threats 1 and 2, but are also the best way to counter Threat 3. The experimenter should examine the experiment environment to see if it is structured to give the new capability a fair chance to demonstrate its advertised strengths. If the experiment is to be a comparison between the old and new capability, it is critical to include the old capability in the pilot test. It is always a good idea to structure some experiment trials where it is expected that the old system may perform equivalent to the new capability and experiment trials where the advantages of the new capability should allow it to excel. Both of these trials should be examined during the pilot test to test these assumptions. If one does not see indications of performance differences between the old and new capability during the pilot test, this should be a strong indication to re-examine the ability of the trial scenario to show a difference.

If the experiment is to examine various levels of the capability (or the same capability under distinct conditions), by design increase the differential between the various levels or the distinct conditions in order to increase the chance of seeing differences in experiment outcomes.

When the primary action and results during the experiment trial action occur within a simulation, the sensitivity of the simulation to differences between the old and new capability should be part of the simulation validation and accreditation effort. New experimental capability such as new sensors, new transporters, or new weapons that are to be simulated can be rigorously tested in simulation prior to the experiment itself. Pre-experiment simulation of the old and new capabilities can also serve to identify trial scenario conditions that will accentuate similarities and differences as discussed above.

3.1.1.4 Threat 4. New Capability Not Employed During Trial

This is the most unfortunate threat in this group. After great effort to counter the first three threats, *e.g.*, getting a fully functional capability on time, providing adequate operator and employment training, and ensuring that the new capability would make a difference; it would be unfortunate if the new capability never had a chance to be employed during the experiment trials. This occurs when the new capability is not the primary focus of the event. This most often occurs when conducting embedded experiments within large operational exercises or training exercises; or when conducting a small "side-experiment" within a larger experiment involving a major operation.

Good practices for preventing Threat 4 include developing a detailed master scenario event list (MSEL) that lists all scenario **injects or events** that are to occur over the course of the experiment trial. Pre-planned scenario events are specifically developed to **drive** the experiment players to deal with specific situations that allow for or **mandate** the use of the new capability. The experimenter continually monitors the trial and

[17] Also known as a "dry run" in some communities.

ensures that all the MSEL events occur. The experimenter should also monitor the experiment players to see if they reacted accordingly to the scenario events. If the players did not attempt to employ the new capability when the MSEL event occurred, was it because they did not **see** the event? This situation needs to be avoided. In order for the new capability to rise or fall on its own merit, it must be employed.

3.1.2 Summary

Good practices associated with the above four threats are not new. They are paradoxical, most obvious but most frequently violated, thereby engendering the most frequently expressed lessons learned in past defense experiments. Why is this so? First, the schedule for defense experiments is fixed to a particular calendar "window" because operational forces need long lead times to commit to participation. New capabilities, however, involving innovative software or hardware configurations seldom meet their optimistic development schedules. As a result, the experimenter is faced with a dilemma: either execute the experiment during the pre-planned window with the capability functionality "as-is" or skip the experiment altogether; because the operational resources, notably the experimental unit and "range time," will elapse at the end of the experiment window. Second, insufficient time is allocated during the experiment window for player training on the new capability and scenario rehearsals because experiment-window durations are minimized to reduce the impact on scarce operational resources. Understanding Threats 1 through 4 and their impact on validity Requirement 1 is the first step in the ability to apply the good practices listed above.

3.2 Experiment Validity Requirement 2: Ability to Detect Change

3.2.1 The Importance of Change

As previously discussed, the most basic paradigm of an experiment is "doing something and seeing what happens." This section focuses on the "**seeing what happens**" and is appropriately titled "detecting change." Detecting change is reflected in observing or measuring an increase or decrease in the "effect" variable after each experiment trial. In defense experiments the experimental effect is called the measure of performance (MoP) or measure of effectiveness (MoE). For the discussion in this section the MoP or MoE will be simply referred to as the effect.

There is a logical order to the four experiment validity requirements. The ability to detect change in the effect from the trial with the new capability when compared to the trial with the old capability is the second logical requirement for a good experiment. If Requirement 1 was not met and the new capability was either not successfully employed or the scenario was not sensitive to its use, then there is no reason to expect that the new capability would produce a change in the trial outcome. Similarly, we will see that if the experiment did not produce an observable difference in the effect variable, then it does not make sense to discuss Requirement 3 (the cause of the change) nor to discuss Requirement 4 (the implications of change to a wider context). Therefore, the ability to detect change is the critical second logical requirement.

3.2.2 Detecting Change is Observing Covariation

The ability to detect change in the effect is concerned with detecting covariation, that is detecting a pattern of change between the treatment **A** and the effect **B**. Covariation occurs when the size of the effect systematically varies with different applications of the treatment: **A** and **~A** are the new sensor and the current sensor respectively[18]. A pictorial representation of covariation is presented as the Experiment X example in Figure 17. If observations of experiment effects (such as targets destroyed, times to detect, amount of supplies delivered) fluctuated widely from trial to trial, then no clear covariation will be discernible (Experiment Y in Figure 17). Clear covariation represents a high signal-to-noise ratio and presents a discernable pattern between the treatment and effect. A low signal-to-noise ratio presents a difficulty in seeing a pattern of covariation within the experiment noise. The ability to detect trial-to-trial changes is called statistical validity. The ability to draw statistical conclusions from an experiment is the ability to detect covariation between different levels of the treatment and the effect. The ability to detect change is statistical power.

[18] A and ~A (A and not A), denote using a new capability and not using it respectively.

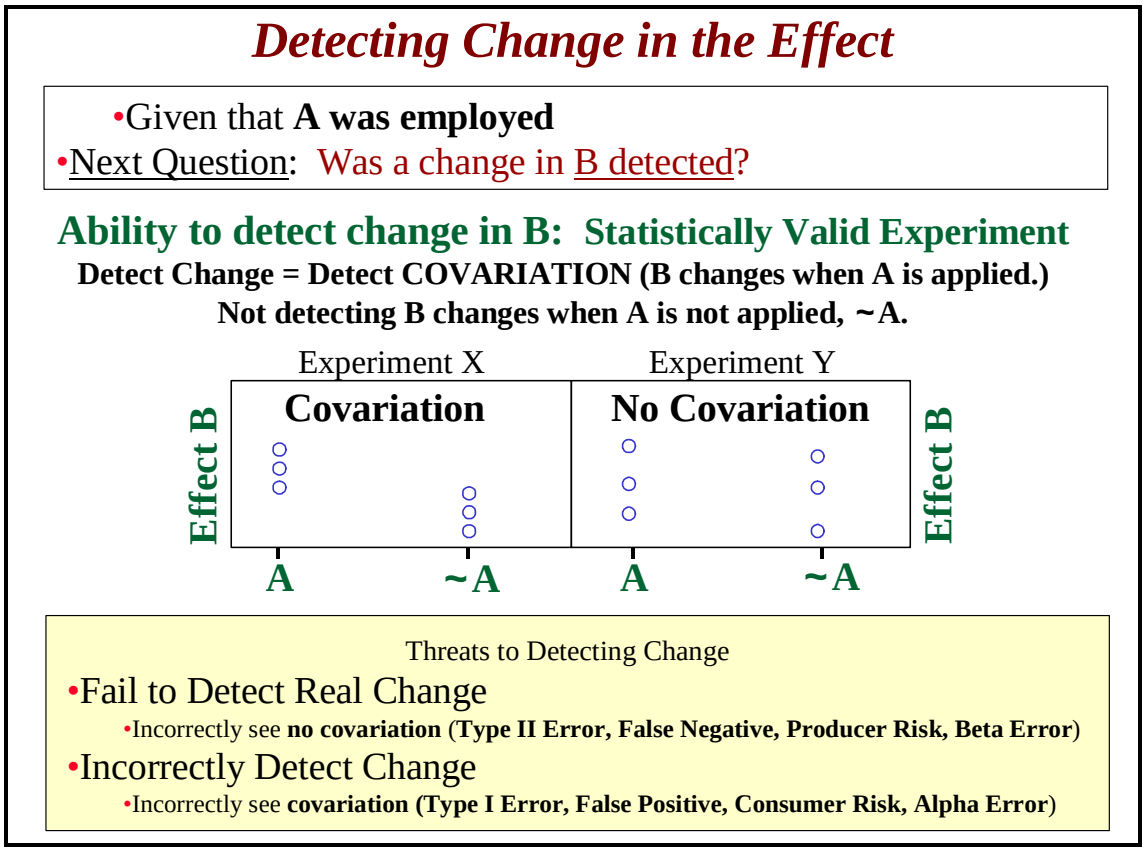


Figure 17 Detecting change in the effect B

Two different mistakes can be made when deciding if change was detected or not. The first mistake is *not detecting real change.* Experimenters mistakenly conclude that **A** and **B** do not covary; when, in reality, they do. That is, they see the no-covariation "Experiment Y" in the computer printout of the data, but "Experiment X" covariation is what really occurred. In statistics this error is referred to as a "*Type II error."* This error is examined first because most defense experiments often have a low signal-to-noise ratio when attempting to measure effects of experimental capabilities in complex military operations in realistic environments. It is often difficult to see a dramatic difference when the new capability is introduced. Therefore, threats to the ability to detect a difference in an experiment are discussed first.

The second mistake is *incorrectly detecting change*. This error occurs when experimenters mistakenly conclude that covariation exists between the treatment and the effect; when, in reality, it does not. This is akin to seeing "Experiment X" (Figure 17) covariation in the computer printout of the data whereas "Experiment Y" no-covariation is what really happened. In statistics this is called a *Type I error* and is discussed second because it pertains to more technical issues of statistical assumptions and error rates.

The six threats to detecting change can be grouped according to whether they increase the risk of the first or second type of error (Figure 18). The five threats to the ability to see real change because of too much noise will be discussed first.

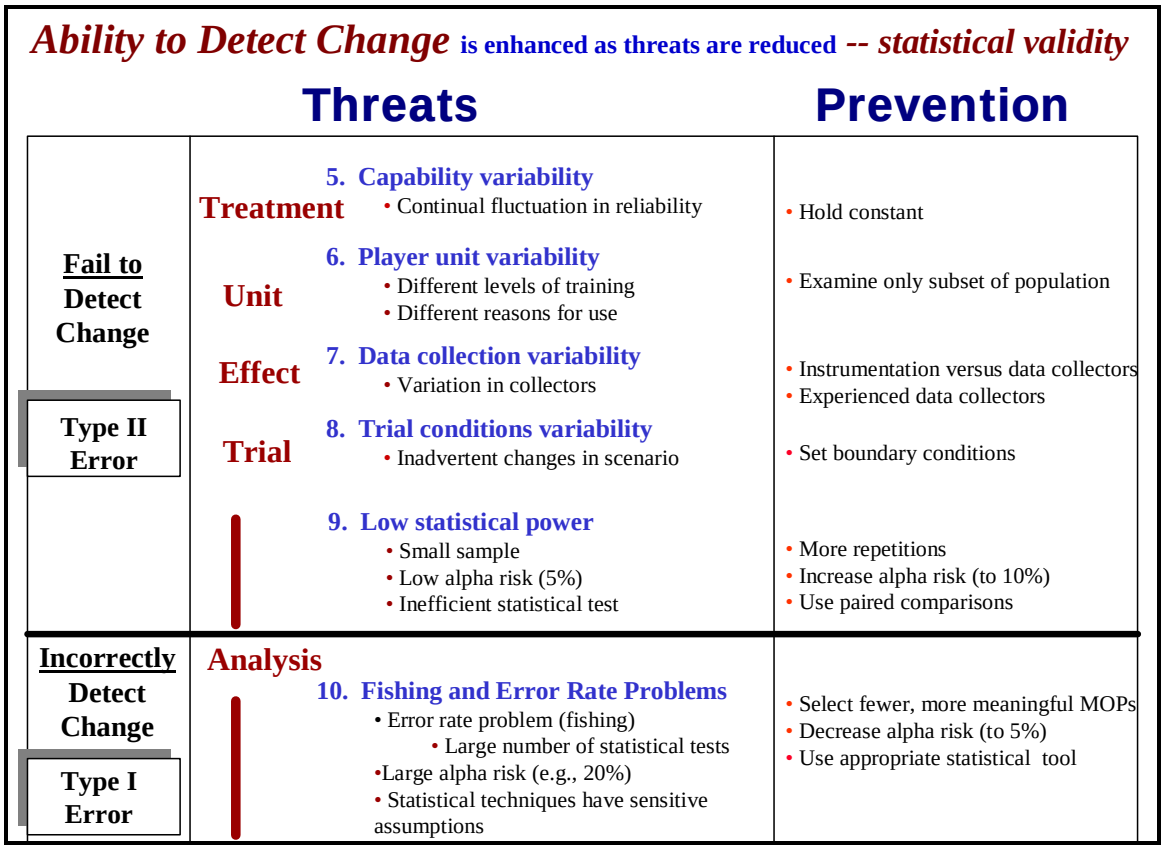


Figure 18 Threats to detecting experiment change

3.2.3 Not Detecting Real Change

This problem arises when experimenters incorrectly conclude that a treatment is ineffective. As an example, suppose that in actual military operations a new sensor system (treatment variable) would produce quite a few more detections; but the experiment did not produce a discernable increase in effectiveness for the new sensor and the experimenter incorrectly concluded that there was insufficient "goodness" or added value for purpose in the new sensor. There was too much noise in the experiment to see the correct signal. The real change was buried in experiment clutter. The ability of defense experiments to produce discernible results is technically referred to as statistical power. The following five sources of experiment noise are the five Type II threats to detecting change.

3.2.3.1 Threat 5. Capability Variability

Noise from capability variability arises from two different situations. The first instance of experiment noise occurs when a capability system has to operate continuously over the course of a lengthy trial. This arises when systems such as communication systems, sensors, and data systems need to be operational continuously and functioning at a constant level over the course of many hours. It is an important good practice to maintain the consistency of an experimental capability during the entire trial. Maintaining the constancy of the capability for long durations is not always easy.

Prototype systems are often unreliable and may stop functioning during a trial. They may also undergo unplanned hardware, software, or training modifications during long trials. This random capability variation within a trial will diffuse the effectiveness of the treatment making it difficult to detect a true change from trial to trial.

Good practices include providing sufficient pre-experiment operating time for immature new technology to ensure it will work consistently for the duration of an experiment trial. For an immature unreliable system, incorporate an experiment-fix-experiment methodology by designing a series of short experiment trials with treatment fixes occurring between trials rather than incorporating capability fixes (changes) during one long experiment trial. In this manner, the capability is held constant during each trial but allowed to improve from trial to trial in a systematic fashion. This experiment-fix-experiment approach now has multiple, sequential capability levels that can be examined separately.

The second instance occurs in experiments where multiple versions of the capability are employed simultaneously within a single trial; for example, giving all members of a platoon a handheld radio to see if that improves overall platoon performance. If each handheld radio functions erratically, any true platoon improvement "signal" may be obscured by the variable performance "noise" within a trial. A good practice is to calibrate all experiment articles for consistency prior to pilot testing so that the new capability is held constant within the experiment trial. Use the pilot test to ensure all copies of the new capability function equivalently. After the experiment the experimenter can assess the extent of capability variability by comparing individual scores across items. When variability is a result of a few outlier cases, the experiment analysis can be performed with and without outliers to determine their impact on the results.

3.2.3.2 Threat 6. Player Variability

Noise from player variability also arises, and certainly occurs, in experiments where multiple individuals or multiple teams of individuals are used to obtain multiple observations (replications) of one treatment condition: for example, using four different side-by-side gun crews to test the accuracy of a new gun at 1000 meters. Non-standardization among different operators, crews, or units increases error variance. Non-standardization occurs when each operator or each team has a different level of training, a different experience level, or different motivation to participate.

It is always best to deal with this threat prior to the experiment. Good practices include increasing standardization across experiment teams prior to the experiment. Standardization among experiment teams can be improved by training everyone to the same level of performance prior to start of the trial. When possible, select similar (homogeneous) players to participate in the experiment to reduce player variability. However, this will compromise Requirement 4, external validity.

After the experiment the experimenter can assess the extent of standardization by comparing individual scores across player teams completing the trial. Variability in these scores can sometimes be statistically corrected using covariance analysis with pre-

experiment training and experience scores. Alternatively, when there are only a few outlier cases, they can be statistically identified and the analysis performed with and without outliers to determine the impact of outliers on the conclusions. The post-experiment statistical corrections are always risky due to the statistical assumptions that accompany them.

3.2.3.3 Threat 7. Data Collection Variability

Many different data collection techniques are available to measure effects in defense experiments. Data collection devices include elaborate instrumentation tapping directly into system data busses; and not so elaborate procedures, such as data collectors, questionnaires, and observations from technically proficient observers, referred to as SMEs. Inconsistencies in any collection device will obscure true change within measurement variance.

Reliable measurement is the principal good practice for countering Threat 7. A reliable measure provides consistent output for a particular stimulus. Data collection measures have been divided into two categories: objective and subjective. Objective measures mean "without human judgment" and include instruments such as laser receivers, electric in-line counters, cameras, software logger algorithms, and so on. Subjective measures, on the other hand, signify "with human judgment" and include player surveys, data collectors to record visual and acoustical events, and subject–matter expert (SME) observers to record and infer why an event happened or evaluate the "goodness" of actions.

It is incorrect to assume that all objective measures are inherently reliable (consistent) and all subjective measures are unreliable (inconsistent). All data collection instruments need to be calibrated to ensure their continued consistency throughout the experiment. It is always a good practice to pretest and calibrate electronic data collection instrumentation to verify consistency.

A good experiment practice is to use objective measures whenever possible. Objective data collection instruments still need to be calibrated. These measuring devices can be calibrated to familiar metrics. For example, a timestamp recorder may be "certified" to vary by no more than plus or minus two seconds.

The techniques for calibrating the consistency of player surveys and human data collectors is less understood but procedures for doing so exist [Kass 1984]. Calibration surveys and data collectors "objectify" traditional subjective measures. Subjective measure still retains human judgment but the human judgment can be made more consistent. Calibrating the consistency of player surveys is called "item analysis." A player questionnaire intended to measure the adequacy or quality of a process or product can be calibrated with respect to consistency. That is, the extent that two individuals with similar opinions will result in similar scores on the questionnaire. This is the meaning of objective measurement of personal (subjective) opinion. Commercial software programs such as SPSS and SAS provide routines that analyze individual questions (items) in surveys to determine their internal consistency with other related items in the survey. In general, increasing the number of related questions about a

particular judgment in a questionnaire increases the reliability of player survey judgments. In this manner questionnaire scales can be calibrated to quantifiable consistency indices, *e.g.*, .85 internal consistency reliability. Using multiple questionnaire items to assess a player response and calibrating these items using item analysis is a good practice for increasing the objectivity of player surveys.

Similarly, the consistency of data collectors can be calibrated by comparing their observations across similar and dissimilar events during training. Data-collector subjective assessment consistency can be enhanced by having individual data collectors to provide multiple component ratings of a single event; for example, rating both the completeness and usefulness of a report. The component assessments are then combined to produce an overall "adequacy" score.

Additionally, the assessments from two side-by-side data collectors providing independent assessments can be combined and averaged to provide a more consistent assessment for the trial event. Averaging component scores of a single collector or averaging across multiple collectors increases the reliability of subjective assessments. Training data collectors to provide consistent responses and averaging across data collector responses are good practices for increasing the consistency (objectifying) of subjective data collector ratings.

3.2.3.4 Threat 8. Trial Conditions Variability

The prevalence in the experiment of uncontrolled variables that impact the effectiveness of the treatment during a trial will artificially increase or decrease the size of the effect for that trial. This unwanted variation may obscure the real difference between trials.

A player unit that experiences different levels of temperature, weather, light conditions, terrain, and threat levels in successive trials will fluctuate in performance during the trial and this **noise** will obscure any potential effect **signal** when compared to another trial. While military robustness may dictate that a useful experimental capability should be able to stand out under any variation in the military environments, many early capabilities may be found to be effective in some, but not all conditions. If all conditions are allowed to impact randomly, the capability potential for high effectiveness in some particular conditions may be obscured in the average.

Early in an experimental campaign, a good practice is to reduce the number of uncontrolled variables to determine under what conditions, if any, an effect can be detected. Additionally, a signal is more likely to be detected in an experiment with a number of shorter trials with constant conditions rather than one long trial having a wide variety of conditions. In such cases, changes should occur between trials rather than within trials.

When constant trials are not achievable (or desirable) and the sources of the differences (variability) between trials can be identified, some reduction in the variance can be accomplished by using statistical designs such as paired comparisons, matching, within-subjects designs, blocking designs, and ANCOVA (analysis of covariance). Each

of these statistical techniques can reduce the size of the error term, thus making the signal (treatment effect) to noise (error variation) larger and more likely to produce a statistically significant result. However, there is a tradeoff in that each of these techniques also decreases the degrees of freedom associated with the denominator of the error term. Thus, these techniques only reduce noise when the reduction in the error variation in the numerator is not offset by reduction of degrees of freedom in the denominator. These techniques work best when the matching, blocking, and covariate variables are highly correlated with the effect.

3.2.3.5 Threat 9. Low Statistical Analysis Power

The risk of failing to detect a real change is known as a Type II error. There are three ways to inefficiently employ statistical analysis that would jeopardize the ability to observe a real change brought on by employment of the new capability. The good practices associated with each of these problems are as follows.

Inadequate Sample Size. There are available techniques for estimating sample size requirements to achieve specific levels of statistical power. The ability of an experiment to detect an effect of some postulated magnitude is known as the power of an experiment. In general, the larger the sample size, the greater the statistical power. While sample size is most often the main consideration for determining statistical power, it is not the only contributor.

Setting Type I Risk Too Low. There is a direct correlation between Type I risk (discussed next) and the current Type II risk problem. If the experimenter focuses solely on preventing the Type I error to prevent seeing a positive result that is solely due to chance, the experimenter runs the risk of creating too stringent a condition that will not allow a small positive result to show up as statistically significant. Allowing a higher Type I risk (accepting more risk by using a risk level of 5 percent rather than 1 percent) correspondingly reduces the Type II risk, thereby increasing the power of the statistical technique. When setting the Type I and II risk levels for statistical analysis, experimenters need to consider the consequences of each.

Inefficient Statistical Techniques. Statistical techniques differ with respect to statistical power. T-tests of paired comparisons have more statistical power than t-tests of independent observations. Parametric techniques are generally more powerful than nonparametric techniques but are more demanding on hypotheses related to input and output variables, or the knowledge required from previous experiments.

3.2.4 Incorrectly Detecting Change

In statistics, a Type I risk is the possibility of incorrectly concluding that **A** and **B** covary leading to the incorrect conclusion that an experiment treatment is associated with a positive result. If the previous Type II threats are the problem of being too conservative, this Type I threat can be characterized as the problem of “too liberal” interpretation of results. It is easier to make this mistake, when a small change in the effect is detected. For example, suppose in a sensor experiment the average effect for the new sensor was 4.6 detections while the current sensor achieved 4.4 detections. Is

this small change an indication of a true difference between capabilities or is this difference due to chance? Of course, the easiest way to incorrectly conclude that a small positive result reflects a true difference in capability is to not conduct statistical analysis of the data.

It is a natural human tendency after conducting an event a small number of times (say three times, *i.e.*, three trials) and observing a positive result two out of three times, to conclude the experimental system is better. However, we know that flipping a coin three times can result in two heads even though heads and tails are equally likely. Computing statistical analysis of experiment data and getting "statistically significant results" indicates that the observed positive result did not occur by chance (as can be found when flipping a fair coin a few times and getting more heads). All experiment results should be subjected to statistical analysis before drawing conclusions about whether the observed change resulted from chance variation or from a difference in treatment capabilities. When conducting statistical analysis, however, the following two threats need to be considered to ensure that the analysis technique itself does not produce the false positive conclusion that the statistical analysis is designed to guard against.

3.2.4.1 **Threat 10.** ***Fishing and Error Rate Problems***

The likelihood of incorrectly detecting a false change increases as the number of statistical comparisons in a single experiment increases. This is relevant when collecting data on many different measures in one experiment; for example, detection times, detection ranges, detection rates, and so on. Binomial probabilities can be used to estimate experiment-wide error. If data for four different measures (k=4) are collected, and each is independent and analyzed in a statistical hypothesis at the 95% confidence level (alpha=.05), then there is only a 81% confidence $[(1\text{-alpha})^k=(1\text{-}.05)^4=.81]$, rather than a 95% confidence, that all four hypotheses will be true. In other words, there is a 19% probability that at least one of the four individual comparisons will erroneously be accepted as positive (incorrectly concluding **A** and **B** covary). A 19% chance of an erroneous conclusion is much higher than the advertised 5% probability. One way to decrease the multiple-comparison error rate is to increase the confidence level for the individual comparisons. A Bonferroni correction is obtained by dividing the desired alpha level by the number of planned statistical comparisons; in this example .05/4=.0125. A conservative alpha level of .0125 instead of .05 for each of the four individual comparisons would increase the overall confidence level for four comparisons from 81% to 95% $[(1\text{-}.0125)^4 =.951]$. Note that the sample size requirement to achieve a 98.75% confidence instead of a 95% confidence can be great. An alternative to correcting for multiple independent comparisons is to conduct a multivariate analysis of variance (MANOVA). It should be remembered, however, any correction for multiple comparisons is conservative and thus makes it more difficult to detect an important small change.

Statistical analysis of data requires that certain assumptions be met to correctly assess hypotheses at a specified risk level. Violating assumptions of statistical tests increases

the risk of a Type I error; although sometimes it can also increase the risk of a Type II error. Not all assumptions are equally important. Analysis of variance (ANOVA) is fairly insensitive to departures from assumptions of normality or equal within-cell variances. Analysis of covariance (ANCOVA), on the other hand, is quite sensitive to its requirement for homogeneous within-group regression slopes. Nonparametric techniques require fewer assumptions than parametric statistics concerning the level of measurement and underlying distribution. During the experiment design stage, evaluating whether field data will meet the assumptions of the planned statistical analysis is based on experimenters' experience with similar type data. After data collection, most assumptions for use of a particular statistical technique can be assessed empirically.

3.2.5 Increasing Experiment Detectability

As Figure 18 indicates, threats to detecting change in the effect arise in all five elements of an experiment. Many experimenters focus on sample size as the key, but from the same figure it can be seen that sample size is only a component of low statistical power and that statistical power is only one threat to experiment variability affecting the ability to detect a real change (Type II error). The good news is that all five of the Type II threats (Threats 5 through 9) can be ameliorated to some extent as discussed above. The key is reducing variability in the experiment execution. There are statistical techniques for estimating the probability of detecting a change of a certain magnitude in a specific effect. This technique is known as "power analysis." The ability of an experiment to detect an effect of some postulated magnitude is known as the power of an experiment. Analysis of experiment power before data collection takes the form of estimating sample sizes needed for statistical comparisons. After data collection, the experimenter can assess the amount of statistical power the defense experiment actually provided.

3.3 Experiment Validity Requirement 3: Ability to Isolate the Reason for Change

3.3.1 Two General Types of Experiment

After the experimenter has reasonable assurance that the new capability will be employed and the experiment is designed to detect a change in the effect if it occurs, the next logical question is whether an observed result **B** is caused by the new capability **A** or is a result of some other influence **C**. For example, suppose the player unit with the new system was more experienced than the unit with the current system at the start of the experiment. The experimenter could not conclude that an increase in performance by the new-system unit over the current-system unit was the result of the new system. The difference may have been a result of the player unit with the new-system beginning the experiment with more experience. Ability to identify the correct cause of any observed change is termed design validity. Threats to design validity are often referred to as problems of *confounding* (Figure 19). Confounded results are

experiment results that may be attributed to a number of alternative, plausible explanations. Confounded results mean that the reason for any observed change in effectiveness cannot be isolated to the intended cause, the new capability. An experiment high in design validity has eliminated or reduced the potential for alternative explanations to observed changes so that the only remaining explanation is the new capability.

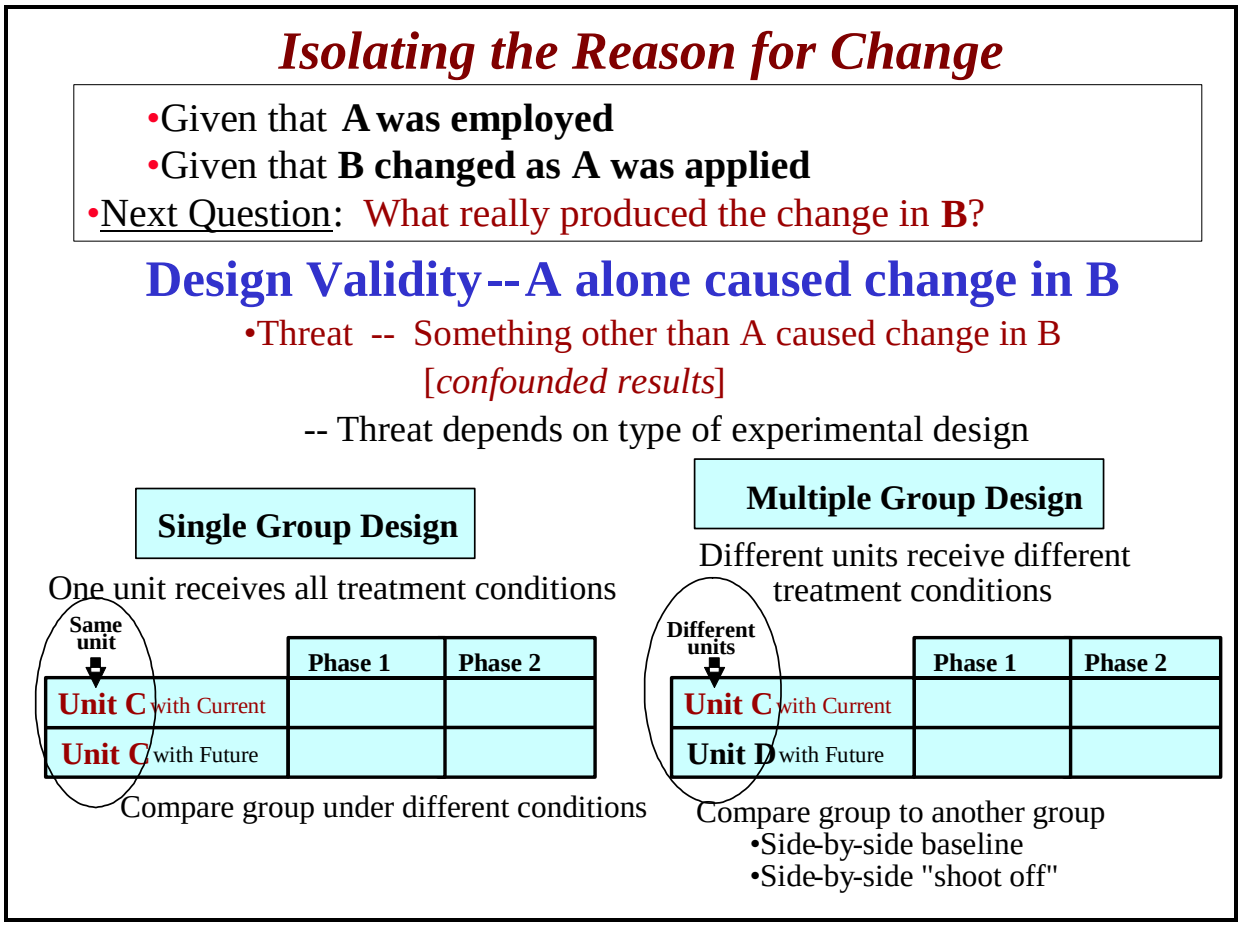


Figure 19 Isolating the reason for change

Threats to the ability to isolate the cause of change can be classified into two different groups: threats affecting single-group experiments and threats affecting multiple-group experiments. Defense experiments can be categorized as either a single- or multiple-group design. There are two types of single-group designs. A single-group experiment comparing the old capability to the new capability will have one player unit use the old capability and then use the new capability in a similar scenario. A second single-unit experiment occurs when there is no comparison to the old capability. A single player unit is trained with the new system and conducts operations with it during the experiment under multiple conditions. In multiple-group designs, on the other hand, at least two different player units are involved in the experiment, each player unit assigned to different treatment conditions. Multiple-group designs are employed when a second player unit operates an alternative system in a side-by-side comparison

experiment. If this alternative system represents the current baseline system, then the second player unit is the control group.

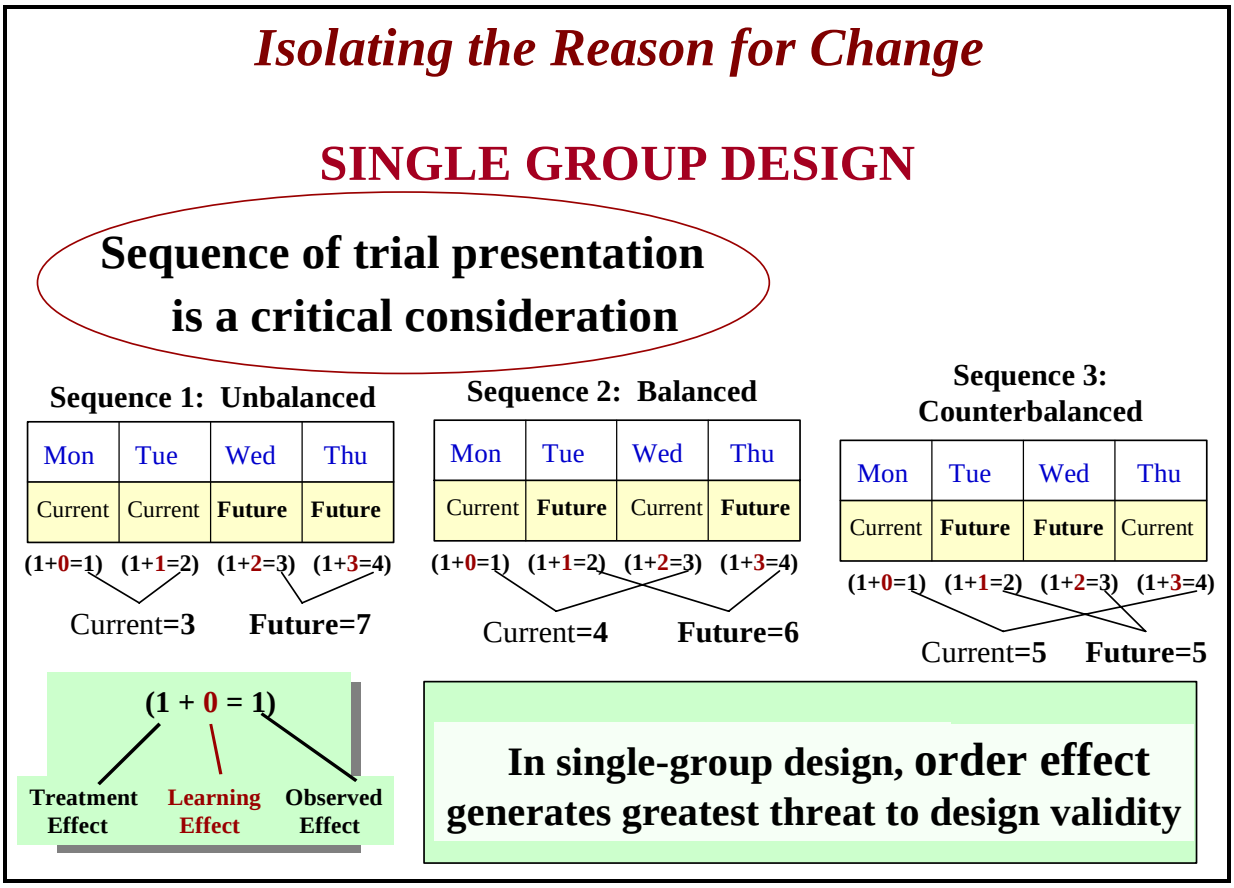


Figure 20 Sequence problem in single-group designs

3.3.2 Single-Group Experiment Causality Determination Problems

The Achilles heel of single-group designs is the problem of order effects. Problems arise when attempting to compare early trials to later trials. Trial order distorts comparisons between trial conditions. A simplified pictorial model can illustrate this problem. In Figure 20 three potential ways to order a sequence of trials are provided as Sequence 1, 2, and 3. The three numbers below each trial quantify the treatment effect, order effect (learning effect), and observed effect. We can hold the treatment effect **A** constant for each trial by giving the treatment effect a quantity of 1 for each trial to see the impact of the trial sequence on what we observe. By giving the treatment effect a consistent quantity of 1 for each trial, we are saying that the treatment (a new sensor system) had the same effect regardless of the condition under which it was operated. Consequently, any differences in the observed trial effect **B** resulted from some other factor **C**.

In this simple example, Factor C is called a learning effect. In Sequence 1 and 2, the observed increase for the current sensor performance is solely the result of the learning effects. Increase in player task proficiency as a result of their experience from one trial

to the next is reflected in the increase of 0, 1, 2, and 3. One method to reduce this order effect is to use a counterbalanced sequence as illustrated in Sequence 3. Order effects need to be closely monitored in experiments because trials are often sequenced to accommodate resource availability rather than experimental design considerations. For example, battlefield smoke trials are usually conducted close together (early or late in the sequence) to coincide with the availability of smoke generators. The following four threats occur when a player unit undergoes experiment conditions in some sequence or order (Figure 21). Sequence threats occur when the experimental unit includes real operators because humans have memories and learn from experience. Computer experiments employing virtual operators are not plagued the same way by these sequence threats because computer players are often memoryless.

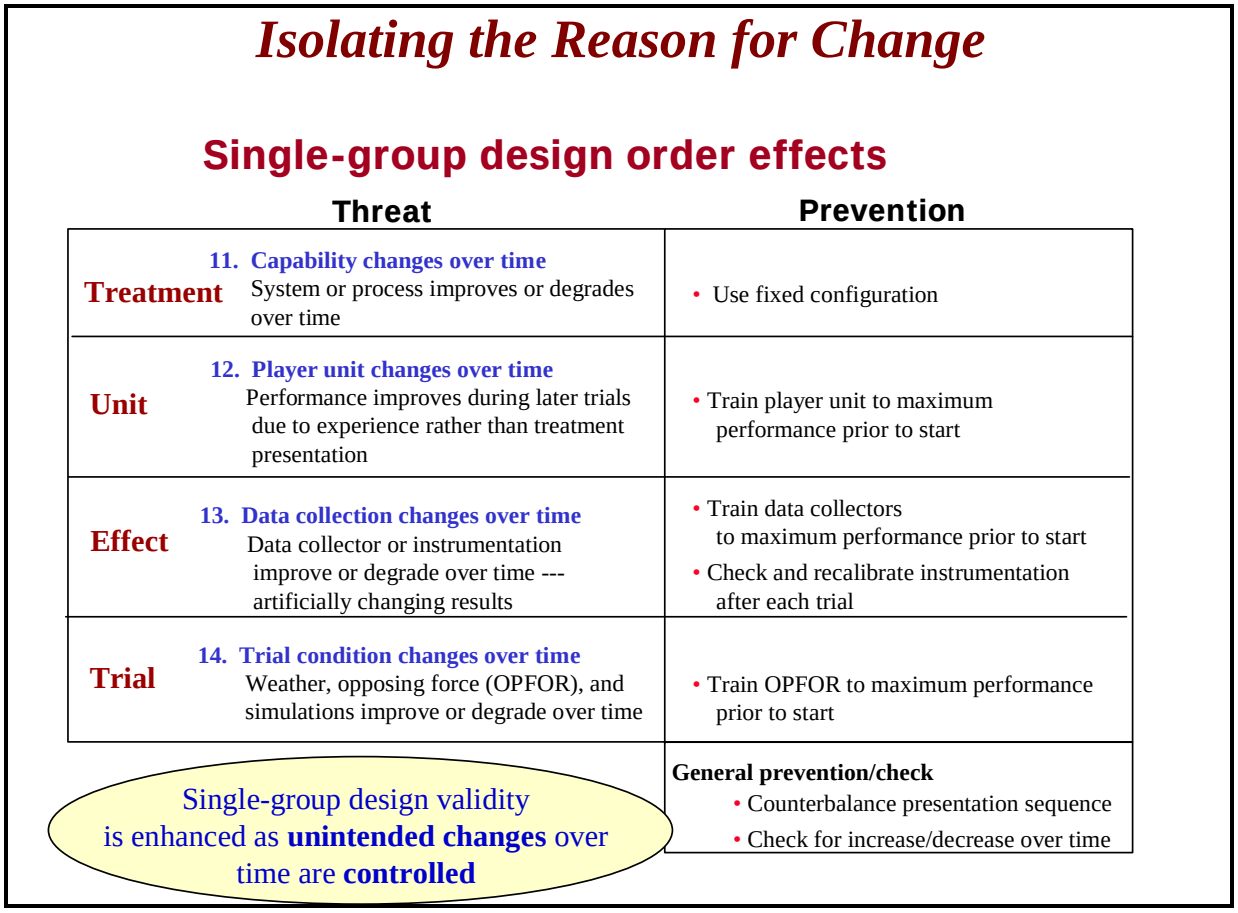


Figure 21 Isolating the reason for change for single-group design order effects

3.3.2.1 Threat 11. New Capability Changes from Trial to Trial

In single-group experiments the functionality of the capability (new system, new process, or new organization) needs to remain constant over time across the different design factors in order to assess whether the new capability is equally effective under different trial conditions that occur later in time. If an intended level of the capability increases or decreases over the course of a single-group experiment that conducts

different treatment conditions in subsequent time periods, then it will be difficult to disentangle the true cause of any detected change.

The primary good practice to prevent Threat 11 is to allow sufficient time in the pilot test prior to the experiment to ensure the stability of the new-capability functionality for the duration of the experiment. During the experiment, continually monitor the functionality to ensure that the "inherent capability" of a treatment does not change during the course of different experiment trials. Monitoring for changes in the treatment, counterbalancing trial sequences when possible, and checking for any increases or decreases in performance over time across successive trials are generally good techniques for reducing this threat.

Sometimes new capabilities, especially experimental future systems, undergo major modifications during a field experiment in order to correct discovered deficiencies in their functionality. These may be hardware, software, or training modifications. An experiment-fix-experiment design encourages and incorporates these modifications. Furthermore the earlier discussion on methods also allows for an alternative fix-experiment-fix design approach. The key question is whether earlier trials conducted prior to the modification need to be rerun in order to make a comparison to the post-fix trials.

3.3.2.2 Threat 12. Experiment Players Change from Trial to Trial

Soldiers, airmen, seamen, and marines participating in field experiments will change during the exercise. If the change is one of maturation, players become more experienced and proficient. This is referred to as a learning effect. If the change is one of degradation, players become fatigued, bored, or less motivated. Player changes over time will produce an increase or decrease in performance in later trials and this change in performance is unrelated to the change in designed treatment conditions. This makes deciphering the real causality of change difficult.

To reduce this threat, good practices such as counterbalanced techniques, as illustrated in Figure 20 should be used when possible. Also, ensure that player units are trained to maximum performance and operate at a steady state. After the experiment is over, check for increasing or decreasing trends over the temporal sequence of trials.

Since the "learning effect" dominates defense experiments (experiment players generally becoming more proficient as the experiment proceeds), the best technique is to counterbalance the sequence (as previously shown in Sequence 3, Figure 20) specifically for this effect. When counterbalancing is not possible, a good practice to **counteract** the learning effect is to conduct the future new-capability trial **first** and the current-capability trial **last**. Any observed improvement for the new capability when compared to the current capability, has "overcome" any learning effects. The experimenter has deliberately biased the sequence of trials so that "learning effects" favor the baseline system. As a result, any performance improvements for the future system can be credibly attributed to the inherent capability of the future system.

Monitor for player attrition, which might impact trial results near end of the experiment. When possible, compute each trial's outcome for only those players who completed all trials. After the experiment, analyze the trial data arranged by time to determine if increases or decreases in performance over time occurred irrespective of the nature of the trial. If temporal increases or decreases are found, analysis of covariance can be used (with caution) to statistically correct for unrelated temporal changes.

3.3.2.3 Threat 13. Data Collection Changes from Trial to Trial

There is always a danger that observed effects may be due to changes in the data collection instrumentation or procedures rather than changes in the test unit performance. As the test progresses, data collectors become more experienced and change their opinions as to what constitutes effective or ineffective responses, or they may become careless and less observant. Similarly, data collection instrumentation may change for the better or worse. Instrumentation technicians may improve their procedure, making it more precise. Conversely, instrumentation may deteriorate if it loses calibration.

Threats to design validity based on data collection changes are reduced by good practices such as: 1- monitoring for changes in data collection procedures, counterbalancing trial sequence when possible, and monitoring for any increases or decreases in performance over time; 2- re-calibrating sensitive data collection instrumentation before the start of each successive trial; and 3- monitoring for data collector attrition or data collector substitution after a trial has started. When possible, a good practice is to compute each trial's outcome for those data collectors who completed all trials to see if their responses differ from those who did not complete all trials.

After the experiment, analyze the trial data arranged by time to determine if increases or decreases in performance over time occurred irrespective of the nature of the trial. If temporal increases or decreases are found, analysis of covariance can be used (with caution) to statistically correct for unrelated temporal changes.

3.3.2.4 Threat 14. Trial Conditions Change from Trial to Trial

This threat represents all of the uncontrolled variables found in the experiment setting such as weather, terrain, light conditions, starting conditions, and free-play tactics. To the extent these variables fluctuate randomly throughout the test, they constitute Threat 8 to detecting change. To the extent, however, they change non-randomly and produce an overall increase or decrease in performance over the sequence of trials, they constitute a threat to single-group design validity by providing alternative causes of change in performance from trial to trial.

Good practices for holding this threat in check include exerting as much control as possible over the trial start and execution conditions, monitoring any changes in the test setting from trial to trial, counterbalancing trial sequence when possible, and checking for any increases or decreases in performance over time across trials.

3.3.3 Multiple-Group Experiment Causality Problems

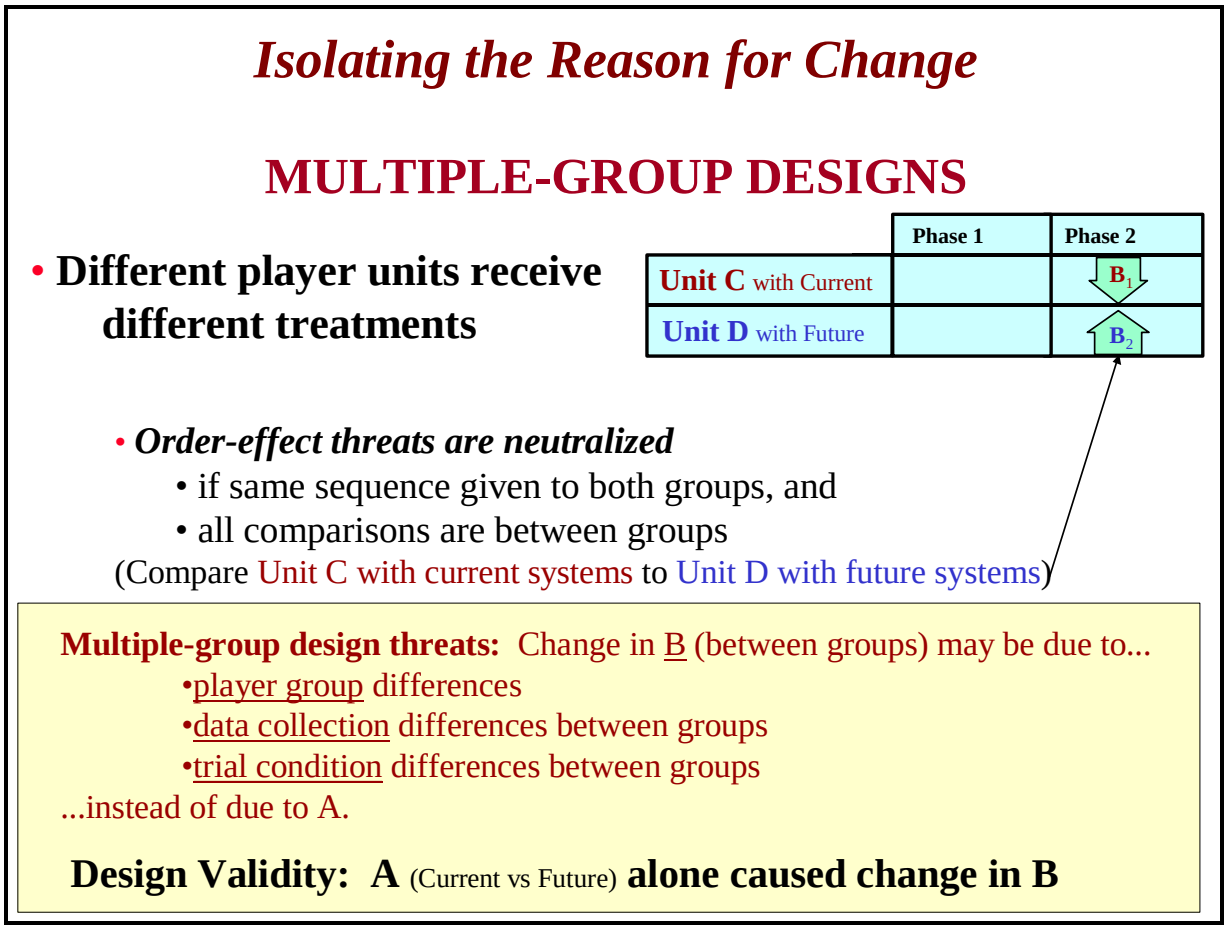


Figure 22 Isolating the reason for change in multiple-group design

In multiple-group designs (Figure 22) the sequence of trials is no longer the primary concern. If both the new-system player unit and the control player unit conduct their day trials first and the night trials last, any artificial increase or decrease in the subsequent night trials will affect both groups. Comparisons between the two groups for performance differences during night trials (or day trials) are immune to order effect threats as long as both groups undergo trials in the same sequence, the rate of change for both groups is similar, and the focus of the analytic comparison is between groups rather than within groups. That is, we are more interested in comparing *new*-system night trials with *old*-system night trials (between-groups comparison) rather than comparing *new*-system day trials to *new*-system night trials (within-groups comparison).

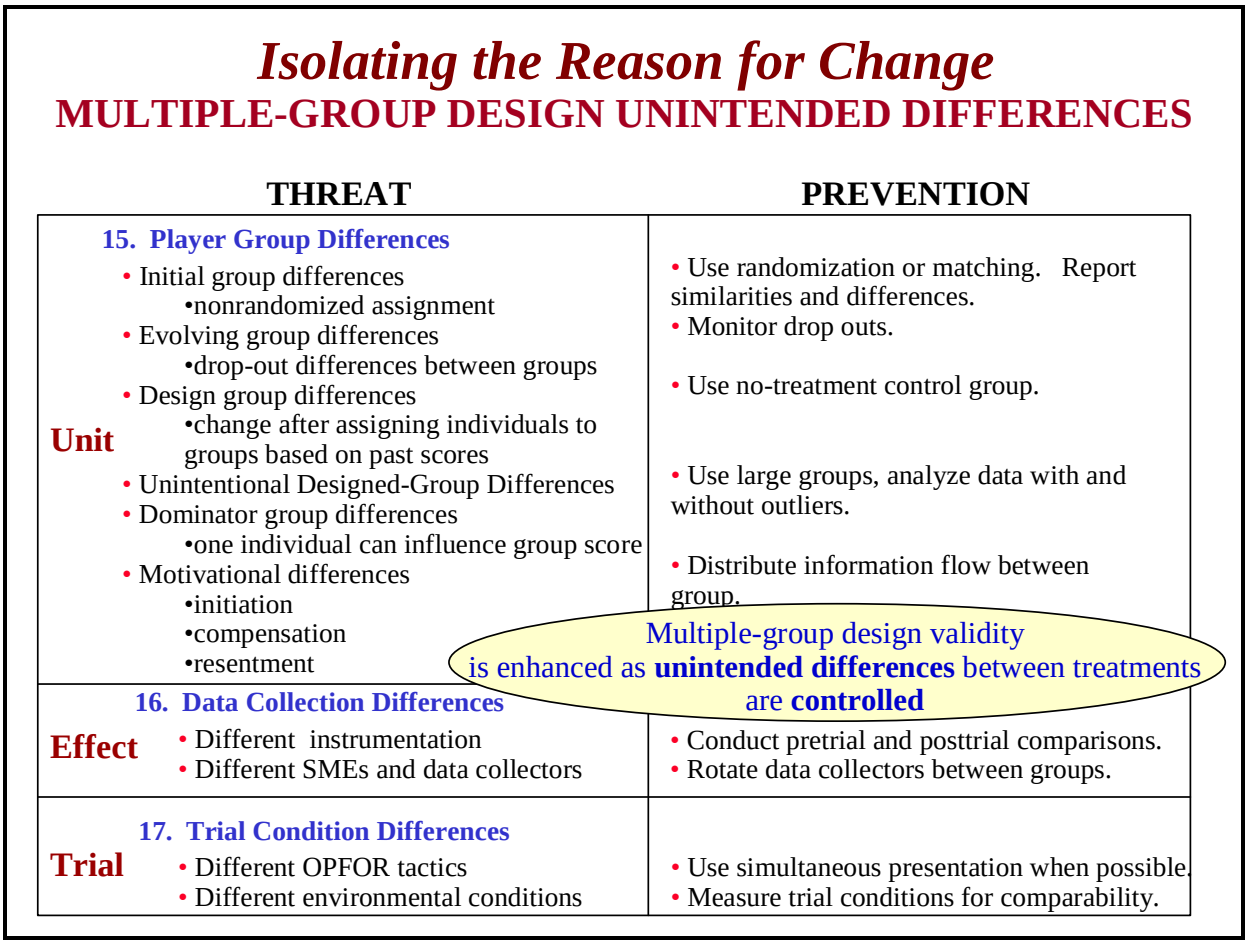


Figure 23 Multiple-group design unintended differences

The primary concern in multiple-group designs is potential confounding due to the inherent association of separate treatments with different player groups. The following three threats (Figure 23) are critical to isolating the true cause of change for between-group comparisons.

3.3.3.1 Threat 15. Player Differences Between Experiment Groups

Inherent differences between player units may result in spurious differences between treatment groups. Assignment of different units to different conditions is necessary when a player unit cannot undergo both treatment conditions sequentially. This occurs frequently in field experiments of new systems since a single player unit cannot be experimented under both the old and new systems. The unit would not be at the same level of experience when it began to use the second system. When different player units undergo different treatment conditions, there is always the danger that any results may be because of some characteristic differences between the units, rather than differences created by the treatment systems. There are six different aspects of grouped differences to be considered.

Threat 15-1. Initial Group Differences. This is the major consideration. Player units may differ at the beginning of the experiment in a way that will influence the outcome. Initial group differences arise because of unequal non-randomized group assignment.

1. The ideal good practice to achieve equal assignment is to measure all of the characteristics of the player units that affect experiment outcome. These characteristics might include years of experience, gender, and rank. Assignment to treatment conditions based on these measured traits is an attempt to make the player groups equal at the start of the experiment. Assigning matched individuals to different treatment groups is seldom possible since soldiers come to the experiment as part of an existing unit and most defense experiments involve integral player units. Assignment based on measured traits, even when doable, is probably not that effective. Those traits most likely to influence the outcome—motivation and leadership—are the hardest to measure.
2. An alternative good practice to matching is random assignment. In an experiment involving a large number of players, for example 50 riflemen, it is possible to randomly assign the soldiers to different treatment conditions, for example, current weapon and future weapon. The advantage to randomization is that it equates the two groups on all characteristics (measurable and non-measurable) that could affect the experiment results. Unfortunately, randomization only works when a large number of experimental units (individual soldiers, teams, crews, and sections) are in the experiment and random assignment does not affect unit integrity.
3. When it is not feasible to equate treatment groups before the experiment, a good practice for accounting for inherent group differences can be facilitated by experiment design manipulations. One technique is to have each group participate as its own baseline. As an example, in a field evaluation of two competing advanced helicopters X and Y, six pilots who flew advanced helicopter X also flew the current baseline helicopter and six other pilots who flew advanced helicopter Y also flew the current baseline helicopter. One of the outcome measures showed that version X performed better than version Y. However, when compared head-to-head in the baseline helicopter, the version-X pilots also performed better than the version-Y pilots. Thus, the correct interpretation is that no performance differences attributable to helicopter differences were found. Performance differences were correctly attributed to initial, inherent group differences.

Threat 15-2. Evolving Group Differences. Treatment groups, assessed as equivalent at the start of an experiment, may not be equivalent at the end of the experiment. This occurs in experiments that continue over a long duration, say several weeks or months, and players in the different treatment conditions drop out at different rates. Dropouts, or “experiment casualties,” are individuals who leave before completion of the experiment for any number of reasons: for example, emergency leave or change of assignment. Artificial group differences may evolve when more players in one experimental condition drop out than in the second condition. A differential dropout rate does not result in initial group differences. Instead, it results in differences between groups after the experiment has started even though the groups may have been equivalent at the beginning of the experiment. A good practice is to monitor experiment casualties in long experiments for their potential impact on group results.

Threat 15-3. Designed Group Differences. Some experiments are designed to begin with nonequivalent groups. This occurs in defense experiments of training devices where soldiers who scored low on some index are assigned to additional training on an experimental new training system. For example, soldiers with low marksmanship scores may be assigned to an experimental laser rifle-training program. The danger in

assigning individuals to treatment conditions based on prior performance is that their performance will change automatically. Individuals with low pre-experiment scores will exhibit an increase in post-experiment scores while individuals with initial high pre-experiment scores will exhibit a decrease in post-experiment scores. This shift toward the middle of the experiment score range (regression toward the mean) occurs in the absence of any additional training and is a result of the measurement error involved in obtaining the initial high and low scores from the pre-experiment. Consequently, players assigned to a training condition based on low scores will show an improvement upon re-experimenting even if the new training system is irrelevant to performance.

A good practice to reduce this risk is to establish a control group. Soldiers with low pre-experiment scores would be assigned randomly to two groups: a control group and the new-training group. The control group would not participate in any remedial training. While both groups will show improvement upon retesting, if the new-training group shows more improvement than the control group, a case can be made for the utility of the new-training system.

Threat 15-4. Unintentional Designed-Group Differences. Group differences can unintentionally be occurring before the formal experiment begins; for example, if only one of two equivalent player units was required to undergo pre-experiment activities. If Unit X is required at the experiment site two weeks early for extra training to run through a practice scenario to develop the associated techniques for employing the new capability, then Unit X will approach the experiment differently than Unit Y.

Threat 15-5. Group Dominator Differences. When treatment groups are small, one operator, one crew, or one team, or one individual may drastically influence the group score for better or for worse. Larger groups are the best remedies. When this is not possible, analysts should examine data for group dominator effects, sometimes referred to as outliers. Group results can be analyzed with and without outliers included to see if conclusions are reversed.

Threat 15-6. Group Motivation Differences. Experiment players will try to figure out what the experiment or exercise is all about and behave accordingly. The threat to design validity occurs when the separate treatment groups are operating under different motivations, thereby confounding (confusing) the interpretation of any treatment differences. There are three variations of this theme:

1. *Imitation.* There is the danger that one group will imitate the other group rather than respond to its own treatment. For example, in an experiment in which manual and automated intelligence analysis systems are compared, the two groups may share information during lunch breaks. Consequently, the group using the manual process may imitate the responses of the group using the automated process. Not only does this exchange of information diffuse any potential performance difference between two groups, the group using the manual procedure no longer reflects an operational unit using only manual procedures. A good practice is to keep competing groups continually separate throughout the experiment.
2. *Compensation.* This is called the "John Henry effect." When individuals become aware of being evaluated in a less desirable or more strenuous condition, they will often push themselves harder to outperform those in the easier condition. Experiment players in a baseline condition may push themselves harder to demonstrate that they are better than the unit selected (with the

accompanying publicity) to receive the new, potentially superior system. Experimentation results would run counter to the hypotheses and would be a result of motivation rather than the intended treatment.

3. *Resentment.* This is the opposite reaction. Experiment players in the less-desirable experimental condition may perform poorly as a result of being selected for this condition rather than the more desirable condition. Their poor performance would exaggerate any actual effect due to the experimental conditions.

Good practices for threats of compensation and resentment are not always easy to find. At minimum the experimenter needs to continually monitor the attitudes and motivations of different groups in the experiment so at least these threats, if operating, can be recognized. Providing equivalent publicity and recognition to all groups in the experiment will help to offset the natural feelings of compensation and resentment.

3.3.3.2 Threat 16. Data-Collection Differences Between Experiment Groups

The same amount of effort to ensure that two different player units are equal should also be taken to ensure that data collection methods for each group are equal. For example, in side-by-side comparison experiments different data collectors are assigned to the different experiment player units. Are the data collectors assigned to the different groups equivalent? Data collectors and the accuracy and reliability of the instrumentation for each group need to be equal. Additionally, the allocation of data collection devices between different experiment groups may reflect the experimenter's expectation. Rosenthal [Rosenthal 2002] has described how the “experimenter's expectancies” concerning the outcome of an experiment may bias the data obtained (and even the subsequent data analysis). Expectations concerning which evaluated system should be better may bias the results if data is collected differently. When this occurs, it is difficult to know whether the reported outcome is a result of the intended treatment or a result of the differences in data collection procedures.

A good practice is to ensure that the new-capability group does not get all of the best instrumentation and most proficient data collectors. The experimentation team, including the analysts, must continually scrutinize their own motivation to ensure that their expectancies are not biasing the data analysis and collection.

3.3.3.3 Threat 17. Trial-Condition Differences Between Experiment Groups

This threat represents the uncontrolled variables found in the experimental setting; such as weather, terrain, tactics, and opposing forces (OPFOR) experience (Red players). To the extent uncontrolled trial variables impact the different experiment groups differently, these influences constitute a threat to experiment validity Requirement 3 by making it difficult to interpret differences in group performance.

This threat is always present in field experiments because two different player units cannot occupy the same terrain and execute the same trial at the same time. There will always be some trial differences. The goal is to minimize any difference that may affect the outcome of the trial. The best practice to minimize this threat is to execute as much of the trial as possible simultaneously for each treatment group. Experiments of detection systems allow simultaneous presentation of targets to all experiment groups.

This ensures that all environmental and most target characteristics are the same for all shooters. To ensure equality of the target aspect angle, a shooter's position can be alternated after each trial. Monitoring any differences in the experimental setting between groups, and counterbalancing the trial sequence between groups when possible, also reduce this threat.

3.3.4 Summary

In summary, the assessment of experiment validity Requirement 3, ability to isolate the reason for change, is a logical assessment. This is in contrast to Requirement 2, ability to detect change, which can be evaluated statistically. Assessment of Requirement 3 requires knowledge of what factors other than the new capability might affect experiment results. Careful consideration and monitoring of the ongoing experiment can neutralize many of the design validity threats. This is the one area where experience in experimental design will pay dividends after the field exercise is completed. Attention to Requirement 3 will allow analysts to interpret results in a clear, unambiguous manner, attributing any changes in the outcome to the new capability alone.

3.4 Experiment Validity Requirement 4: Ability to Relate Results to Actual Operations

3.4.1 Importance of Relating Results to Actual Operations

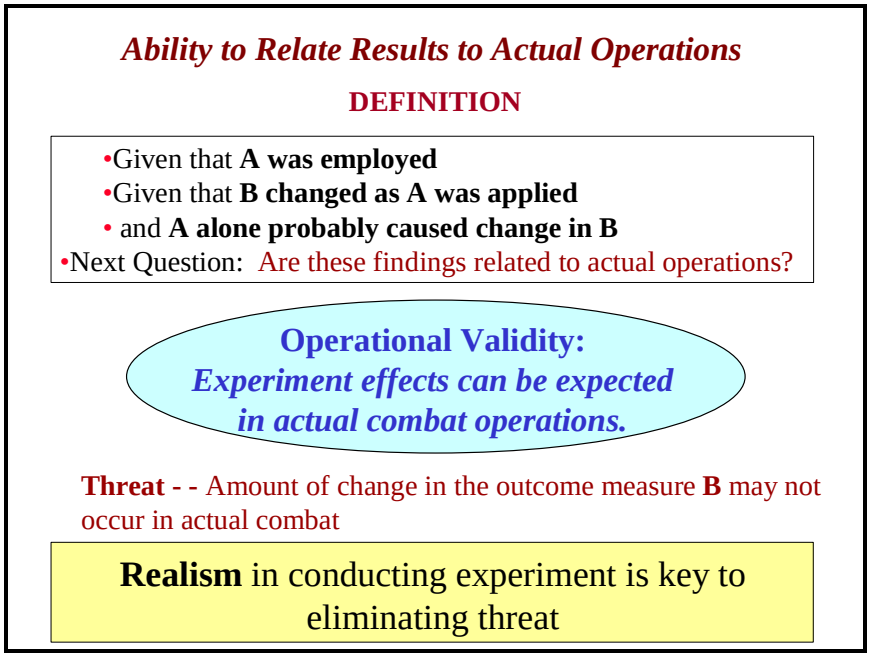


Figure 24 Threat to experiment operational validity

Let us now suppose that the experimenter was successful in employing the new capability, detecting change, and isolating the cause. Now the question is whether the experimental results are applicable to operational forces in actual military operations. The ability to generalize experiment results to the operations of interest is termed

operational validity. This fourth experiment validity requirement is the easiest to understand but the most difficult to achieve. It is easy to understand that a defense experiment ought to represent actual military operations (Figure 24). It is difficult to achieve because many operational conditions of importance are difficult to represent in the experiment environment. The more operational conditions represented in the experiment, the easier it is to provide evidence that experiment results will be applicable to an operational unit in an operational situation.

3.4.2 Threats that Diminish Experiment Generalizability

Experimental results are only useful to the extent they say something about the real world. *Generalizability* is the scientific term for the ability to apply results outside the experiment context. Ability to relate results pertains to experiment realism. The threats to Requirement 4 limit the realism of the experiment itself making it more difficult to generalize, or translate, from the experiment to military operations in real-world operations. The following four threats illustrated in Figure 25 limit the ability to generalize experiment results.

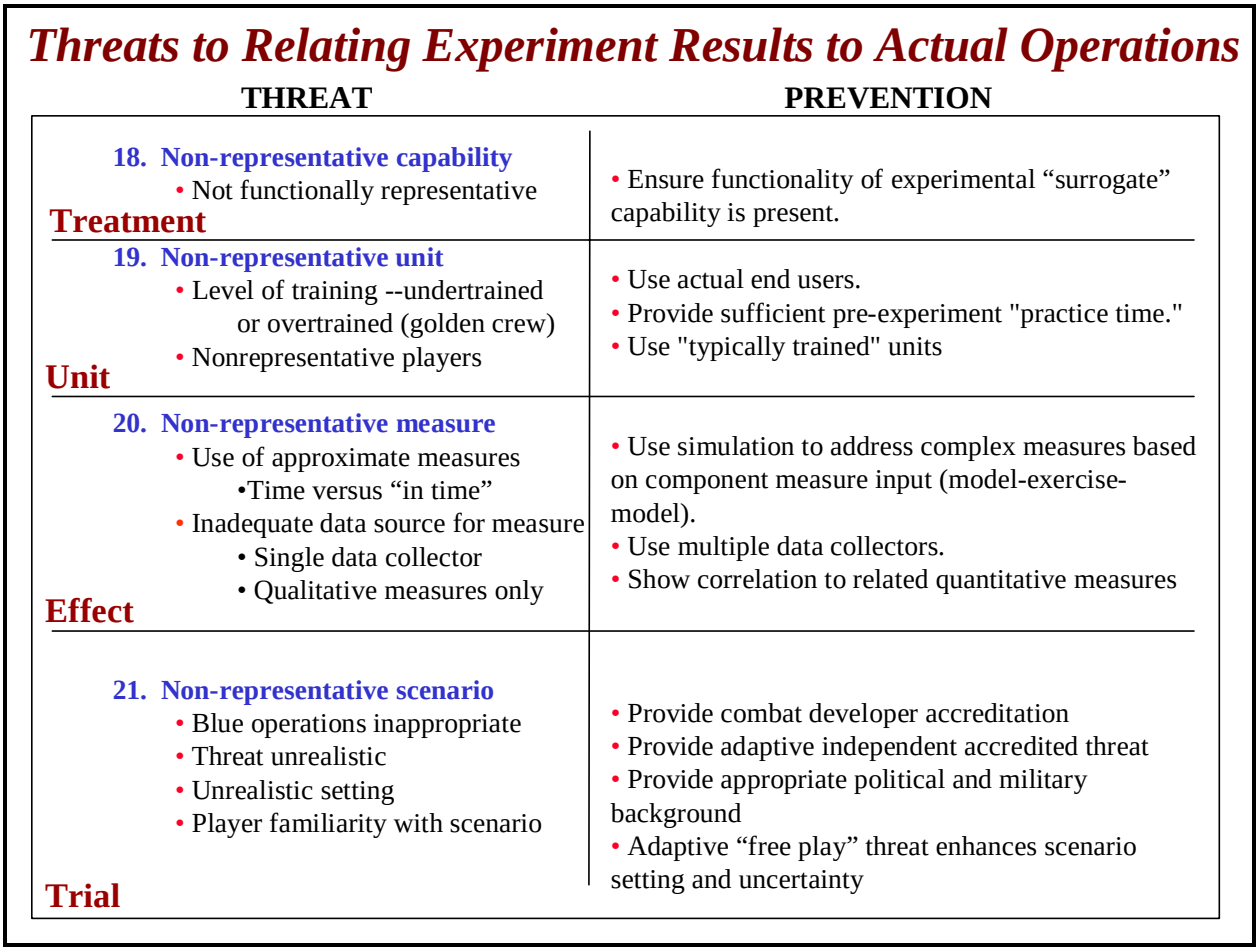


Figure 25 Threats to the generalizability of experiment findings

3.4.2.1 Threat 18. Non-representative Capability

Future systems in defense experiments are rarely sufficiently mature to give confidence in the representativeness of their future functionality. First, new capabilities continually evolve before, during and after the experiment. As the capability evolves post-experiment, it will be difficult to match the experiment results to its evolutionary functionality. Second, and more importantly, new capabilities are dependent on surrogates during experimentation and the question is the representativeness of the surrogates to future functionality. In experiment validity requirements 1, 2, and 3 the "internal validity" of the experimental capability concerned its employability, variation in producing effects, and potentially changing functionality between trials. In Requirement 4, the question concerning the experimental capability concerns the "external validity" of the experiment. To what extent is the experimental capability sufficiently representative of the future "real" capability to conclude that the experiment findings are relevant to the use of this future capability?

Very early-idealized surrogates tend to be overly optimistic in representing future capability. Importantly, however, these optimistic surrogates are useful in examining the worth of pursuing a particular capability experimentation campaign. These experiments investigate whether an optimized capability can markedly improve warfighting effectiveness. If the experiment results are negative, there may be sufficient reason to not explore further. If positive results, a case can be made for further experimentation on more realistic surrogates to get more accurate estimate of potential effect.

Interestingly, as subsequent surrogates become more realistic, sometimes referred to as prototypes, they may tend toward underestimating the potential future capability. As the surrogates incorporate more and more of the software, hardware, and process modules of the "final" configuration, there will be inevitable functionality deficiencies brought on by the immaturity of the development software, hardware, processes, and integration problems. The interpretation of experiments with "under-representative surrogates" that produce low effects is much more difficult. Were the low effects due to the poor representation of the prototype and a more functional prototype would have produced better results? Capability proponents will always be accused of wishful thinking. The more realistic the surrogate, the more time has to be devoted prior to the experiment to ensure that it has sufficient and stable functionality or the experiment will not be interpretable.

A good practice is to accurately report the strengths and limitations of surrogates and prototypes used in the experiment. When defense experiments are used as the final event to decide if a new capability should be deployed to the operating forces, it is critical to use fully functional prototypes to get accurate estimates of their effectiveness. On the other hand, use of surrogates with major limitations is permitted, even encouraged, in early experimentation in the concept development cycle. These early surrogates permit a preliminary look at the system's potential military utility, help develop potential human factors requirements, and help identify potential failure modes to facilitate an experiment-fix-experiment paradigm. Early experimenting with

surrogates and prototypes also provides critical information to influence design decisions. However, the limited capability of early experimenting to relate conclusions from prototype systems to production systems in actual operations needs to be recognized and accounted for in later experimentation.

3.4.2.2 Threat 19. Non-representative Experiment Unit

How well do the experiment players represent operators and operational units that will eventually employ the experimental capability? There are three related issues in this threat: the prior experience of the experiment players, their level of training on the new capability, and their motivation for participating in the experiment.

A good practice to enhance experiment generalizability is to select experiment players directly from an operational unit that will eventually employ the capability. Often, however, defense experiments use reservists, retired military, or government civilians due to unavailability of operational forces. This is not a major threat when the experimental task represents basic human perception or cognition. However, if the experiment task represents a military task under combat conditions, the absence of actual experienced military personnel would jeopardize the applicability of any observed effects.

Even when operational forces are available as the experimental unit, the experimenter has to be concerned about the appropriate level of training on the new capability. If the experiment unit is undertrained or overtrained, the true capabilities of soldiers in a typical unit will be misrepresented. Undertraining results from compressed schedules to start the experiment and inadequate training development for new concepts or new systems. Overtraining arises when player units undergo unique training not planned for units that will receive the fielded systems. Overtraining, like undertraining, is difficult to avoid.

The good practice is to ensure the experiment unit is well qualified to operate the experimental systems and experimental concept so that the systems and concept will be given a fair evaluation. The temptation is to overtrain the experiment unit to ensure success. An overtrained experiment unit is referred to as a “golden crew.” The challenge is to produce a well-trained, typical unit rather than an overtrained or undertrained unique experiment unit.

Participant motivation is always of concern in defense experiments. Since motivation affects performance, the concern is the extent the participant’s motivation during the experiment represents the motivation expected in the actual environment that should be represented by the experiment. Constructing a realistic experiment setting, as discussed later as a counter to Threat 21, is important to approximating the conditions under which the experiment players will perform. In the actual environment, it is expected that military personnel will work extremely hard to achieve their mission under any condition. In the experiment, this same motivation needs to occur and most often it does because participants are professionals and want to excel.

Three potential problems can occur, however, that can produce under or over motivation yielding unrealistic low or unrealistic high results. When personnel are assigned to participate in the experiment as “an additional” duty and it is perceived to be unrelated to their real mission, motivational problems can occur. In this case, participants may “under perform” out of lack of interest or resentment.

A second problem is that players may “over perform” due to being in the spotlight of an experiment. This is known as the "Hawthorne effect" where it was found that factory workers increased productivity, not because of different experimental illumination levels in the workplace; but because the workers were being observed. The Hawthorne effect is more likely to occur in highly visible experiments that have continual high-ranking visitors. In this instance, the players are motivated to make the capability “look good” to please the audience even though the capabilities may not be that effective.

The third area is to avoid inducing “experimenter expectancies” in the experiment groups where they perform according to the expectation of the experimenter (also known as Pygmalion effect). If the experimenter expects the control group to do less well than the new-capability group, the control group may perceive this and perform accordingly.

It is always a good practice to continually monitor the motivation of the participants. Sufficient time has to be allocated to explain the importance of the experiment and their contribution to the effort emphasizing that the success of the experiment is not whether the capability produces a positive result but that it was thoroughly and realistically employed so that it can be honestly evaluated.

3.4.2.3 Threat 20. Non-representative Measures

Ensuring representative measures is easier when examining the effects of new capabilities on relatively simple military outcomes such as target detections, targets killed, attrition, transit time, and so on. The primary concern here is measurement bias. Is the measure of these relatively simple and straightforward effects not biased? A biased measurement is one that tends to provide an output that is over or under representative of the true value. Measurement precision[19], in this context, means that the output is unbiased: the measure does not measure to the left or right of the true value. A biased data-collection device or measure would over or under represent the effect of the new capability and thus the effectiveness of the capability in the experiment would not represent its future potential, for better or worse, in the operational environment. A good practice for ensuring the precision, non-bias, of simple measures is pilot-testing the data collection instrumentation to ensure its accuracy.

[19] The reader may recall that the earlier discussion under Threat 7, defined measurement precision as consistency, or reliability of output. Here the meaning of precision is non-biased measurement or accuracy. Both consistency and non-bias are essential to measurement precision. The consistency aspect of measurement precision applies to Experiment Requirement 2, ability to detect a result, finding a consistent signal in a sea of noises. The non-bias aspect of precision applies to Requirement 4, ability to relate the results. Detecting a “consistent signal” that is offset (biased) from the actual signal is a threat to Requirement 4, relating results, because the “offset signal” was an experiment artifact.

Non-biased representative measures are more difficult to achieve when the new capability is attempting to achieve a complex result, such as information superiority, improved planning, better decisions, increased situational awareness, better collaboration, or mission success. These complex operational concepts are difficult to define and, not surprisingly, difficult to measure in actual operations and in defense experiments. There are **two** general good practices to develop representative experiment measures of complex outcomes. Both of these good practices have strengths and weaknesses.

1 Combining concrete components of complex effects. Overall unit effectiveness, for example, may be definable in terms of concrete, measurable variables such as loss-exchange ratio, rate of movement, and time to complete a mission. A weighted or un-weighted composite score of the components can be combined to represent the complex effect. There are several problems with this approach.

One problem is that component measures may not covary in a similar fashion. In some instances, a slow rate of movement may be associated with a low loss ratio. In other instances, it could be associated with a high loss ratio. While the individual component variable scores can be reported, these scores by themselves do not address overall unit effectiveness that is the measure of interest. An alternative approach is to select a single component measure that represents the highest level of interest in the complex variable.

A second problem is the "halo effect." When measuring multiple components, analysts need to ensure individual components are measured independently of each other. If all of the components are measured in the same manner, any covariation among the component indices cannot be disassociated from the influence of its "method of measurement." This is problematic whether the sole data source for all component measures is a SME rater, a questionnaire, or electronic instrumentation. For example, if a single rater provides estimates for a unit's ability to maneuver, to collect intelligence, to engage the enemy, and these three estimates are combined into a unit effectiveness score; the covariation of these component measures may be artificially high due to a "halo effect." Any inaccuracy in the single data source (a single rater) induces the same error in each component score resulting in an inflated component covariation. To avoid this halo effect, a good practice is to collect component data using independent sources (raters, participant surveys, instrumentation) whenever possible.

2 Measure complex effects with overall subjective rating. A knowledgeable SME can provide an overall rating or assessment to provide a "score" for the complex variable of interest. This alleviates the problem of defining, measuring, and combining data from component measures. However, use of subjective rating brings its own set of problems: inconsistency and, even if consistent, because a consistent inaccuracy is an undesirable bias in the scores. The problem of inconsistency and associated good practices was discussed previously under Threat 8. The problem of potential bias in subjective assessments will be discussed here. For our purposes, a bias judgment is one that is "consistently off the mark" whereas an inconsistent judgment is one that is "sometimes on and sometimes off the mark." There are good practices for calibrating

and enhancing the non-biasness, accuracy of subjective ratings similar to those for improving consistency discussed previously under Threat 7.

Good practices for calibrating the objectivity of subjective ratings to estimate the extent of individual bias in subjective ratings is to continually assess inter-rater agreement of independent experts observing the same event. Second, it is important in training to allow them to observe predetermined “good” and “poor” practice events to determine if their assessment differentiated. During the experiment execution it is important to collect objective quantitative component scores in addition to the composite rating provided by the SME. Confidence increases in the subjective ratings to the extent they correlate to the independently obtained component measures. Another good practice for increasing the “objectivity” of “subjective” ratings is to employ several raters independently and combine their individual scores into a single overall assessment. And finally, the veracity and generalizability of SME ratings rest on the operational experience and credibility of the raters.

3.4.2.4 Threat 21. Non-representative Scenario

How realistic is the experiment scenario for the Blue- and Red-force experiment participants?

Realistic Blue-force Operations. Many factors make it difficult for the experimental unit and opposing forces to use realistic TTP during an experiment. Additionally, modifying current Blue-force tactics to incorporate the new capabilities and countering opposing new capabilities often follows rather than precedes new capability development. Even when new techniques and procedures have been developed, adequately training is difficult due to surrogate shortages until experiment execution. Additionally, terrain, instrumentation, or safety restraints during experiment execution may preclude appropriate tactical maneuvering during field experiments.

Good practices include allocating sufficient time for training the experiment unit and threat unit in appropriate tactics with the new capability. Tactical units can assist the experimenter in developing realistic operational plans that provide for appropriate force ratios, missions, and maneuver space and time.

***Realistic Setting*.** It is impossible to create conditions during a field experiment that approximate the noise, confusion, fear, and uncertainty of combat. A good practice for offsetting the potential lack of player apprehension during experiment trials is increasing the realism of player participation. The use of lasers to simulate engagements increases the realism of tactical engagements. Other good practices include allowing the experiment to continue for many hours or days to generate fatigue-associated stress.

Over time experiment players can anticipate and prepare for scenario events. Directing a unit to an assembly area during continuous operations to calibrate instrumentation is a signal to the unit that a battle will soon occur. Surprise has evaporated. Additionally, player units that undergo the same scenario over successive trials know what to expect. Anticipation of scenario events decreases apprehension and promotes non-

representativeness of unit reactions. Good practices allow for maximum free-play and sufficient scenario space to promote player uncertainty, player creativity, and sufficient opportunity to explore and attempt to solve the warfighting problem.

Realistic and Reactive Threat. Representation of threat tactics and equipment in the experiment is a special difficulty. Captured threat equipment is not always available for field experiments and training operational units to emulate threat tactics is a low priority except at centralized training centers. It is difficult to imagine what would the adversary do in any given situation. It is all too easy to imagine and rationalize what a given nation would do in a similar situation. History has shown, however, that irrational leaders do exist and we should not always prepare for the rational, mirror-image adversary.

A good practice to enhance threat realism is to conduct field experiments at the national training centers, when possible, because they can provide realistic, well-trained threats. When not conducting defense experiments in the field, good practices include using threat experts from the national agencies to assist in designing the future threat in the experiment scenarios and to monitor the conduct of the threat during experiment execution. Additionally, the threat has to be given maximum free-play to respond to and even preempt, if possible, Blue-force employment of the new experimental capability. The development and employment of an intelligent, determined opposing force is one of the best counters to the threat of non-representative scenarios.

3.4.3 General Good Practices to Enhance Experiment Relevancy

Experiments can never be perfect representations of actual combat operations. Meeting Requirement 4, however, depends on approximating the operational conditions to which the conclusions of the experiment are pertinent. All experiments are approximations to operational realism and can never fully represent actual operational conditions. To formally assess operational validity (see Figure 26), the analyst would need to examine data from a series of similar experiments involving different units and different environments. Field experiments for the sake of replicating findings are seldom funded. Consequently, the assessment of operational validity rests on judgments as to the representativeness of the system, the measures, the player unit, the scenario, and the site conditions under which the experiment was conducted.

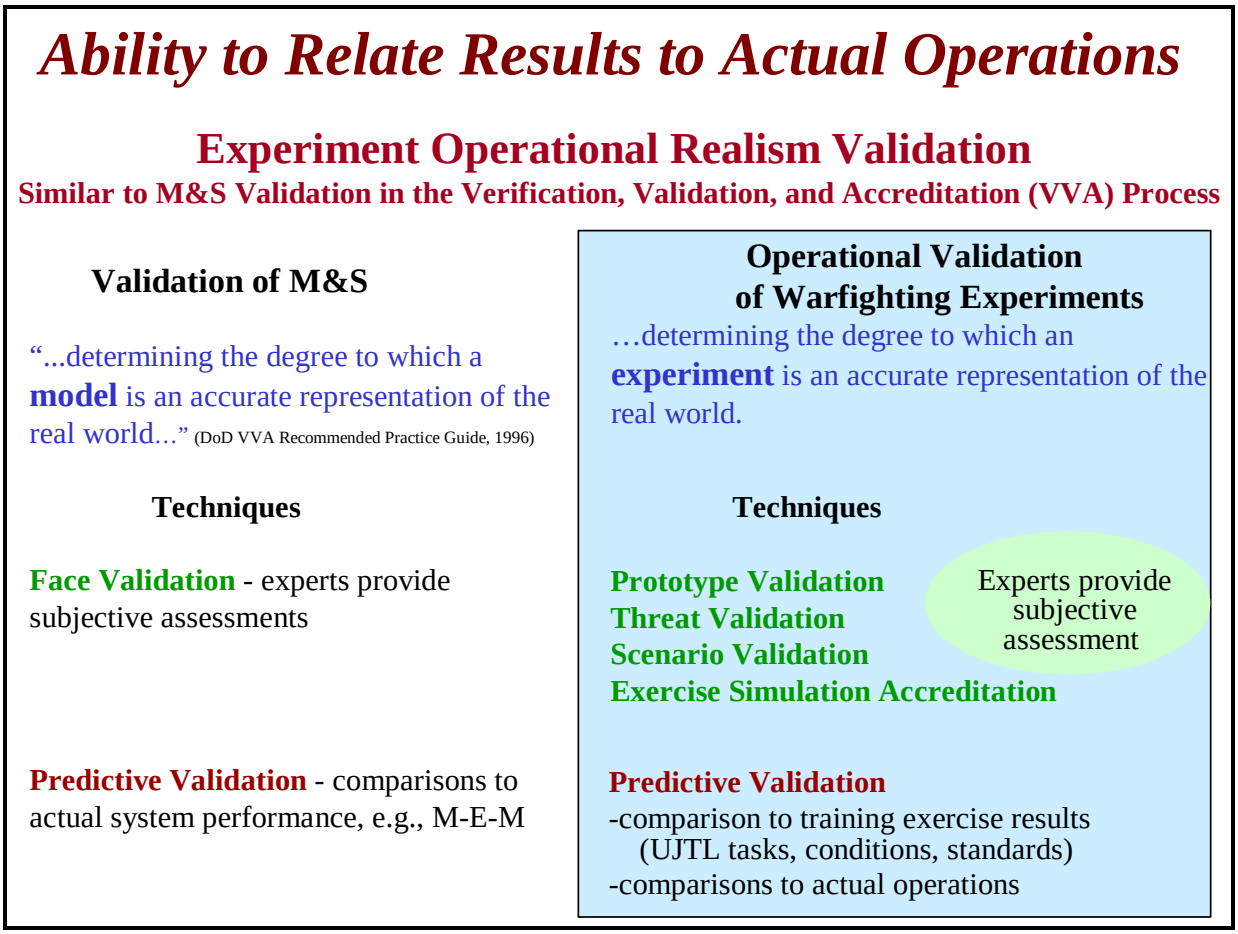


Figure 26 Ability to relate results to actual operations[20]

Many of the good practices for validating the representativeness of the experiment environment are similar to the techniques used in the validation of M&S, especially the idea of "face validity." In most cases, experts from inside and outside the defense organizations are employed to certify and validate the prototypes' capabilities, the scenario and the treat play in the scenario, and any experiment simulations. Where possible, some "predictive validity" techniques may be employed to the extent conditions in the experiment scenario that can be related to real-world exercises, deployments and operational lessons learned.

[20] The quotation is from The DoD VVA Recommendation Practice Guide [DoD Modeling and Simulation Office (DMSO) 1996].

3.5 Summary of Good Practices to Meet the Four Experiment validity requirements

This section summarizes the good practices discussed as counters to the 21 threats to the four experiment validity requirements.[21] These are not presented as "cook book" solutions to designing an experiment. As discussed in the previous section, it is impossible to satisfy all four experiment validity requirements simultaneously because the requirements seek to achieve contradictory goals in defense experiments: maximization of experiment statistical power and control on one hand, and maximization of free-play and real-world operations on the other hand. An understanding of the rationale for the four experiment validity requirements permits the experimenter to make knowledgeable and rational tradeoffs among the good practices to maximize the applicability of the knowledge that can be gained from a single experiment within the context of a campaign. This allows campaigns to address successively complex questions.

These good practices are selective. They only pertain to the threats to defense experiment validity. Good practices involving the mechanics of agency organization, planning, and reporting of defense experiments are critically important to the success of a campaign but are not included here. However, these agency good practices for including stakeholders, peer reviews, having experienced practitioners, and allocating sufficient time and resources to plan, execute, and report an experiment certainly have implications for designing valid experiments by countering the threats to the four experiment validity requirements.

And finally, the following good practices are not exhaustive. They are provided as examples and aides to better understand the 21 threats to experiment validity. Understanding the specific threats to validity and their importance to the logic of defense experimentation allows the experimenter "on the ground" to be creative in finding more innovative methods for countering specific threats. Each defense experiment agency already has a list of useful experiment practices. These lists of good practices (do's and don'ts) by experienced practitioners can now be partitioned to reinforce and expand the good practices provided below. The discussion in the previous sections of Principle 3, this chapter, provides a common framework for organizing and understanding the good practices gained by different practitioners. The framework relates all good practices that promote experiment validity to a thematic logic, the mnemonic numbers "2, 3, 4, 5, and 21" for defense experimentation. This logic allows experimenters to understand the relative importance, the interrelationships, and the tradeoffs required in using their own good practices to design better defense experiments.

[21] The four experiment requirements, the threats to validity and the good practices to address the threats are adapted from an expansion of those described in [Shadish *et al.* 2002]. Please refer to the introduction text to Principle 3 for details on the use of this work.

3.5.1 Techniques to Counter Threats to Requirement 1: *Ability to Use the New Capability*

Threat 1: New Capability Does Not Function.

1. Schedule frequent demonstrations of the new capability prior to the experiment. These demonstrations should take place in the experiment environment.
2. Prior to the experiment, ensure that new command, control, and communications (C3I) systems interoperate with the other systems in the experiment. Systems that interoperated in the designer's facility almost surely will not when brought to the experiment.

Threat 2: Experiment Players Cannot Use or Employ the New Capability Effectively.

3. Provide sufficient practice time for players to be able to operate and optimally employ the system. Not only does the new functionality need to be available ahead of time, but also the techniques, tactics, and procedures (TTPs) and standard operating procedures (SOPs) need to be developed concurrently with the new capability and available prior to the pilot test.

Threat 3: New Capability Cannot Impact Experiment Outcome.

4. Conduct full-dress rehearsal pilot tests prior to the start of experiment trials to ensure the experimental capability in the hands of the user can produce the anticipated outcome.
5. If the experiment is to examine various levels of the capability (or the same capability under distinct conditions), by design increase the differential between the various levels or the distinct conditions in order to increase the chance of seeing differences in experiment outcomes.
6. If the experiment is to be a comparison between the old and new capability, it is critical to include the old capability in the pilot test also to see if performance differences will occur.
7. In a comparison experiment, design some experiment trials where it is expected that the old system should perform equivalently to the new capability and trials where the advantages of the new capability should allow it to excel. Both of these trials should be examined during the pilot test to assess these assumptions.
8. New experimental capabilities that are to be simulated can be rigorously tested in the simulation prior to the experiment itself. The sensitivity of the simulation to differences between the old and new capability should be part of the simulation validation and accreditation. Pre-experiment simulation of the old and new capabilities can also serve to identify trial scenario conditions that will accentuate similarities and differences between the old and new capabilities.

Threat 4: New Capability Not Adequately Exercised During the Experiment.

9. Develop detailed master scenario event lists (MSELs) that depict all the scenario events and scenario injects that are to occur over the course of the experiment trial. These pre-planned scenario events and scenario inputs "drive" the experiment players to deal with specific situations that allow for, or mandate, the use of the new capability during the trial.
10. Experimenters need to continually monitor not only that the MSEL occurred but also that the experiment players reacted accordingly. If the players did not attempt to employ the new capability when the MSEL event occurred, then ensure that the players actually "saw" the scenario event.

3.5.2 Techniques to Counter Threats to Requirement 2: *Ability to Detect Change*

Threat 5: New Capability Varies (Unreliability) Within an Experiment Trial.

A: For a single new-capability system that has to operate continuously over the length of a trial:

11. Provide sufficient pre-experiment operating time for immature new technology to ensure it will work consistently for the duration of an experiment trial.
12. For an immature unreliable system, incorporate an experiment-fix-experiment methodology by designing a series of short experiment trials with treatment fixes occurring between trials rather than incorporating capability fixes (changes) during one long experiment trial. In this manner, the capability is held constant during each trial but allowed to improve from trial to trial in a systematic fashion. This experiment-fix-experiment approach now has multiple, sequential capability levels that can be examined separately.

B: For multiple new-capability systems in a single trial:

13. Use the pilot test to ensure all copies of the new capability function equivalently.
14. After the experiment the experimenter can assess the extent of capability variability by comparing individual scores across items.
15. When variability is a result of a few outlier cases, the experiment analysis can be performed with and without outliers to determine the impact of outliers on the analysis.

Threat 6: Experiment Players Vary Within an Experiment Trial.

16. It is always best to deal with this threat prior to the experiment. Consistency among experiment player responses can be improved prior to the experiment by thoroughly training everyone to the same level of performance before the start of the trial.
17. When possible, select similar (homogeneous) players to participate in the experiment to reduce player variability. However, this will compromise Requirement 4, external validity.
18. After the experiment the experimenter can assess the extent of player variability by comparing individual scores across players.
19. Variability in player scores can sometimes be statistically adjusted using covariance analysis with pre-experiment training and experience scores. Post-experiment statistical corrections are risky due to the statistical assumptions that accompany them.
20. When variability is a result of a few outlier cases, the experiment analysis can be performed with and without outliers to determine the impact of outliers on the analysis.

Threat 7: Data Collection Randomly Varies Within Experiment Trials.

21. Use objective data collection measures when possible that have been calibrated. Pretest data collection instrumentation to verify reliability (consistency).
22. Questionnaire scales can be calibrated using techniques such as item analysis to quantifiable consistency indices, *e.g.*, .85 internal consistency reliability. In general, increasing the number of related questions about a particular judgment in a questionnaire and combining these related items into an "overall judgment score" increases the consistency of player survey judgments.
23. Increase the objectivity (reliability, consistency) of subjective data collection procedures by adequately training data collectors. Data collectors can be objectively "calibrated" by comparing their observations across similar and dissimilar events during training.

24. Consistency of subjective assessment across events is enhanced by having a data collector provide multiple component ratings (or scores) of a single event, and then using the component assessments to produce an average assessment score.

25. A more consistent assessment can be obtained by combining or averaging individual assessments of two or more side-by-side observers who provide independent assessments.

Threat 8: Trial Conditions Randomly Vary Within an Experiment Trial.

26. An experiment result is more likely to be detected in experiments with a number of shorter trials with a constant condition within a trial but not between trials, than having only one long trial with a wide variety of conditions.

27. When the sources of the trial variability can be identified, some reduction in the variance can be accomplished by using statistical designs and techniques such as paired comparisons, matching and within-subject designs, blocking designs, and analysis of covariance (ANCOVA).

Threat 9: Low-Power Statistical Analysis Decreases Detections of Real Difference.

28. Use an adequate sample size. There are available techniques for estimating sample size requirements to achieve specific levels of statistical power. In general, the larger the sample size, the greater the statistical power.

29. Accept more risk by setting statistical requirements lower, *e.g.*, setting the statistical-rejection level at 90% risk instead of 95%. Setting too stringent a statistical risk will not allow small positive results to show up as statistically significant.

30. Use efficient statistical analysis techniques. Parametric techniques are generally more powerful than nonparametric techniques but they require more assumptions.

31. Use efficient experiment designs such as matching, stratifying, blocking, or within-subject designs. Efficient experiment designs and statistical techniques can reduce the sample size requirement to well below the standard notion of 30.

Threat 10: Fishing and Error Rate Problems Increase Chance of Incorrectly Detecting a False Change.

32. The probability of incorrectly concluding that a chance outcome is a positive change decreases as the statistical risk is decreased (*e.g.* setting the statistical-rejection level at 95% or 99% instead of 90%).

33. The likelihood of incorrectly detecting a false change increases as the number of statistical comparisons in a single experiment increases. Decrease the multiple-comparison error rate by increasing the required confidence level for each individual comparison, *e.g.*, 98% versus 95%.

34. Violating assumptions of statistical tests can increase the chance of incorrectly detecting a false change; but can also decrease the chance of detecting a real change. Analysis of variance (ANOVA) is fairly insensitive to departures from assumptions of normality or equal within cell variances. Analysis of covariance (ANCOVA), on the other hand, is quite sensitive to its requirement for homogeneous within group regression slopes. Nonparametric techniques, while less efficient than parametric techniques, require fewer assumptions than parametric statistics concerning the level of measurement and underlying distribution.

3.5.3 Techniques to Counter Threats to Requirement 3: *Ability to Isolate the Reason for Change in Single-Group Experiments*

Threat 11: New Capability Changes from Trial to Trial.

35. Allow sufficient time for the pilot-testing prior to the experiment to ensure the stability of the new-capability functionality.
36. Monitor that the functionality of the new capability does not change over the course of succeeding experiment trials where it is intended to be constant.
37. When experimental systems, especially future ones, undergo major modifications during a field experiment in order to correct discovered deficiencies in their functionality, consider whether trials conducted prior to the modification need to be rerun in order to make valid comparisons with the post-fix trials.

Threat 12: Experiment Players Change from Trial to Trial.

38. Monitor for player changes over the course of succeeding trials. Players may become more experienced and proficient, due to learning effect, or they may become fatigued, bored, or less motivated. Player changes over time will produce an increase or decrease in performance in later trials unrelated to the new capability.
39. Counterbalance the sequence of trials (*e.g.* NG-CG-CG-NG) so a sequential learning effect will affect the new-capability group (NG) and the control group (CG) to the same extent.
40. In general, conduct new-capability trials before the control current-capability trials. Any observed improvement for the new capability when compared to the current capability, has "overcome" any learning effects.
41. Ensure that players are trained to maximum performance and operate at a steady state prior to experiment start.
42. Monitor for player attrition which might impact trial results near the end of an experiment. When possible, compute each trial's outcome for only those players who completed all trials.
43. After the experiment, analyze the trial data arranged by time to determine if increases or decreases in performance over time occurred irrespective of the nature of the trial. If temporal increases or decreases are found, analysis of covariance can be used (with caution) to statistically correct for unrelated temporal changes.

Threat 13: Data Collection Changes from Trial to Trial.

44. Continually monitor for changes in data collection procedures to ensure consistency.
45. Re-calibrate sensitive data collection instrumentation before the start of each succeeding trial.
46. Monitor for data collector attrition or data collector substitution after the trial has started. When possible, compute each trial's outcome for those data collectors who completed all trials to see if their responses differ from those who did not complete all trials.

Threat 14: Trial Conditions Change from Trial to Trial.

47. Exert as much control as possible over the trial execution conditions to ensure consistency from trial to trial.
48. When new conditions occur that cannot be controlled, delay start of trial. When delay is not an option, record the trial differences and report the estimated impact on results.

3.5.4 Techniques to Counter Threats to Requirement 3: *Ability to Isolate the Reason for Change in Multiple-Group Experiments*

Threat 15: Player Differences between Experiment Groups

49. With large treatment groups, randomly assign individuals to different groups when possible. This is not possible when treatment groups must be organic units.
50. With small treatment groups, use pair-wise matching when individual assignment to different groups is possible and pre-experiment data on all individuals is available for matching purposes.
51. Use each group as its own control when random assignment is not possible. Each treatment group should use the new capability and the old capability.
52. Avoid giving the new-capability group "extra preparation" for the experiment which would create artificial group differences (trained group difference).
53. Monitor for differential player dropouts from the different groups over a long experiment to avoid evolving artificial differences between groups as the experiment progresses.
54. Establish a "no treatment" control group when players are assigned to a particular experiment group based on low (or high) scores. Because of "regression toward the mean" players with initial low scores will show an improvement upon subsequent retesting even if the experimental treatment is irrelevant to performance.
55. Monitor for "dominator effects" in small experiment groups where one individual may drastically influence the group score for better or for worse.
56. Monitor for "imitation effects" where one group will imitate the other group rather than respond to its own experiment treatment.
57. Monitor for "compensation effects" (John Henry effect) where individuals in less desirable or more strenuous conditions will push themselves harder to outperform those in the easier condition. If the less desirable condition is the baseline control group, their over-compensation may equal any potential improvement in the new-capability group.
58. Monitor for "resentment effects" where individuals in the less-desirable experimental condition may perform poorly as a result of being selected for this condition rather than the more desirable condition.

Threat 16: Data Collection Differences between Experiment Groups

59. Ensure that the new-capability group does not get all of the best instrumentation and most proficient data collectors.
60. Experimentation team, including the analysts, must continually scrutinize their own biases to ensure that their "experiment expectancies" do not bias the data collection and analysis.

Threat 17: Trial Conditions Differences between Experiment Groups

61. Execute the trials for each treatment group simultaneously (same day, same time, same location, same targets, *etc.*) to the extent possible. Experiments of detection systems allow simultaneous presentation of targets to all experiment groups.
62. When the different treatment groups cannot undergo their respective trials simultaneously, ensure that the trial conditions are as similar as possible, *e.g.*, same day, same time, *etc.*
63. When simultaneous trials are not possible, counterbalancing the trial sequence between two groups when possible (GP1-GP2-GP2-GP1) with Group 1 (GP1) as the new-capability group and Group 2 (GP2) the control group.
64. Monitor and report any differences in the experimental setting between groups.

3.5.5 Techniques to Counter Threats to Requirement 4: *Ability to Relate Results to Actual Operations*

Threat 18: Non-Representative Capability

65. Be aware of and report the strengths and limitations of surrogates and prototypes used in the experiment.
66. Surrogates with major limitations are encouraged early in the concept development cycle for the preliminary examination of the system's potential military utility, to help develop potential human factors requirements, and to influence design decisions. However, the limited capability to relate conclusions from prototype systems to production systems needs to be recognized and accounted for in later experimentation.
67. Use fully functional prototypes when experiments are used as the final event to decide if the new capability should be deployed to the operating forces.

Threat 19: Non-Representative Experimental Unit

68. Select experiment players directly from an operational unit that will eventually employ the capability.
69. Use students, retired military, or government civilians when operational forces are unavailable and the experimental task represents basic human perception or cognition.
70. Avoid the temptation to overtrain the experiment unit to ensure success. An overtrained experiment unit is unrepresentative and referred to as a “golden crew.”
71. Avoid undertraining by ensuring the unit is trained sufficiently to represent an experienced operational unit.
72. Explain the importance of the experiment to the players and their contribution to the effort to ensure the new capability can be thoroughly and fairly evaluated.
73. Monitor to ensure participants do not “under perform” out of lack of interest or resentment. This may occur when personnel are assigned to participate in the experiment as “an additional” duty and it is perceived to be unrelated to their real mission.
74. Monitor to ensure players do not “over perform” due to being in the spotlight of an experiment. This is known as the "Hawthorne effect." This effect is more likely to occur in highly visible experiments that have continual high-ranking visitors. In this instance, the players are motivated to make the capability “look good” to please the audience even though the capabilities may not be that effective.
75. Avoid inducing “experimenter expectancies” in the experiment groups where they perform according to the expectancies of the experimenter (also known as Pygmalion effect). If the experimenter expects the control group to do less well than the new-capability group, the control group may perceive this and perform accordingly.

Threat 20: Non-Representative Measures of Effectiveness

76. Measure simple objective effects (time, detections, rate of movement, *etc.*) with data collection instrumentation calibrated for precision (non-bias accuracy) by pilot testing instrumentation prior to the experiment.
77. Measure complex effects (information superiority, mission success, situational awareness, *etc.*) as the weighted or un-weighted composite score of concrete components that can be measured objectively.
78. Measure components of complex effects with alternative independent methods to avoid a “halo effect.”

79. Measure complex effects with overall subjective expert ratings.
    a. Estimate the objectivity of subjective ratings through inter-rater agreement of independent experts observing the same event.
    b. During training, have raters observe predetermined “good” and “poor” practice events to determine if their assessments differentiated.
    c. Increase confidence in the subjective ratings by correlating them to independently obtained objective component measures.
    d. Employ several raters independently and combine their individual scores into a single overall assessment.
    e. The veracity and generalizability of expert ratings rest on the operational experience and credibility of the raters.

Threat 21: Non-Representative Scenario

80. Ensure realistic Blue-force operations.
81. Develop realistic tactics, techniques, and procedures (TTP) for the new capability prior to the experiment.
82. Allocate sufficient time for training the experiment unit in appropriate tactics with the new capability.
83. Ensure a realistic scenario environment.
    a. Approximate the noise, confusion, fear, and uncertainty of combat where possible. Allow the experiment to continue for many hours or days to generate fatigue-associated stress.
    b. Allow for maximum free-play and sufficient scenario space and events to promote player uncertainty, player creativity, and sufficient opportunity to explore and attempt to solve the warfighting problem.
    c. Increase tactical realism of player participation by use of lasers to simulate battlefield engagements.
84. Ensure a realistic and reactive threat.
    a. Conduct field experiments at national training centers when possible because they can provide realistic, well-trained threats.
    b. Use threat experts from the national agencies to assist in designing the “future threat” in the experiment scenarios and to monitor the conduct of the threat during experiment execution.
    c. Allow the threat maximum free-play during the experiment to respond to and even preempt, if possible, Blue-force employment of the new experimental capability.
    d. The development and employment of an intelligent, determined opposing force is one of the best counters to the threat of non-representative scenarios.

Principle 4.

Defense experiments should be integrated into a coherent campaign of activities to maximize their utility

Principle 4 describes the needs for, and foundations of, integrated analysis and experimentation campaigns (designed, coherent sequences of experiments and other methods of knowledge generation) based on metrics according to the problem characteristics, complexity and definition.

Integrated analysis and experimentation campaigns provide a coherent framework for addressing capability development problems. A well-designed campaign will combine a range of diverse analytical methods, each with its own unique strengths and weaknesses. These are integrated in a manner to exploit their strengths, while providing coverage to help mitigate their weaknesses, akin to combining diverse systems into mission capability packages (see Principle 7). Results are related in a progressive manner to resolve the problems and increase confidence.

Campaigns include a management and communication framework, and an analytical program. The analytical program is conservative in the sense that it retains a problem formulation and analytical phase, but is radical in the sense that these stages are iterative and the campaign evolves based on the cumulative results of the analytical activities. Both phases are informed by individual analytical activities, experiments or other activities. Specific activities may also be included to initially decompose the problem and to integrate the results.

Principle 4. Defense experiments should be integrated into a coherent campaign of activities to maximize their utility

The reader is encouraged to apply and adapt this Principle. However, the examples herein are based on the specific perspective and experience of different lead-nation authors with contributions from other participants. Therefore they may require supplementary effort to relate them to national perspectives.

Principle 4 introduces the concept of an integrated analysis and experimentation campaign in which a large capability development problem is coordinated and managed under an analytical umbrella to design, manage and review the coordinated sequence of activities used to attack a particularly large issue.

4.1 Campaigns

In project management terms, an integrated analysis and experimentation campaign can be defined as:

"*A portfolio of projects designed to achieve a set of business objectives which benefit from a consolidated approach and where deliverables of each project are integrated into one overall program. These projects are likely to be linked both logically and by resources, they are likely to provide deliverables which are required by other projects and often, as projects are completed, this translates into a revised set of corporate objectives.*"[22]

This definition contains the essential elements of a campaign of analysis and experimentation in which the components, experiments and studies are considered as projects.

Campaigns use a mix of defense experiments and parallel studies to understand the problem's context, the associated warfighting concept and the capabilities required. The product of a campaign is advice to decisionmakers on the utility and versatility of the concept and the capabilities required to achieve the concept. Campaigns can be used to analyze issues at all levels from joint and combined operations to platforms and components.

The use of experimentation in helping decisionmakers to understand a particular problem is rarely a single activity. Typically a problem, at whatever level, is best addressed through a matrix of analytical tools and activities, where each activity provides information related to specific issues, context for subsequent activities and a comparison to previous work. An integrated campaign using a variety of techniques

[22] Consolidated from: http://www.e-programme.com/articles/proj_def.htm

ensures that weaknesses in one technique can be mitigated by others. Where information correlates between activities it increases confidence, where it diverges it provides guidance for further investigation. It is only when all activities are brought together in a coherent manner and the insights synthesized, that the overall problem under investigation is advanced as a whole.

Campaigns seek to set up a deliberate framework of activities with which to address a given issue or problem. Through careful design and management, a campaign should seek to resolve the issue under study in the most effective manner, ideally minimizing the resources and time expended in coming up with the solution. Thus we have the concept of a campaign being a carefully coordinated process, rather than a random or *ad-hoc* set of activities, that itself undergoes a rigorous process of design, management, execution and analysis, as would any individual activity within the campaign. Campaigns are characterized by both an analytical and management framework.

Such campaigns can address force development issues at any level. Here are two possible mappings among many of the problem space. The first one is used in Figure 27 below. The second one uses the following levels: technological (*e.g.*, systems of systems), tactical, operational, as well as strategic. As examples, in Australia:

- at the **technological** level, helicopter operations within a combined arms team; surface and sub-surface platforms for maritime operations; and the JSF within the air control system;
- at the **tactical** level, amphibious and airmobile task groups;
- at the **operational** level, the capability balance required to achieve the Future Warfighting Concept; and finally
- at the **strategic** level, the Effects Based Operations concept being developed in conjunction with many government agencies.

An ideal campaign will integrate the events conducted in all the levels. Ideally the campaigns run by various agencies will be integrated and the results iterate in both directions through the various levels. In this way, for example, experimentation at the strategic level provides the context and problem definition for subsequent analysis at the tactical level and the detailed results of experimentation and analysis at the technological level would constrain subsequent tactical level experimentation.

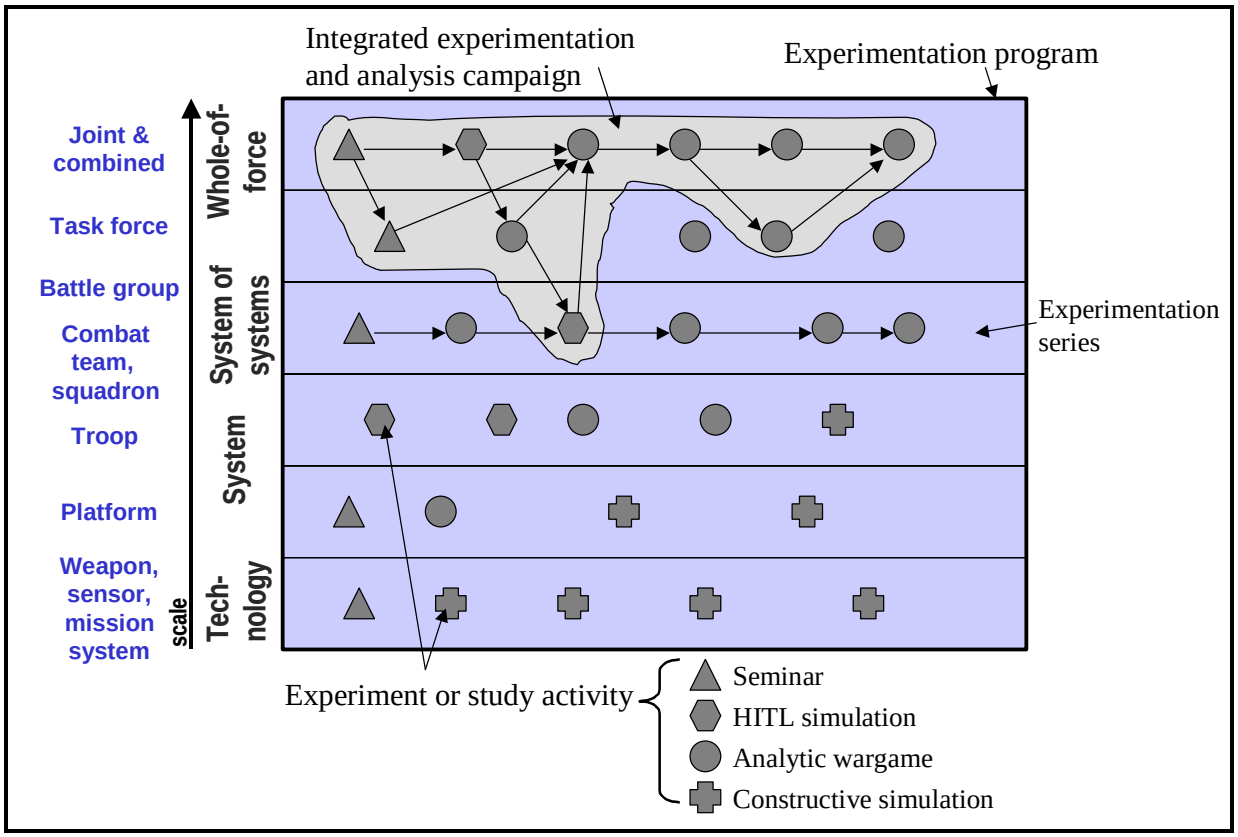


Figure 27 Australian example of campaigns[23]

The term campaign may be applied at any of the different levels, from the operational level, for example the Australian Army's Army Capability Management Plan; through capability-specific work addressing, for example, the introduction into service of the Armed Reconnaissance Helicopters (ARH) within Australia addressing issues such as the impact on a task force and battle group, C2, troop and squadron tactics, techniques, & procedures (TTPs); to the lower levels, for instance related to a given systems acquisition process in which the sponsorship and stakeholder membership is much simpler. At whatever level, an appropriate sequence of campaign activities, as illustrated in Figure 27 for the Australian Organisation, is required to address different aspects of a problem and to accumulate validity with regard to its conclusions.

[23] In Australia, the use of integrated analysis and experimentation campaigns (IAECs) is well established, and the use of the terms Program, Campaign and Series are used to distinguish between different kinds of IAEC, although the principles remain the same.

4.2 Foundations of Integrated Analysis and Experimentation Campaigns

Integrated analysis and experimentation campaigns must operate according to two key principles.

1. The choice of analytical tools should be dictated by the fidelity required for the problem to be addressed. In general, low-fidelity, low-resource models best address high-level "broad brush" questions, while higher fidelity tools, for example human-in-the-loop simulations, are more appropriate for more narrowly focused questions.
2. Experimentation that is focused on specific questions is more likely to yield useful insights than exploratory events. Moving to the point at which such specific questions can be framed must be *a priori*y of the process.

The basis for campaigns is **learning by doing**. Rather than conducting one-off experiments, the aim is to build up a rich understanding of future possibilities by "living and breathing them" over a period of time, using feedback gained to guide future paths.

Implicit in learning by doing is the concept of **building knowledge**. The notion of conducting **parallel studies** supports the learning by doing philosophy, but carries significant implications for staffing and timescales. Every effort must be made to **link past, present and future**—evaluation of the current force must feed into the exploration of future force options and concepts. Similarly, understanding of our past history can also help illuminate our future paths. The emphasis should be on **iterative** activities, where the results of one are fed into the design of the next in a rolling campaign of experiments and analysis activities.

Campaigns must encourage innovation. The emphasis will be on initially exploring innovative concepts rather than highly focused verification testing of well-defined concepts.

Campaigns must **be credibly relevant** and this is achieved through the involvement of decisionmakers and a warfighter partnership in the preparation, conduct and review of experiments. The outputs of experimentation must ultimately carry weight if they are to influence the wider debate.

Methodological power is achieved by understanding that **socio-technical systems** are being analyzed. It is fundamental that the wide definition of "system" is used, including human and organizational aspects in addition to the technical aspects. A toolbox of methods is therefore required as no single analytical/investigative technique is sufficient to generate credible results for these complex problems. Different methods and tools have different methodological strengths and weaknesses, such as seminar wargames, constructive, human-in-the-loop and abstract simulations. A philosophy of triangulation[24], should be used to examine particular topics across a number of

[24] Triangulation is borrowed from navigation terminology and means to reduce the area of error by taking three independent reference points to assess a position.

dimensions. Central to experimentation is a wargaming philosophy. Playing against an agile and intelligent enemy provides a more powerful learning environment than analysis alone can provide (as with constructive simulation without a smart reactive opposing force), and seminar and analytic wargames provide a good balance between the physical and psychological aspects of warfare at the expense of statistical rigor. This rigor is gained through iteration and the use of a range of tools.

4.3 Why Use a Campaign

An integrated analysis and experimentation campaign will be required for a variety of reasons. There may be resource or political reasons why a campaign is preferred to a single activity, although more often it will be because, without a coordinated campaign, the problem or issue under investigation simply cannot be satisfactorily resolved. A campaign allows the problem to be tackled in a coordinated, manageable manner with a variety of analytical techniques and allows a degree of iteration and synthesis between activities that help ensure that the overall problem is satisfactorily addressed. The problem may initially be ill-defined and a sequence of activities will allow assessment and adjustment as the problem is refined.

Some of the analytical reasons for using a campaign approach are described in the following sub-sections.

4.3.1 Problem Characteristics

The principal analytical reason for using experimentation is the nature of the problem. Problems may be described using two characteristics: system complexity and the nature of its internal systems interactions [Flood and Jackson 1991: p. 31-43]. Problems can be defined as either simple (few, well-defined components and a closed, constant overall system) or complex (ill-defined components and an open, evolving overall system). The nature of the internal interactions is defined as unitary (fully aligned with common objectives), pluralist (some divergence of "interests" but with common objectives), or coercive (no common interests or objectives) as indicated in Table 1.

	Unitary	Pluralist	Coercive
Simple	Machines	Coalitions	Prisons
Complex	Organisms, cybernetics	Cultures, commerce	Warfare

Table 1 Example of problem domains[25]

Defense problems that require an experimentation approach tend to be complex and coercive. The systems that provide solutions must also:

1. be adversarial, that is they operate against one or more systems (the adversary) that either directly oppose, contest, or compete with the goals of the first system.
2. be socio-technical; hence the components are ill-defined.

[25] Modified from the following reference [Flood and Jackson 1991].

3. operate in a wide range of physical environments such that the environment affects the systems' component characteristics, hence it is representing an open system.

The importance of an opposing force, itself a socio-technical system, means the system is coercive. The socio-technical nature of the system and the interaction between the components and the environment characterize the system as complex. If the problem presented to the analyst does not fulfill these criteria, then a campaign may not be an appropriate approach.

4.3.2 Increasing Confidence

A campaign allows a gradual build-up of the knowledge surrounding the problem or issue under investigation, leading to increased confidence that the findings are valid. In addition, a number of experimental activities increases the sample size, leading to more confidence in any statistically based outcomes. Finally, more activities allow more participants and more user engagement, again resulting in a greater degree of confidence in the outcomes.

4.3.3 Problem Complexity

Many problems that might be explored through experimentation are simply too complex to be dealt with in a single activity. A well-focused experiment will necessarily constrain a large number of variables in order to ascertain linkages between cause-and-effect. Complex problems may have far too many independent variables and ill-defined constraints to be handled in a single activity. A campaign permits the problem to be tackled in a multi-stage manner, so that individual elements can be explored in turn, before re-immersing the elements back into the wider context. Campaigns allow the results of single activities to be synthesized into meaningful advice across the entire problem.

4.3.4 Synthesis of Military and Analytical Skills

The key component of the process is to immerse human decisionmakers in an environment that challenges existing paradigms through the actions of an intelligent enemy. Within this environment, a synthetic operational experience is provided to the players and assessed through the After-action review or Report (AAR) in a similar manner to a "normal", or real operation, as well as providing a wide range of subjective and objective data. A campaign enables the application of many different techniques, generating opportunities for analytical and military skills to be applied to the problem.

4.3.5 Problem Definition

In a static strategic context, with a known operational concept, military judgment is usually sufficient for problem definition because of a deep, real-world, professional experience base. When the strategic environment is uncertain and unprecedented, and the impact of technology unknown, the experience base is usually too narrow to confidently conduct the problem definition. Within the campaign therefore we must build a "synthetic experience base" and the process of scientific inquiry is used to

increase our confidence in the problem definition. The selection of the experimental force and the conditions for its test are important products of the early stages of a campaign (problem formulation), because this will provide the new experience base for military judgment.

4.3.6 Tool Selection

A campaign requires a range of activities in addition to those required by its experiments alone (see Principles 3 and 7). One of the key stages of a campaign plan is to work out the most appropriate tool or method to tackle a given aspect of the problem under study. Methods available may include historical analysis, traditional operational analysis (OA) or operations research (OR) studies[26], and spreadsheet modeling or seminar activities. If experimentation is deemed appropriate for the particular stage of the problem, then a process of experimental design should be followed in order to select the most appropriate form of experimentation method, *i.e.*, field experiment, analytic wargame, constructive simulation or human-in-the-loop simulation. It is important to realize that the strengths of one method may be used to mitigate weaknesses in another, such that over a whole campaign, the 21 threats to experimentation (see Principles 2 and 3) can be managed.

4.3.7 Other Considerations

In an effort to coordinate major activities across significant periods, an integrated analysis and experimentation campaign plan must incorporate the following characteristics:

Identify decision points for which a body of knowledge is required. Analysis and experimentation efforts should be focused on providing specific knowledge in time for required decisions:

1. Determine the critical information requirements for each decision point.
2. Track the development of the required body of knowledge.
3. Plan the series of mutually supporting events that develop the necessary body of knowledge.
4. Establish standards and methods of enforcement for the conduct of analysis throughout all activities.
5. Establish standards and methods of enforcement for the selection of tools and the development of the technical environment to support analytical and experimental activities.

[26] The **R** (research) has been largely replaced by **A** (analysis). The Navy now talks more of **OA** (Operational Analysis) than **OR** with **OA** often being considered an adjunct to modeling and simulation. The result has been an emphasis on quantification and metrics at the expense of understanding the problem. Like so many other debilitating trends, this one developed largely in response to what **decisionmakers** have demanded. What we often have now is **advocacy analysis**, where much time and effort are spent to provide justification of a position or decision based on having more and **better** numbers and **metrics** than your critics. This often occurs by focusing on a very narrow slice through a problem that is often far removed from the true context of the overarching problem. http://www.strategypage.com/prowg/default.asp?target=or.htm

To establish clarity of purpose and execution, a campaign plan must:

1. Specify the objectives and intent of each event within the campaign plan and define the products required from that event.
2. Ensure each event meets the necessary analytical and technical standards.
3. Coordinate the use of scenarios and input data across all events.
4. Ensure results are properly analyzed and interpreted and devoid of institutional bias.
5. Ensure results are shared and products disseminated.

4.4 Campaign Analysis

Analysis within the campaign process should focus on bringing together all the discrete pieces of analysis that were generated by the activities within the campaign. It is a process of assembling the bigger picture from all the components that have resulted from the detailed experiments and other studies. This requires a coherent set of campaign level metrics within which each of the assembled pieces may be related. It is likely that the final campaign-level analysis involves a great deal of conjecture and assessment related to how all the component pieces fit together, and a final campaign-level finding will most likely contain a number of alternative proposals and findings, rather than a single, objective result. The campaign output is there to inform the decisionmaker, not to provide a single, irrefutable finding in itself.

4.4.1 Campaign Metrics

A campaign requires a coherent analytical framework across all activities within the campaign. In addition, there may be metrics imposed by the sponsor to determine the performance of the campaign process itself (as opposed to the study under consideration).

4.4.1.1 High-level MoM

Identification of high-level measures-of-merit (MoMs) should start with ideal measures of the desired benefits or effects before considering what can be practically generated by analysis (the latter may force the use of surrogate MoMs, but these must be clearly related to the desired measures).

A structured analysis of potential benefits[27] should be carried out as a basis for constructing appropriate MoMs. Mapping techniques, such as cognitive and causal mapping, (also known as influence diagrams), are a good way to express the various relationships within the problem space and to identify “chains” of analysis (*i.e.*, links among the independent variables and between the independent and dependent variables). These lead to a resultant structure in terms of independent and dependent variables, and hence to high-level MoMs.

[27] The structured analysis of benefits is a logical process that seeks causally to map lower-level MoMs that can be related to investments or other actions to higher-level MoMs that can be valued directly by decisionmakers.

4.5 Context and Scenarios

One of the early products developed within a campaign is the context that can drive the following stages of the process. The context itself will vary according to the level and scale of the issue under consideration, however the following sections may be typical of the context related to a high-level CD&E type problem.

The scenarios that may be applied throughout a campaign will then be derived from these contexts. It is not necessary to use the exact same scenario for each stage; indeed there are good reasons to vary the scenario throughout the campaign in order to develop a more generalized solution. However the bounds on the scenario should be consistent throughout the campaign in order to provide some degree of rigor and validity to the whole problem resolution process.

4.5.1 Context[28]

Military Context. The military context of the study includes geopolitical parameters that bound the problem space, such as:

1. The geographic, oceanographic, and climatic characteristics of the possible theatres of operation.
2. The possible effects required of, and constraints on, military operations and their possible consequences in the other domains of national power (diplomatic, economic and information).
3. Possible national and coalition partners, their goals and constraints.
4. Possible adversaries and the characteristics of their political, military, economic, social, information and infrastructure (PMESII) systems.

Analytic Context. The analytic context of the study includes:

1. Aim and objectives of the analysis, including the decisions to be supported,
2. Generic warfighting issues[29], and
3. Relevant previous studies.

4.5.2 Other Aspects[30]

Military Aspects. The military aspects of the problem include:

1. The concept of operations to achieve the national objectives.
2. The missions and tasks that must be undertaken to achieve the desired effects.
3. The mission capability packages required for the missions.
4. The operational conditions under which the capabilities must operate.

Analytic Aspects. The analytic aspects of the problem include:

1. Issues to be addressed (formulated problems and hypotheses);

[28] Adapted from [NATO 2002].
[29] Generic warfighting issues include key systems, doctrine, Tactics, Techniques, and Procedures (TTP), organizational structures, and key assumptions (*e.g.*, system performance parameters).
[30] Modified from [NATO 2002].

2. Assumptions and constants;
3. High-level MoM;
4. Independent variables (controllable and uncontrollable); and
5. Constraints on the values of the variables (domain and range).

Principle 5.

An iterative process of problem formulation, analysis and experimentation is critical to accumulate knowledge and validity within a campaign

Principle 5 argues for the criticality of an iterative process of problem formulation and analysis to accumulate knowledge and validity within an integrated analysis and experimentation campaign by bounding the problem, issues and assumptions.

Force development generates systems and capabilities to deal with problems that are, by their nature, complex and coercive in that they:

1. are adversarial: the military system operates against one or more systems.
2. are socio-technical; hence their components are ill-defined.
3. must operate in a wide range of physical environments by which they are affected.

Consequently the key aspect of the process, that of problem formulation, should aim to decompose force development problems into components that can be addressed with specific analytical techniques or studies (be they mathematical modeling or historical studies for example), or integrated analysis and experimentation campaigns. Events within individual defense experiments can then either be controlled and manipulated experimentally to isolate cause-and-effect or at the least observed without interference to establish associative relationships (as is often the case when training exercises are used for analysis).

Problem formulation is about decomposing the problem to the point that elements can be defined in terms of tasks, issues and analytical techniques to ensure appropriate techniques are employed. Campaigns should ensure that further research is organized and modified in a coherent manner by revisiting the deconstruction based on the information gained from each activity. In this sense problem formulation is never fully complete and the activities may change as the campaign progresses. Additionally the analysis continually accumulates validity and can provide information to decisionmakers at any stage of the process.

Principle 5. An iterative process of problem formulation, analysis and experimentation is critical to accumulate knowledge and validity within a campaign

The reader is encouraged to apply and adapt this Principle. However, the examples herein are based on the specific perspective and experience of different lead-nation authors with contributions from other participants. Therefore they may require supplementary effort to relate them to national perspectives.

5.1 Problem Characteristics

The principal analytical reason for using integrated analysis and experimentation campaigns is the nature of the problem as described in Section 4.3.1. Complex coercive problems require appropriate representation of their adversarial nature and dynamics, which can be achieved through the use of experimentation.

5.2 Problem Formulation

The initial stage of any campaign is problem formulation. Effective problem formulation is fundamental to the success of all analysis, but particularly at the campaign level because the problems are normally ill-defined, complex and coercive, involving many dimensions and a rich context. Problem formulation involves decomposition of the military and analytical aspects of the problem into appropriate dimensions. Decomposition cannot normally be achieved without detailed analysis using a matrix of tools such as seminar wargames and experiments, supported by analytical studies and operational experience. Detailed analysis also assists in the reconstruction of the problem segments and interpretation of results.

The problem formulation phase should identify the context of the study and aspects of the problem-related issues.

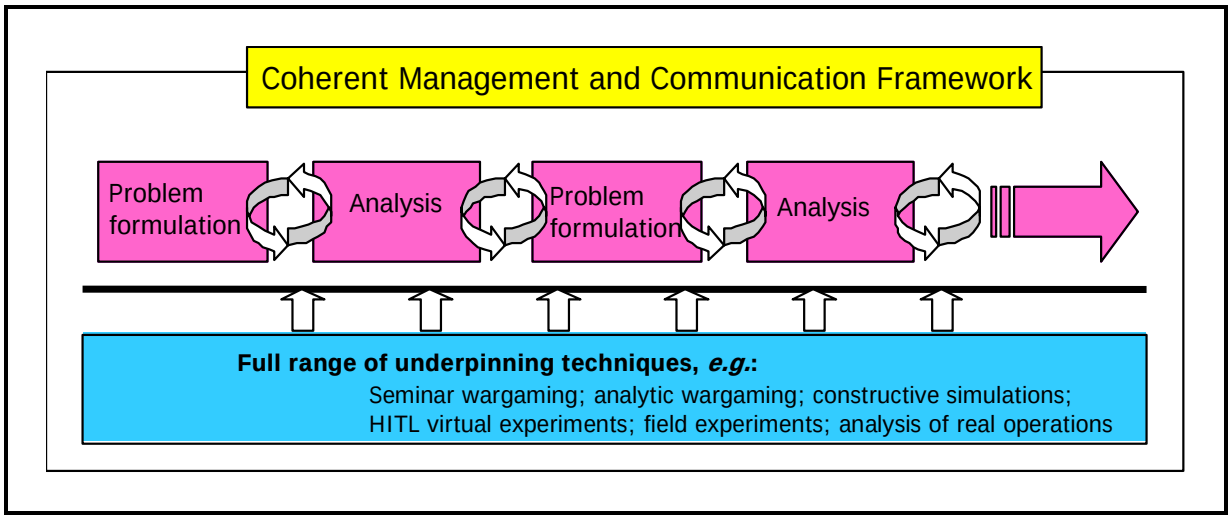


Figure 28 Problem formulation and analysis within a campaign

Figure 28 shows the role of problem formulation within an integrated analysis and experimentation campaign. The problem is being defined and refined throughout the entire campaign in an iterative cycle that never really completes until the campaign itself completes. The process of problem formulation and analysis undergoes constant review to reshape the direction of the campaign and to ensure that the real issue or concept is being addressed.

5.3 Problem Formulation Process[31]

Explicit problem formulation must precede construction of concepts for analysis or method selection. This is not a trivial exercise, especially at the campaign level. Problem formulation is the phase of the analysis that generates the hypotheses for subsequent analysis. However problem formulation does not end there; it is constantly reevaluated and reassessed during the campaign to ensure that the objectives are met.

The principles of explicit problem formulation are:

1. Proper resourcing of problem formulation activities will improve the overall efficiency and quality of the campaign.
2. A key risk in designing campaigns is allowing the problem formulation process to focus prematurely on subsets of the problem because they are: a) interesting; b) familiar; c) pre-judged to be critical; or d) explicitly requested by the customer. This requires great discipline by the study team, especially where the team's previous experience is biased in favor of particular parts of the problem space. The assessment team needs access to subject matter experts (SMEs) from a broad range of disciplines (*e.g.*, social scientists, historians, and regional experts in operations other than war (OOTW) assessment).
3. An understanding of the decisions to be supported by the analysis and the viewpoints of the various stakeholders (*e.g.*, customers, users, and suppliers) is essential to clarifying campaign issues.

[31] Modified from [NATO 2002].

4. A careful review of previous work must be carried out as a valuable source of ideas, information, and insight. This review should also serve to identify pitfalls and analytical challenges.
5. Problem formulation must not only provide problem segments amenable to analysis, but also a clear and valid mechanism for meaningful synthesis to provide coherent knowledge about the original, larger problem. The formulated problems (hypotheses) are thus an abstraction of the real problem that can be defined in terms of dependent variables that relate to this real problem and coherent settings for the independent variables that can be interpreted in terms of decisions and actions by the customer.
6. Problem formulation must be broad and iterative in nature, accepting the minimum of *a priori* constraints and using methods to encourage creative and multi-disciplinary thinking. It must be recognized that change is inevitable in many dimensions (*e.g.*, understanding of the problem, requirements, technologies, co-evolution of concepts of operation, command concepts, organization, doctrine, and systems). Thus the assessment process must anticipate and accommodate such change.
7. Campaign-level problem formulation must look beyond the next experiment or activity to the overall campaign goals, and not focus just on the immediate study. It is formulating a set of hypotheses and questions that together can answer the bigger issues under study. A separate process of problem formulation should occur within each experiment or activity focused on that specific phase of the problem (see Principle 4).

5.4 Issues in Problem Formulation[32]

5.4.1 Bounding the Problem / Issues and Assumptions

In dealing with fuzzy or uncertain boundaries, the problem formulation process needs to explore and understand the significance of each boundary before making (or seeking from customers) assumptions about it. This involves keeping an open mind, during the early stages of problem formulation, about where the boundaries lie and their dimensional nature. This is difficult because it makes the problem modeling process more complicated. A call for hard specification too early in the problem formulation process must be avoided. In the end, of course, the problem must be formulated in order to solve it, but formulation should be an output from the first full iteration, not an early input to it.

The problem may be formulated from multiple perspectives, each with different boundaries, some overlapping, and thus embrace the richness and complexity of the problem at hand. Importantly, the broader problem context in which these perspectives reside must be understood and represented so as to justify the selection of supposedly important elements of the problem. Such a mapping will also assist with the interpretation of results out of the analysis stage of the problem.

In formulating the problem, we are trying to bound a complex system. This is partly a process of understanding boundaries that exist in reality (*e.g.*, mission statements and geographical areas) and partly imposing artificial boundaries in order to illuminate the structure of the problem and constrain the scope of the analysis. To avoid the trap of over-specification, boundaries (especially self-imposed ones) should be kept porous,

[32] Modified from [NATO 2002].

allowing for cause-and-effect chains to flow through the external environment of the portion of the complex system that the boundaries define.

5.4.2 Problem Formulation Tools

It is useful to identify, develop (if necessary), and apply appropriate tools to support problem formulation. Representative tools and techniques include: techniques for supporting expert elicitation, influence diagrams, causal maps, system dynamics models, and agent-based models.

Wargames, and in particular seminar wargames, have an important role in problem formulation. In wargaming it is possible to balance the physical and psychological aspects of the problem by using warfighters as the players and adjudicating their actions using simulations. Most importantly wargaming introduces an adversary early in the problem formulation process, providing a stressful environment to explore the concept and develop the hypotheses for subsequent analysis. Although human-in-the-loop simulations and live simulations also introduce a human adversary, they are frequently too expensive and unwieldy for the problem formulation phase.

Tools and approaches used for problem formulation must be consistent with other tools and techniques likely to be considered for the subsequent analysis in order to produce a sensible **multi-methodology** approach to the entire problem and its solution.

Principle 6.

Campaigns should be designed to integrate all three scientific methods of knowledge generation (studies, observations and experiments)

Principle 6 advocates the integration of all three scientific methods of knowledge generation of GUIDEx. This requires additional planning for experiment design, execution and analysis, combined with campaign definition, and analysis for studies and supplementary observations. However it maximizes the quality of results from a campaign.

In about 400 BC, Plato, and other philosophers investigated the meaning of knowledge and the means to obtain it. Their method was primarily a *rational-deductive* process. Later Ptolemy and Copernicus focused on precise observations and explanations of the stars. Their methods were *empirical-inductive*, however, they were not experimenters. When scientists turned from the heavens to investigating earthly objects, they uncovered a new paradigm for increasing knowledge. Since they could manipulate those objects, new answers to questions about them were obtainable (See Principle 1). Francis Bacon and Galileo pioneered *experiments* to answer the question "If I do this, what will happen?"

Campaigns should be designed to integrate all three avenues to knowledge generation: *rational-deductive*, in the form of studies, in particular operations research and historical research; *empirical-inductive*, in the form of precise observation of real-world events in particular operations and exercises; and *experiments*, manipulation of events to isolate cause-and-effect. This guide principally addresses experiments and their role within capability development. Study and observational techniques are not included in this guide except to show their role in the overall campaign. Numerous other documents and books are available on the conduct of studies, operations research, historical studies and observational techniques.

Principle 6. Campaigns should be designed to integrate all three scientific methods of knowledge generation (studies, observations and experiments)

The reader is encouraged to apply and adapt this Principle. However, the examples herein are based on the specific perspective and experience of different lead-nation authors with contributions from other participants. Therefore they may require supplementary effort to relate them to national perspectives.

6.1 Formulating a Campaign Plan

Historically there have been three broad methods of accumulating knowledge. The rational-deductive (studies) approach, or pure logic (logic), without reference to the real world, practiced by Socrates and Plato; the empirical-inductive (observations) which focuses on precise observation of the real world, practiced by Ptolemy and Copernicus; and the empirical-deductive (experiments) where objects are manipulated and measured, introduced by Francis Bacon and practiced by Galileo who pioneered *experiments* to answer the question "If I do this, what will happen?"

Given the scope of GUIDEx, studies and observational techniques [Rosenbaum 2002] will not be discussed further in this document. However, for an extensive guide to empirical observational, and measurement methods in defense, readers are referred to the ABCA Analysts Handbook [ABCA 2004], which grounds these approaches within the military exercise context.

The aim of an integrated analysis and experimentation campaign is to integrate a range of warfighting knowledge generation methods, from analytical studies (rational-deductive), to operations observations (empirical-inductive)[33], up to experimental measurements (empirical-deductive), into a coherent package that addresses a complex capability development problem. The phases of campaign design are the same as for any evaluation, which are problem formulation and analysis. The complexity is that after the completion of each activity the problem formulation is reassessed and adjusted and subsequent activities may be redesigned. Additional planning for experiment design, execution and analysis combined with problem definition and analysis for studies must be integrated. As a result a campaign plan is a flexible instrument, with a supporting risk-management framework and an iterative approach to constantly review and reshape the remainder of the campaign to ensure that the overall goals are achieved (Figure 29).

[33] Induction allows the generation of new hypotheses to be tested.

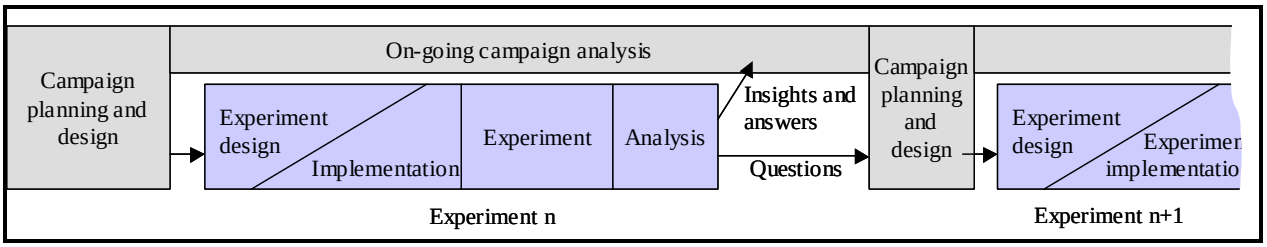


Figure 29 Campaign stages

Both the level and scope of planning in campaigns are quite different from those of experiment design. What distinguishes an experiment from any other form of study activity is the requirement to generate some link between cause-and-effect (see Principle 1). Thus one view of the process of generating an integrated analysis and experimentation campaign plan is to come up with a range of possible causes that relate to the problem under study and to attempt to align these with measurable effects that can then be studied in an experiment.

This naïve view of a campaign plan does not take into account the reality that a campaign is likely to include activities other than defense experiments. In all likelihood, seminars, workshops, historical analysis, and the like, will also be required as part of the campaign to support and help inform the experiments that will ultimately address the overall question. The campaign plan process must take these other activities into account within its design phase. The ultimate aim is to synthesize the outputs from all activities into coherent advice to the decisionmakers. Figure 30 shows an example campaign, the Australian RTA trials from 1997, which consisted of a whole variety of analytical methods, including experiments.

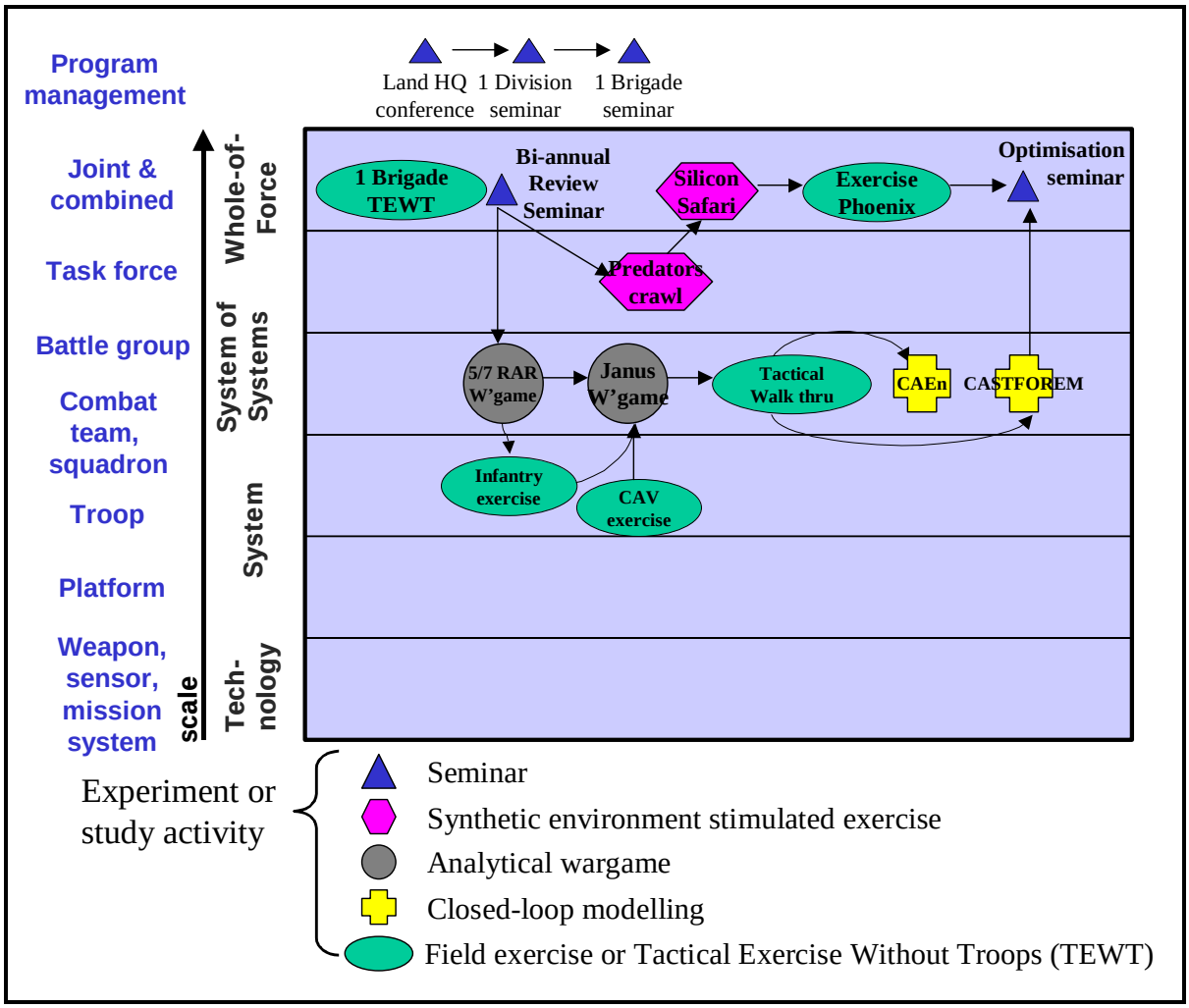


Figure 30 Example of a campaign: Experimentation Program, RTA 1997

The initial stages of planning a campaign are concerned with defining the overall issue or problem that the campaign is to address. Typically a sponsor has sought a campaign to help inform a high-level policy decision. The first stage of the campaign design is to investigate the details of the problem in order that it can be broken down into tractable-sized chunks. These chunks will then begin to form the outline of the set of experiments or other activities within the campaign. In undertaking this process, it is often a good idea to begin with an initial broad-brush study that will cover the problem space at a reduced level of detail to help identify those specific issues, scenarios and tasks that warrant further study. The campaign would then proceed through a number of activities, each focused on a single aspect of the overall problem space. At some stage it may be necessary to bring ideas and findings from these studies together and a larger regrouping activity might be employed to identify how the concepts develop when combined. Toward the end of a campaign a large-scale activity is typically used to both validate overall findings, and also to provide an opportunity to demonstrate the outcomes to stakeholders. Thus a typical campaign may resemble that shown in Figure 31.

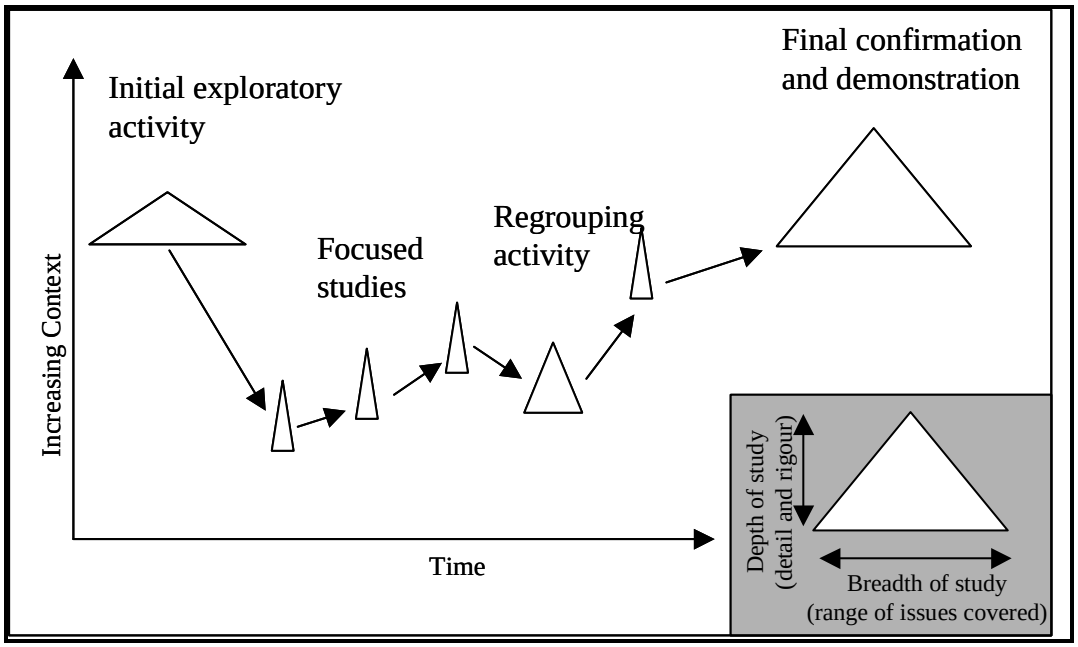


Figure 31 A typical campaign

Thus the process of campaign design typically involves a number of experimentation and study activities in order to help scope and refine the issues that the campaign is addressing.

The role of the campaign designer is to constantly re-evaluate the progress of the activities to ensure that the appropriate outcomes are generated. At each stage, the progress of the campaign is reassessed to ensure that it is heading in the correct direction. Each single activity within the campaign is not only generating some analytical answer that forms a piece of the overall puzzle, but is also a stage in the problem definition process to make sure that all the puzzle pieces will be generated by the end of the campaign. Thus it is common, and indeed expected, for a campaign to require redirection and refocusing throughout its course. It is through this process that the stakeholders can build confidence that the campaign has indeed explored sufficient options and conditions to give the study findings the required degree of rigor.

6.1.2 Method Selection

One critical component of campaign design relates to allocating elements of the problem to appropriate methods or methodologies of solution, as part of a general strategy to accumulate validity. Problem appreciation necessarily involves allocation of sub-elements of the problem (without wishing to imply a reductionist approach) to suitable techniques. Figure 32 illustrates the notion that the processes of the generation of suitable “bite-size” chunks and their allocation to techniques are intimately linked. However, only the question of allocation is explicitly discussed here.

Total Systems Intervention (TSI) [Flood and Jackson 1991], advocates the characterization of problems into six types (via two orthogonal axes (as shown in Table 1, page 111): simple or complex to represent the problem itself; and unitary, pluralistic

and coercive to represent the problem's stakeholder environment). A methodology similar to TSI was trialed to inform the choice of techniques (such as seminar wargames, constructive simulation, and field events) that might be used for specific elements or aspects of the total problem.

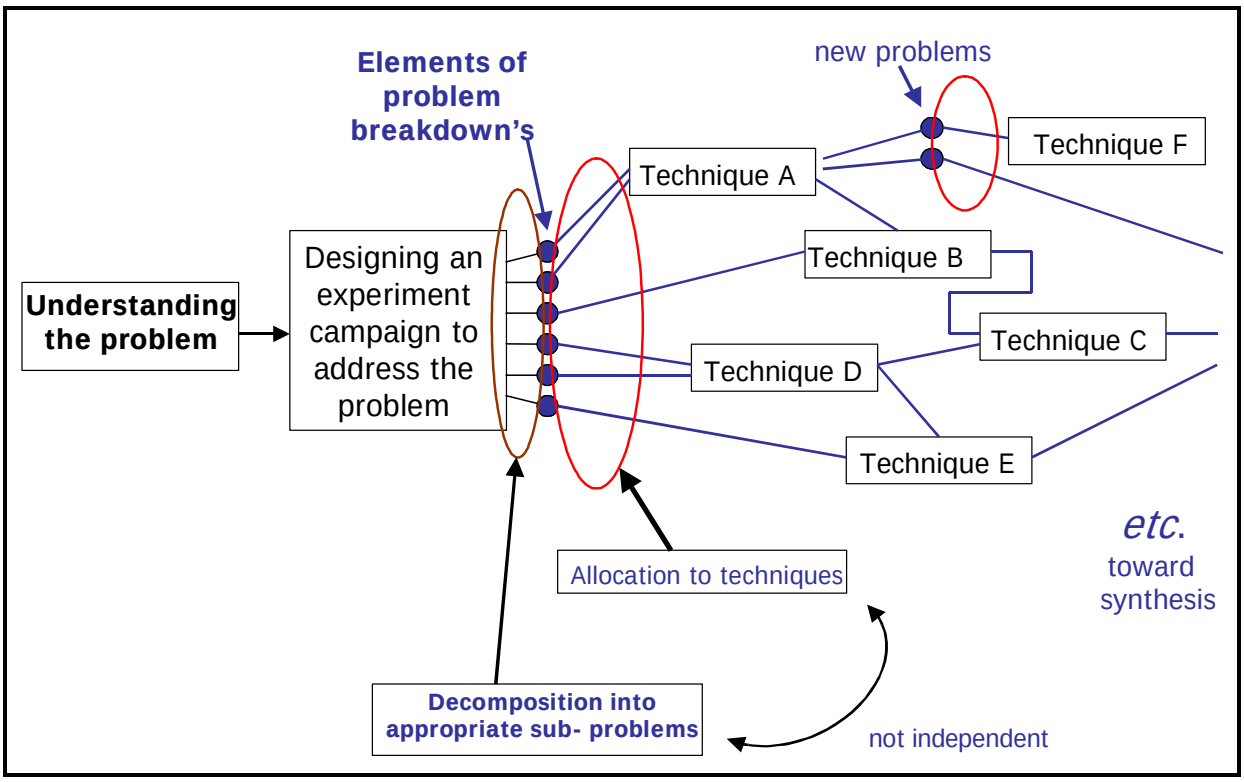


Figure 32 Campaign deconstruction

The intention is not to be prescriptive with regard to the choice of specific techniques, but to inform decisions with regard to the options. In fact, current thinking in the systems and operations research fields advocates the use of multiple, hybrid or multi-methodologies, very much akin to the thinking behind integrated analysis and experimentation campaigns, as a means to build on strengths and mitigate weaknesses. Based on consensus views gained from surveys of the experimentation practitioner communities, the sub-domains of "strength" (with regard to applicability and validity) were identified for the various techniques available, to go beyond the issue of whether or not a technique could be used in a given context.

The proposed methodology (abstracted in Figure 32 and Figure 33) extends TSI by characterizing outputs as well as inputs. In other words, it also provides advice on whether a technique is likely to meet requirements with respect to the nature of results.

The proposed method discriminates problem types according to:

1. **scale**—from sub-entity/platform level through teams to national/international,
2. **complexity**—from unambiguous cause-effect, to multi-level, highly interdependent (tangled) problems,
3. **clarity**—from loosely defined, poorly understood, to well-posed and unambiguous problems, and

4. **scope**—from a problem in which 95% of the solution is fixed and defined, to one in which there is great freedom to co-evolve all inter-related systems.

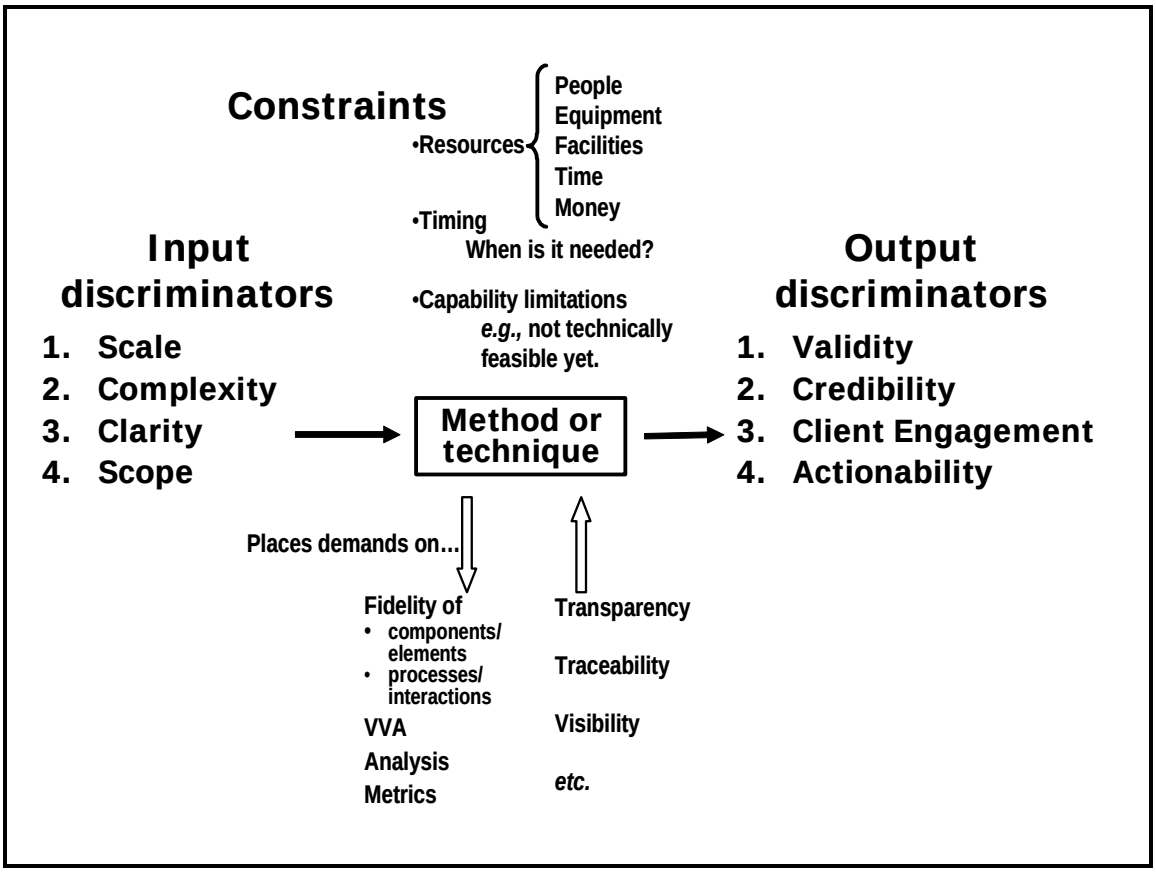


Figure 33 Proposed model for characterizing experimentation methods or techniques

It discriminates outputs according to:

1. **validity** (or veracity)—which represents the spectrum of required results from "quick and dirty" through expert judgment, sensitivity-tested, to full operational testing,
2. **credibility**—which represents the quality (validity) of the representation of stressors, or how realistically and rigorously the system was stressed,
3. **client engagement**—which represents the level of solution ownership held within the stakeholder group based on their active involvement in generating solutions (through intermediate steps), and
4. **actionability**—which represents the spectrum from the case where results inform further experimentation to the other extreme in which results impact directly on decisions within the client community.

Figure 33 shows that the model allows for decisions about the utility of a technique based on its feasibility with respect to: time and resources requirements on further development of experimentation capability; satisfying verification, validation, accreditation and analysis (VVA) needs; and the requirements of transparency and traceability.

This set of discriminators was reviewed alongside administration of surveys to various practitioner groups. While there is no requirement that the discriminators be independent (in fact some overlap is desirable) the final model was ultimately a compromise between the desire to map or represent inputs and outputs, and the practicalities of asking survey subjects to rank the strengths and weaknesses of the list of techniques. From initial survey results, it appears that there is some consensus, within expert practitioners regarding particular techniques, on the “signatures” of those techniques in terms of the problems they are well suited to deal with, and the kinds of results they deliver well.

Principle 7.

Multiple methods are necessary within a campaign in order to accumulate validity across the four requirements

Principle 7 shows how understanding of the four experiment validity requirements detailed in Principle 3 is essential to understanding the strengths and weaknesses of the primary methods used for defense experiments:

1. The strength of experiments using **analytic wargames** resides in the ability to detect any change in the wargame outcome, provided there are major differences in the strategies used. Additionally, to the extent that operational scenarios are used and actual military units are players, such wargaming may reflect real-world possibilities. A major limitation is the inability to isolate the true cause of change because of the myriad of differences between playing two different campaigns against a reactive threat.
2. Experiments conducted using **constructive simulations** allow repeated replay of the same battle under identical conditions while systematically varying capabilities, tactics employed, or levels of threat. Experiments using constructive simulations with multiple runs are ideal for detecting change and isolating the cause of that change. Because modeling complex events requires many assumptions, such as valid models of human behavior, critics may question the applicability of results to operational situations.
3. **Human-in-the-loop simulations** represent a broad category of real-time simulations with which humans can interact. In human-in-the-loop experiments, military subjects receive real-time inputs from simulations, make real-time decisions, and direct simulated forces or platforms against simulated threat forces. The use of actual military operators and staff allows the experiment designer to better reflect warfighting decisionmaking than experiments conducted purely with constructive simulations. However, once humans make decisions, variability increases, making it more difficult to isolate the reason for changes.
4. **Live simulation** is conducted in the actual environment, with actual military units and equipment and with operational prototypes. Usually only weapon effects are actually simulated. As such, the results of experiments in these environments, often referred to as **field experiments**, are highly applicable to real situations. Good field experiments, like good military exercises, are the closest thing to real military operations. A dominant consideration however, is the difficulty in isolating the true cause of any detected change since field experiments include much of the uncertainty, variability, and challenges of actual operations; in addition they are seldom replicated due to costs.

The best strategy is to construct an integrated analysis and experimentation campaign using multiple methods so that the weaknesses of any one method are compensated by the strengths of another. This provides the strongest case of accumulated validity in a campaign.

The model-exercise-model (M-E-M) paradigm is a special case of employing multiple methods to increase rigor. On the one hand it explicitly integrates the strengths of constructive simulation (*i.e.*, “model”) and, on the other hand, any of the methods that involve human interaction (*i.e.*, “exercise” in a generic sense). This technique is especially useful when resource constraints prohibit conducting side-by-side baseline and alternative comparisons during wargames and field experiments.

Principle 7. Multiple methods are necessary within a campaign in order to accumulate validity across the four requirements

The reader is encouraged to apply and adapt this Principle. However, the examples herein are based on the specific perspective and experience of different lead-nation authors with contributions from other participants. Therefore they may require supplementary effort to relate them to national perspectives.

The real power of the four requirements and the 21 threats to validity, arranged under these requirements presented in Principles 2 and 3, is that they allow the experiment designer to understand and evaluate the strengths and weaknesses of different defense experiments. All experiments have strengths and weaknesses. There is no such thing as a perfect experiment, in the laboratory or in the field. Knowing the strengths and weaknesses of particular experiments **in advance** of experiment execution allows the experimenter to decide which experiment strengths are important for a particular experiment. It also allows the experimenter to more realistically apprise the "stakeholders," those with an interest in the experiment outcome, of what any one particular experiment will return for their investment. Defense experiments can provide a wealth of empirical support for transformation decisions, but no single experiment can do it all, as this section explains.

7.1 No Such Thing as a Perfect Experiment

Any framework for organizing our lessons learned on good design techniques must bear sufficient level of practicality to be useful. Consequently, this section will discuss pragmatic implications of the four-requirement experiment validity framework presented in the preceding sections.

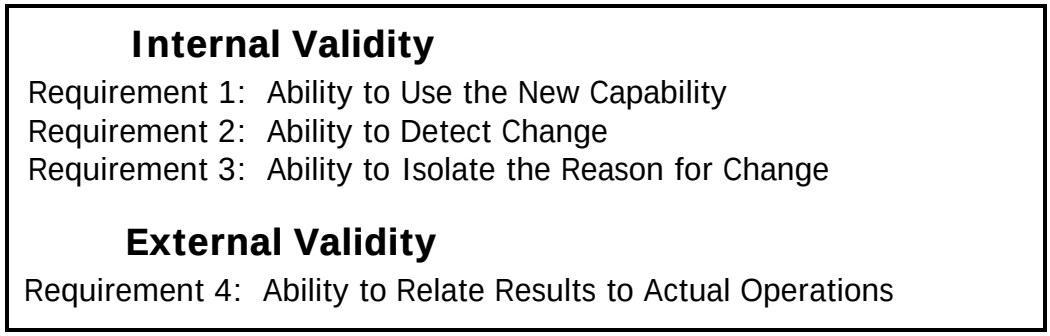


Figure 34 Classification of the four requirements in terms of validity

The first three experiment validity requirements represent the internal validity (Figure 34) of the experiment, the ability to determine if a causal relationship exists between two variables. The fourth requirement represents the external validity of an experiment,

the ability to generalize the cause-and-effect relationship found in the experiment environment to the operational military environment.

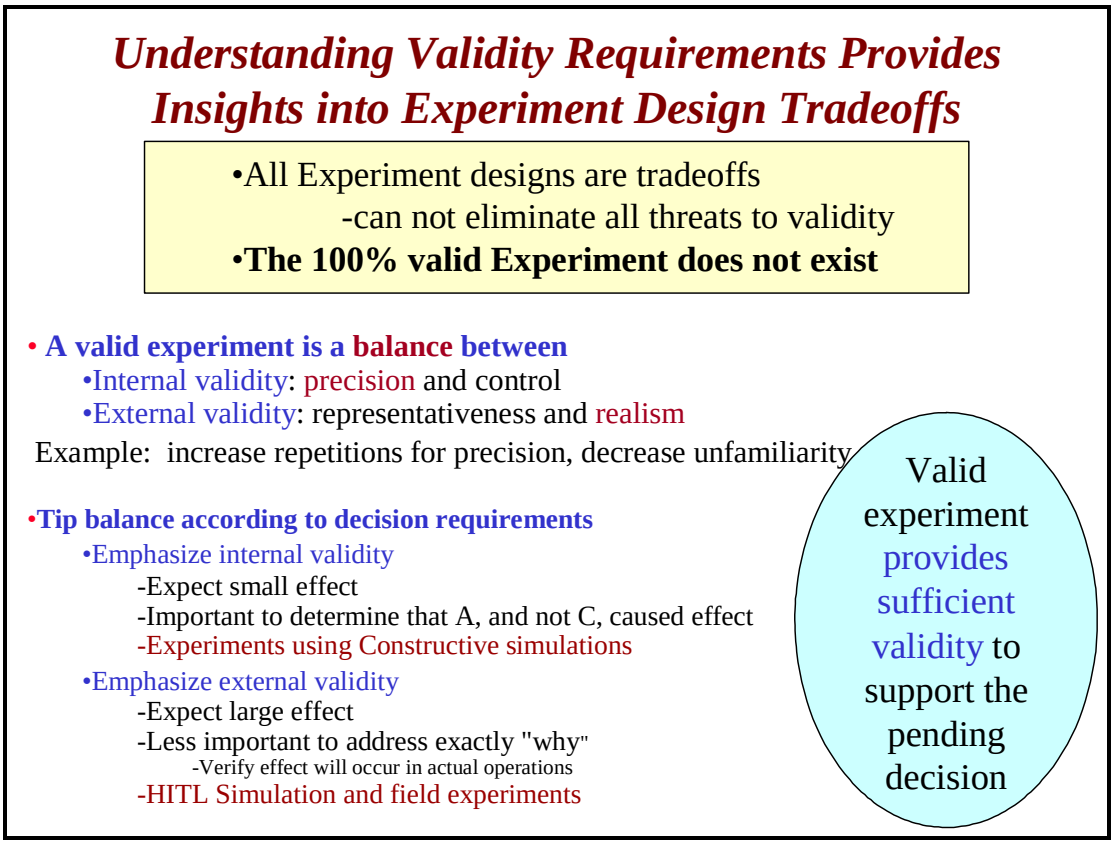


Figure 35 Design tradeoffs for valid experiments

One of the first implications of these four experiment validity requirements is that 100 percent validity is not achievable. The four experiment validity requirements cannot be fully satisfied in one experiment. Satisfying one, often 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 internal and external validity requirements (Figure 35).

Precision and control increase internal validity (ability to detect and isolate change) but often lead to decreases in external validity (ability to relate results to actual operations). Experiments that emphasize free-play exercises and uncertainty in scenarios, represent conditions found in real operations and thereby satisfy Requirement 4, the ability to generalize, *i.e.*, relate results to real operations. Experiments emphasizing control of trial conditions and sample size can satisfy the internal validity Requirements 2 and 3, the ability to detect and isolate change.

The idea that there are no 100% valid experiments and the presentation in Principle 3 of a long list of good experiment techniques to support the four experiment validity requirements may make it appear that defense experimentation is too hard.

Shadish [Shadish *et al.* 2002], however, wrote that experimenters need to be cognizant of validity tradeoffs and explicit about priorities when designing experiments.

> "This [long list of validity threats] might lead readers to wonder if any single experiment can successfully avoid all of them. The answer is no. We cannot reasonably expect one study to deal with all of them simultaneously, primarily because of logical and practical tradeoffs among them that we describe in this section. Rather, the threats to validity are heuristic devices that are intended to raise consciousness about priorities and tradeoffs, not to be a source of skepticism or despair. Some are more important than others in terms of prevalence consequences for quality of inference, and experience helps the researcher to identify those that are more prevalent and important for any given context. It is more realistic to expect a program of research to deal with most or all of these threats over time. Knowledge growth is more cumulative than episodic, both with experiments and with any type of research. However, we do not mean all this to say that single experiments are useless or all equally full of uncertainty in the results. A good experiment does not deal with all threats, but only with a subset of threats that a particular field considers most serious at the time" [Shadish *et al.* 2002: p. 96].

Experiment priorities can differ. 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.

7.2 The Importance of Requirement 3—Ability to Isolate the Reason for Change

In most defense experiments, indeed in most experiments of any kind, a case can be made for special attention and consideration to satisfying Requirement 3. The **ability to isolate the reason for change** can be considered the *sine qua non* (**necessary reason**) of conducting an experiment [Shadish *et al.* 2002: p. 99]. Resolving the "cause-and-effect" clause is essential to interpreting the experiment. If one cannot ascribe the observed change to some cause with some degree of certainty, the experiment is uninterpretable.

> "That is, do an experiment and have no interest in internal validity (cause and effect) is an oxymoron. Doing an experiment makes sense only if the researcher has an interest in a descriptive causal question, and to have this interest without a concomitant interest in the validity of the causal answer seems hard to justify" [Shadish *et al.* 2002: p. 99].

Internal validity, especially Requirements 2 and 3, *i.e.*, detecting a change and isolating the reason for change, clarifies why a specific level of performance that was observed is critical to all defense experiments. A very realistic field test may be conducted; but in the end, if the experimenter cannot, with some degree of assurance, make a case for or against the new capability, then the experiment can turn out to be an expensive training exercise for the player units. A case for a capability can be made when

something different happens in an experiment and this difference is solely due to the introduction of the new capability.

To ensure sufficient level of Requirement 3 validity, some operational realism may need to be sacrificed. In an evaluation of a new gas mask for tank crews, for example, a data collector may replace one of the crewmembers, such as a loader. While this detracts from crew integrity, it provides data for evaluating the mask's effectiveness at specific times during operations. Similarly, a scenario calling for continuous tactical operations may have to be interrupted periodically to check and realign data-collection instrumentation. In a final example, to ensure that two player units are at similar levels of proficiency in a multiple-group design experiment, one unit may require more training to equal the other unit, even though all units are not equal in the operational forces.

The point of these examples is to illustrate that Requirement 3, ability to isolate the reason for change, is most often the critical reason for conducting defense experiments. This is not to say that the other requirements never rise in importance. The next two sections will show that they do. It is critical to reach a balance and every effort should be made to minimize the impact of increasing one requirement over any of the other three.

7.3 Rigorous Experimentation Requires Multiple Methods

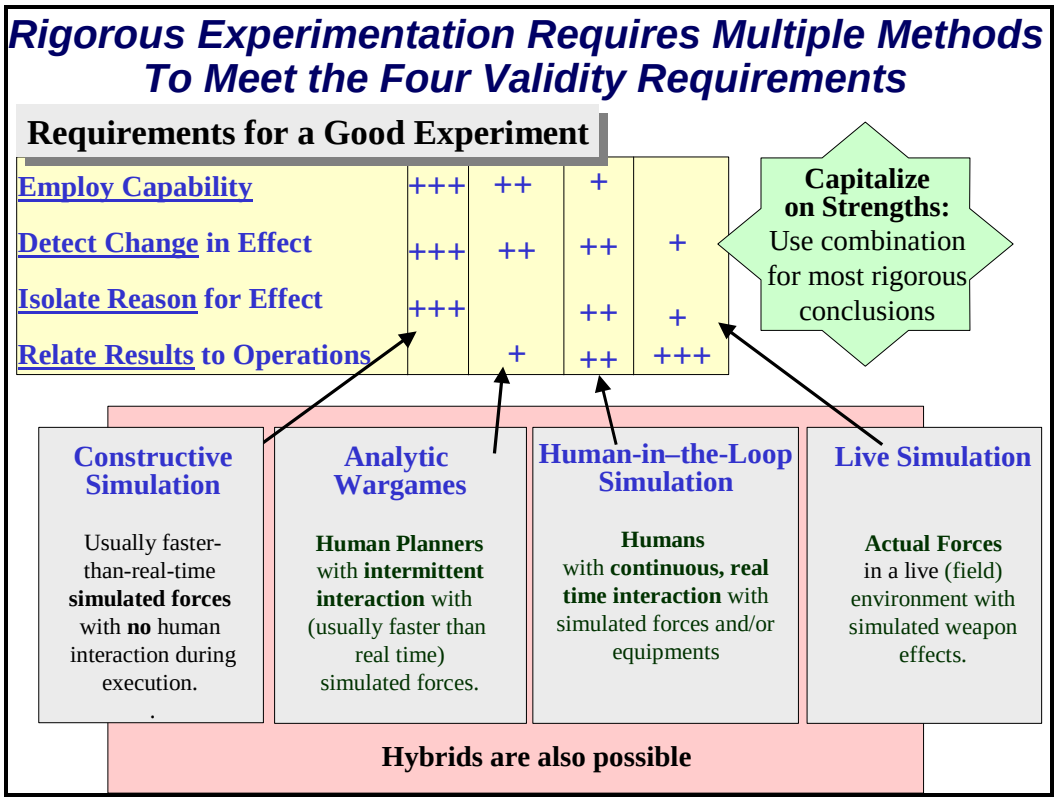


Figure 36 All experiment campaigns must strive for a balance among the four experiment validity requirements.

Most defense experiments use some form of simulation, which can be grouped into one of four general methods, as illustrated above: *constructive simulation, analytic wargames, human-in the-loop simulation, and live (field) simulation*. Each of these four methods has its own strengths and weaknesses with respect to the four experiment validity requirements discussed previously. Since one particular method cannot satisfy all four requirements, an integrated analysis and experimentation campaign requires multiple methods.

Constructive simulations are those in which no human intervention occurs in the play after designers choose the initial parameters and then start and finish the simulation. Constructive simulations are a mainstay of military analytical agencies. They allow repeated replay of the same battle under identical conditions, while systematically varying parameters—the insertion of a new weapon or sensor characteristic, the employment of a different resource or tactic, or the encounter of a different threat. Experiments using constructive simulations with multiple runs are ideal to detect change and to isolate its cause. Because modeling complex events requires many assumptions, including those of variable human behavior, critics often question the applicability of constructive simulation results to operational situations.

Analytic wargames 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 a simulation to help determine the outcome. A typical analytic wargame might involve fighting the same campaign twice, using different capabilities each time. The strength of such wargames for experimentation 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, analytic wargames 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.

Human-in-the-loop simulations represent a broad category of real-time simulations with which humans can interact. In a human-in-the-loop defense experiment, military subjects receive real-time inputs from the simulation, make real-time decisions, and direct simulated forces or platforms against simulated threat forces. The use of actual military operators and staffs allows this type of experiment to reflect warfighting decisionmaking better than experiments using purely constructive simulation. However, when humans make decisions, variability increases, and changes are more difficult to detect and consequently to attribute to the cause.

Live simulation is conducted in the actual environment, with actual military units and equipment and with operational prototypes. Usually only weapon effects are actually simulated. As such, the results of experiments in these environments, often referred to as field experiments, are highly applicable to real situations. Good field experiments, like

good military exercises, are the closest thing to real military operations. A dominant consideration however, is the difficulty in isolating the true cause of any detected change since field experiments include much of the uncertainty, variability, and challenges of actual operations; in addition they are seldom replicated due to costs.

7.4 Emphasizing Different Experiment Validity Requirements during Concept Development

Since no single experiment will totally satisfy all four experiment validity requirements, a comprehensive analysis and experimentation campaign should include a series of individual successive activities that emphasize different experiment validity requirements. As potential capabilities advance through the concept and prototype development stages, the following considerations are useful in selecting which experiment validity requirements to emphasize.

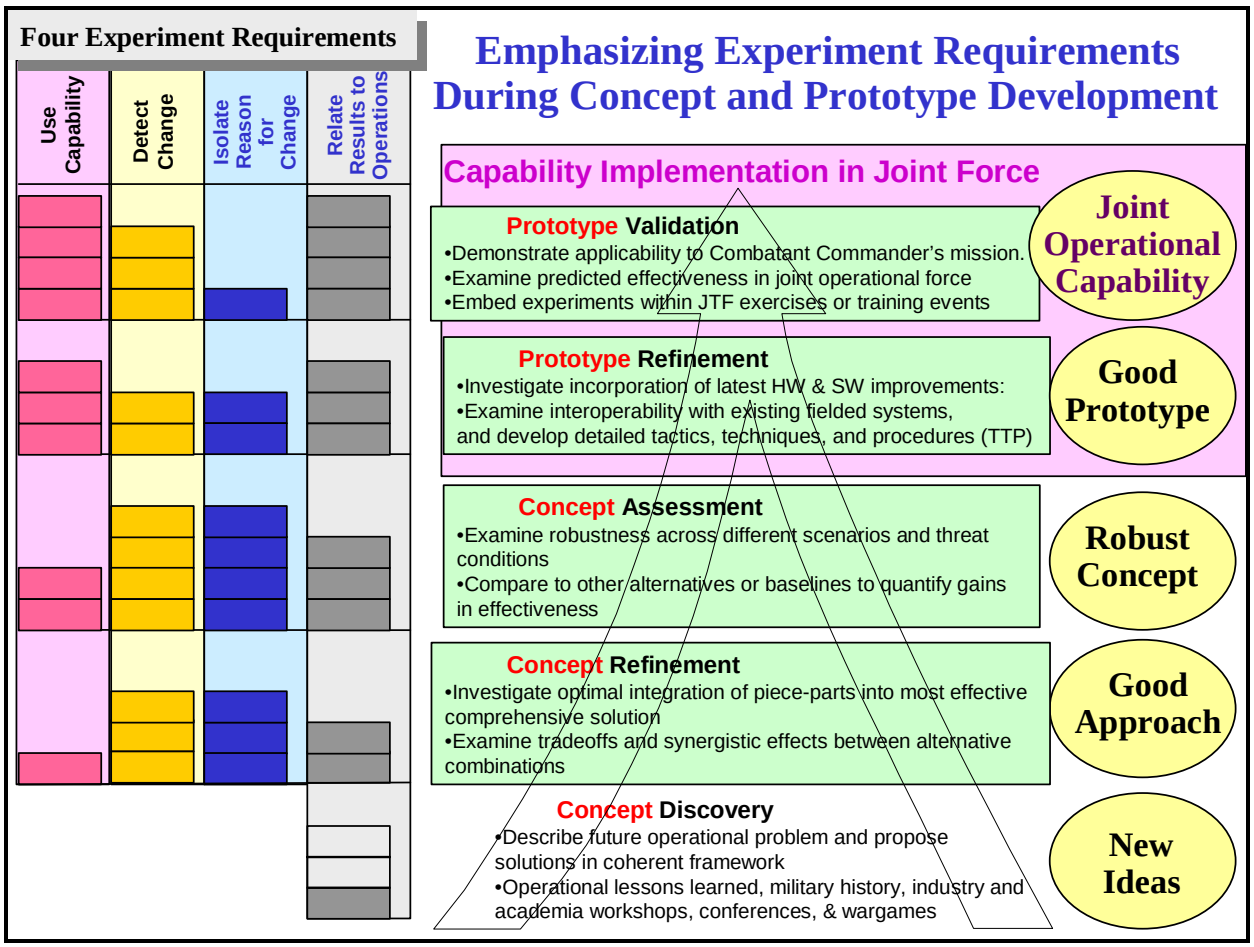


Figure 37 Progression from concepts to prototypes for successful experimentation campaigns

Concept Discovery. The primary consideration during concept discovery is relevance and comprehensiveness. To what extent do initial articulations of the future operational environment include a comprehensive description of the expected problems and propose a full set of relevant solutions? Relevance however, should not be over stressed. It is important to avoid eliminating "initially strange solutions" that subsequent experimentation should investigate for effectiveness.

Concept Refinement. Finding an initial set of potential capabilities that empirically show promise is most important in concept refinement. These early experiments examine idealized capabilities (future capabilities with projected characteristics) to determine if they lead to increased effectiveness. Initial experiments during concept refinement are dependent on simulations to represent simulated capabilities in simulated environments. Thus accurately isolating the reason for change is less critical to allow for "false positives." Allowing some false solutions to progress to be examined in later experiments with more realistic environments is more important than eliminating potential solutions too quickly. The concept refinement stage is dependent on experiments supported by methods such as constructive simulations, analytic wargames, and human-in-the-loop simulations. Sometimes, simple field experiments can be constructed to investigate whether future technologies will lead to a dramatic difference in operations by employing highly abstract surrogates; for example, designating that a hand-held clipboard provides exact enemy locations.

Concept Assessment. Quantifying operational improvements and correctly identifying the responsible capabilities is paramount in providing evidence for concept acceptance. Concept justification is also dependent on experiments with better-defined capabilities across multiple realistic environments. Experiments conducted using constructive simulations can provide statistical defensible evidence of improvements across a wide range of conditions. Experiments using human-in-the-loop simulations and field experiments with realistic surrogates in realistic operational environments can provide early evidence for capability usability and relevance. Incorporation of the human decisionmaker into human-in-the-loop simulations and field experiments is essential to the concept development process. Early in the concept development process, the human operators tend to find new ways to solve problems.

Prototype Refinement. In experiments during the prototype refinement stage, one should anticipate large effects or its implementation might not be cost effective. Accordingly, the experiment can focus on the usability of working prototypes in a realistic experimental environment. Isolating the real cause of change is still critical when improving prototypes. 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 simulations and field experiments with realistic surrogates for the prototype in realistic operational environments provide the experimental context for assessing gains in effectiveness when considering capability refinement and employment. Human decisionmakers may find unexpected ways to use and employ new technology effectively.

Prototype Validation. Applicability to the warfighting operational environment is paramount in prototype validation. 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 the combatant commander to employ it. Uncovering the exact causal chain is less important. In prototype validation, human decisionmakers ensure that the new technology can be employed effectively. Experiments during prototype validation are often embedded within exercises and operations (see Principle 9).

7.5 Employing Multiple Methods to Increase Rigor

This Principle has already presented the implications of tradeoffs among the four requirements when designing an individual experiment and provides a way to compare the inherent strengths and weaknesses of different methods available to choose from. The four validity requirements also provide the rationale for the necessity of experiment campaigns and provide a guide for developing integrated analysis and experimentation campaigns. Since a single experiment cannot meet all four requirements, a campaign consisting of a series of experiments and other analytical activities (Principle 4) can be designed to accumulate decision validity across the four requirements over time. This Principle now presents a specific example of one campaign paradigm—the model-exercise-model paradigm[34] (M-E-M)—in which experiments conducted using constructive simulations (the model), human-in-the-loop simulations, live simulations and analytic wargames (exercise) are combined to make up for the deficiencies in the four requirements exhibited by any one of these methods when used alone.

For example, when large analytic wargames and field exercises are used to conduct an experiment to investigate the effectiveness of new capabilities, the results are often disappointing. Because these exercises are player resource intensive, there are few opportunities to examine comparisons of alternative capabilities or to examine different scenario variations. In this situation, the utility of analytic wargames and exercises is enhanced within the model-exercise-model paradigm. The paradigm consists of conducting experiments using constructive simulations prior to the wargame or exercise and then following these events with a second set of post-exercise experiments using constructive simulations.

The early experiments using constructive simulations examine multiple, alternative Blue-force capability configurations and baselines. This methodology allows experimenters to determine the Blue configuration that provides the most robust potential benefit across different Red-force scenarios. A more realistic human-in-the loop simulation, analytical wargame, or field experiment where independent and reactive Blue- and Red-force decisionmakers and operators engage, then reexamines this superior configuration and scenario.

[34] Also called model-wargame-model paradigm.

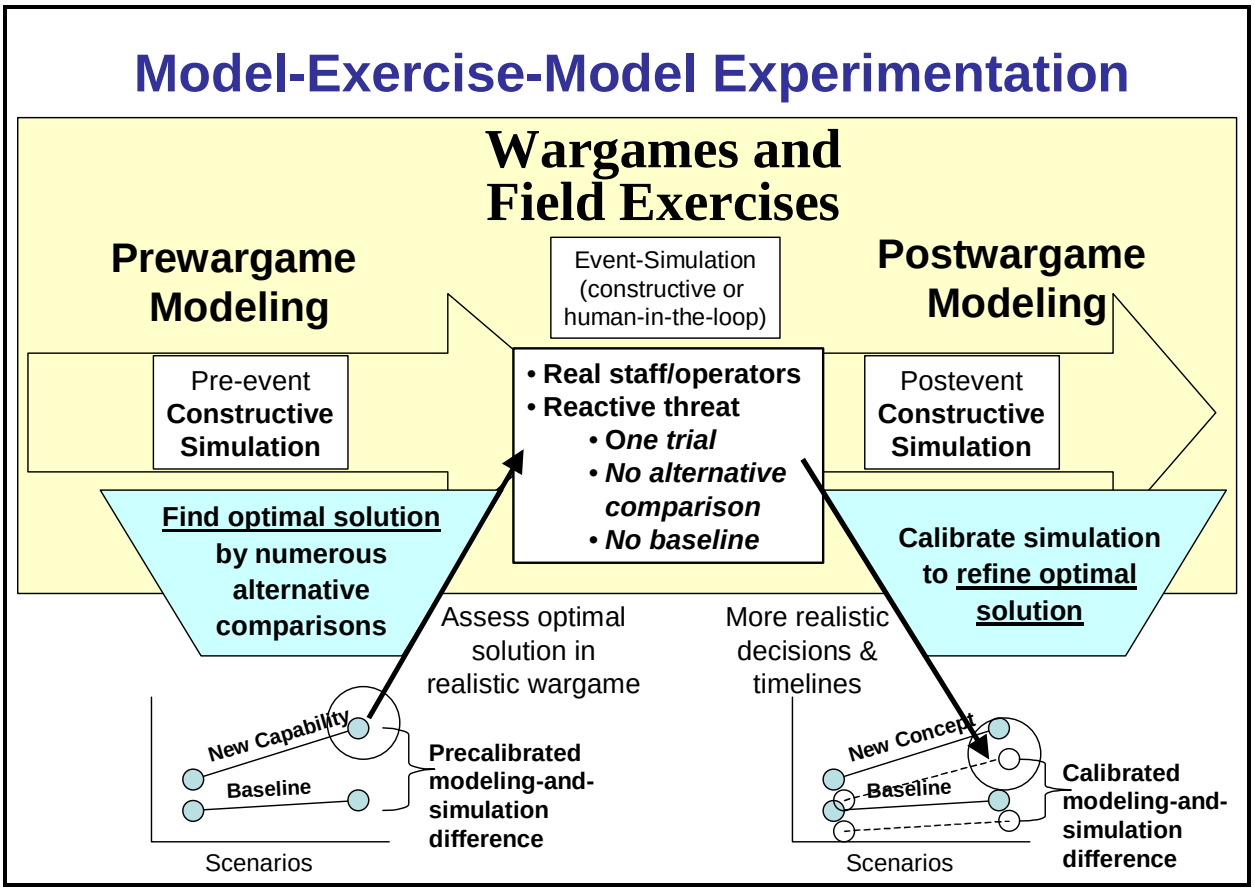


Figure 38 Model-exercise-model or model-wargame-model workflow

Pre-exercise Constructive Simulation. The early experiments using constructive simulation use the same order of battle and capabilities that are anticipated to be played in the exercise. These experiments examine multiple, alternative, Blue-force capability configurations and baselines. This pre-exercise simulation allows experimenters to determine the most beneficial Blue-force configuration of capabilities for different Red-force scenarios. It also helps to focus the data collection during the exercise by pinpointing potential critical junctures to be observed during the follow-on exercise.

Exercise. The exercise itself can focus on realistically executing the "best" scenario identified in the pre-event simulation. The "best scenario" is one where the simulation indicated that the new capability dramatically improved Blue's outcome. In the exercise phase, with independent and reactive Blue- and Red-force decisionmakers and operators engaged, the exercise allows the re-examination of this optimal configuration and scenario with more external validity than in the model phases. The scenario that provides the best opportunity for the new capabilities to succeed is chosen because exercises include the "fog of war" and traditionally the capability does not perform as well in the real environment as it does in the constructive 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.

Post-exercise Constructive Simulation. Experimenters use the results of the exercise to calibrate the original constructive simulation for further post-exercise simulation analysis. Calibration involves the adjustment of the simulation inputs and parameters to match the simulation results to those of the exercise (*e.g.*, via a wargame), thus adding credibility to the simulation. Correspondingly, rerunning the pre-exercise alternatives in the calibrated model provides a more credible interpretation of any new differences observed in the simulation. Additionally, the post-exercise calibrated simulation improves analysts' ability to understand fully the implications of the exercise results by conducting "what if" sensitivity simulation runs. Experimenters examine what might have occurred if the Red or Blue forces had made different decisions during the exercise.

This model–exercise–model paradigm increases overall experiment validity by combining a constructive simulation's ability to detect differences among alternative treatments with an analytic wargame, human-in-the-loop simulation or field experiment's ability to incorporate human decisions that better reflect the actual operating environment. This paradigm also helps to optimize operational resources by focusing the exercise event on the most critical scenario for useful results, and by maximizing the understanding of the exercise results through post-exercise sensitivity analysis.

7.6 Summary

Explicating experiment validity into four experiment validity requirements is quite useful when designing defense experiments to support concept or prototype development. This validity framework depicts the implications of tradeoffs when designing an individual experiment and provides a way to compare the inherent strengths and weaknesses of different methods available to choose from. It also provides a foundation for improving experiments executed during operational exercises. Just as importantly, the four requirements provide the rationale for the necessity of campaigns and provide a guide for developing campaigns. Since a single experiment cannot meet all four requirements, a campaign consisting of a series of analysis and experiment activities can be designed to accumulate decision validly across the four requirements over time. This Principle has also provided a specific example of one campaign paradigm—the model-exercise-model paradigm—in which experiments using constructive simulations (model), human-in-the-loop simulations, live simulations and wargames (exercise) are combined to make up for the deficiencies in each of the four requirements that were exhibited by each on their own.

Principle 8.

Human variability in defense experimentation requires additional experiment design considerations

Principle 8 provides an insight into the effects of human variability on defense experiment observations since an understanding of the impact of human variability on experimental design and outcomes is a first step toward its mitigation.

The implications arising from using human subjects in defense experimentation are generally overlooked. Most, if not all defense experiments examine impacts on socio-technical systems but experimental designs rarely cater sufficiently for the human element. Because humans are unique, highly variable and adaptable in their response to an experimental challenge, they are more than likely to introduce a large experimental variability. In addition, humans will have different experiential baselines in terms of, for example, training and trainability, and unlike technology, will become tired and possibly demotivated. The experimental design and the data analysis and collection plan must recognize and accommodate human variability. Human variability will be much larger than would be predicted if the socio-technical system were treated as technology. What is overlooked is that this variability provides important information on why a socio-technical system responds to a challenge in a particular way. Indeed there is an argument that human variability should not be minimized, as this would lose important information. High variability may indicate a fault in the system under examination or in the experimental design.

An understanding of the impact of human variability on experimental design and outcome is a fundamental skill required by all experimenters.

Principle 8. Human variability in defense experimentation requires additional experiment design considerations

The reader is encouraged to apply and adapt this Principle. However, the examples herein are based on the specific perspective and experience of different lead-nation authors with contributions from other participants. Therefore they may require supplementary effort to relate them to national perspectives.

8.1 Introduction

The implications arising from using human subjects in defense experimentation are sometimes overlooked. Most if not all defense experiments examine impacts on socio-technical systems but experimental designs often fail to cater sufficiently for the human element. In the context of this guide a socio-technical system is defined as an interacting collection of human and non-human parts. A socio-technical system is not a technical system with human "users," the human parts are integral rather than "bolt-ons." The important characteristic of a socio-technical system is that the behaviors arise from cycles of interactions between and within human and non-human parts.

Because humans are unique, highly variable and adaptable in their response to an experimental challenge, they are more than likely to introduce large experimental variability. In addition, humans will have different experiential baselines in terms of, for example, training and aptitude, and unlike technology, will become tired and possibly demotivated. The experimental design and the data analysis and collection plan must recognize and accommodate human variability, which will be much larger than would be predicted if the socio-technical system were treated purely as technology. What is overlooked is that this variability also provides important information on why a socio-technical system responds to a challenge in a particular way, see [Mathieson 2001; Mathieson and Dodd 2004]. Often, human variability can be accounted for by classification of individuals into groupings afforded by personality type, *e.g.*, Myers-Briggs types[35] or authoritarian *versus* non-authoritarian. Authoritarian types of people may behave differently with certain kinds of information technology, for example they may be less tolerant of uncertainty in information displays. Indeed, an argument can be made that human variability should not be minimized, as this would lose important information. High variability may indicate a fault in the system under examination or in the experiment design.

[35] http://skepdic.com/myersb.html An instrument for measuring a person's preferences, using four basic scales with opposite poles. The four scales are: (1) extraversion/introversion, (2) sensate/intuitive, (3) thinking/feeling, and (4) judging/perceiving.

8.2 Impacts of Human Variability

The positive and negative impacts of "human variability" can be visualized by considering them in the context of the four experiment validity requirements. This is shown in the following table, which is not exhaustive.

Experiment validity requirements	Key Human Characteristic	Positive Impact	Negative Impact
Ability to use the new capability	Adaptability	High adaptability in that humans can effectively employ the new system as designed and employ it adaptively in ways not originally envisioned or predicted.	Low adaptability in that humans have a difficult time in understanding and adapting or employing the new capability.
Ability to detect a change in the effect	Variability	It is possible to take natural human variability as a factor in the analysis and compare its effect with the effects of the deliberately manipulated variables.	Variability in a whole range of human factors introduces more noise than expected, which may impair the ability to detect any change due to the deliberately manipulated variables.
Ability to isolate the reason for the change in the effect	Cognitive and linguistic ability	Subjects have an understanding of why they reacted in a specific manner and can communicate this to the experimenter.	Subjects are misled as to the process and are not aware of why things turned out as they did.
Ability to relate the results to actual operations	Representative-ness	Subjects are typical of the warfighters expected to use the capability in the future.	Subjects are **not** typical of the warfighters expected to use the capability in the future.

Table 2 Impact of human variability on GUIDEx four requirements to valid experiments

An understanding of the impact of human variability on experimental design and outcome is a fundamental skill required by all experimenters. This is to ensure that maximum benefit is gained from defense experiments in relation to the experimental requirements above. Humans are variable and this variability manifests itself in physiological, anthropometric and psychological characteristics and the impact of these factors need consideration. Examples of these are given in the table below.

Social Science	Individual Characteristics
Physiology	Fatigue, endurance
Anthropometry	Weight, body size
Psychology	Cognition, intellect, team dynamics, leadership
Sociology	Cultural and social characteristics and interactions.

Table 3 Example of domains of variability due to humans in experiments

8.3 Experimental Design Considerations

8.3.1 General

Experimenters need to ensure that full account of all human variability in an experiment is considered. Conventional experimental design focuses upon minimizing human variability that may create **noise** upon the variables being measured. Although this is a powerful way to isolate variable impact and relationships, it has some fundamental drawbacks; mainly that it inevitably reduces the external validity of the experiment. For example, if subjects are chosen as having similar rather than different levels of aptitude, little information is obtained on individual differences in being able to use a new capability. One of the advantages of this additional information about individual differences is that training can be adapted to cater for all levels of aptitude.

If it is possible to achieve significant differences between treatments (should they exist) at the desirable effect size, without artificially constraining human variability, then this should be done. However, this does not occur very often because defense experiments frequently struggle to detect significant effects, due to limited sample sizes and large within-treatment variability (human or otherwise). If this is the case, then it is a good idea to add a human variability element to the experiment with only one or two treatments. In other words, execute the main experiment with a single group (if you have a single-group design) but also execute one or more treatments with different subject groups to establish the effect of human factors.

Regardless of the experimenter's ability (or desire) to control human variability, it is important to measure it. This is to determine if detected effects can be explained in

terms of human variability rather than treatment changes. For example in a within-subjects design with a number of treatments, it may be possible to measure learning effects within each treatment, and from that estimate any confounding effect on the treatments that learning had on the whole experiment. This will increase the complexity of the experimental design since the data analysis will need to incorporate human variability measures into the analysis in ways to measure their impact upon the main variables.

The experiment also needs to be designed to reduce or eliminate human variability due to fatigue or boredom. These are especially important to consider in an experiment with a repeated measures design using a small sample, where the likelihood is that each subject will be tested in all treatments. Although randomizing the order of treatments may reduce practice and learning effects, randomized blocking of treatments is a more effective method of reducing fatigue and boredom, especially if a large number of treatments are involved. The content of the task needs careful planning in that there is sufficient workload for the subjects, not only to increase the amount of data collected but also reduce boredom experienced by subjects. However, the workload should not be excessive as to encourage fatigue. Breaks should be incorporated where possible to reduce fatigue. The design of the experiment must ensure the reduction of boredom and fatigue.

8.3.2 Representative Sample

The reasons for having subjects with different levels of aptitude, whether this be due to experience, prior training, *etc*., is to ensure having a representative sample of the population in the experiment. Measuring different levels of aptitude will enhance knowledge of the impact of this variability and not only does this provide a better representative sample to relate the findings to actual operations, more importantly it can aid in enhancing the success of using a new capability.

8.3.3 Method of Data Collection

Objective measures can only help to observe actions and performance of subjects. Subjective measures, however, are important to ascertain the mental processes of why subjects have behaved in such a manner, enabling a clearer understanding and isolation of reasons for change in effect. This information may be important especially if a subject adapts to using a capability in a way not considered by the experimenter. Asking subjects why they have changed their behavior can enhance understanding of maladaptive ways of using a new capability. Consideration needs to be given to the timing of subjective interviews, particularly whether they should take place soon after the action occurs (for example the end of a day's play, or the end of a single trial run) or at the end of the experiment. The former may be obtrusive to the subjects and may impact the results, with the latter being affected by memory decay, motivation, *etc*.

8.4 Data Re-use

The conduct of defense experiments can be costly and care needs to be taken to re-use data collected in experiments as much as possible. Simulations and models provided by Operational Analysis are being used across the TTCP nations to support procurement submissions for equipment. The re-use of human variability data from experiments to support these activities has been limited to date because human variability has been "controlled out" in the experiments themselves. The impact of this has been that human variability is commonly not represented in Operational Analysis simulations and models. By increasing the complexity of design of defense experiments so that human variability is not artificially constrained, it will be possible to ensure that the human element is accounted for during the procurement of equipment in socio-technical systems.

8.5 Subject-induced Variability

Having attempted to accommodate and understand the complexities of human variability, and how it can be used to the benefit of the experimenter, there is still the problem of subject-induced variability. The dismounted combatant is the most adaptable element of the fighting force and this will have an impact on test equipment and experiment design. Soldiers will utilize equipment for purposes for which it was not designed. This will introduce an element of unpredictability and uncertainty to the conduct of, and data collected from, an experiment.

In this instance the variability may arise from a range of factors, which cannot be controlled. An obvious source is the motivation of the subjects. This has been discussed in previous chapters but the difficulty remains in how to maintain motivation. A number of potential courses of action are open to the experimenter and these may include a greater involvement of the subjects in the experimental team, for example discussing the conduct of the trial or by a more positive approach to after-action reviews, counseling for individual comments.

Human performance, and hence variability in the experiment, is uniquely sensitive to the physical environment. Heat, cold, rain, altitude, dehydration, poor nutrition, sleep loss/deprivation or excessive consumption of alcohol, among other factors may collectively, or individually serve to impair performance and increase variability. While the majority of these factors are unlikely to be encountered in simple desk-top wargaming, in large field experiments they are likely to become major issues. Some of these constraints may be alleviated by good experimental design while others, such as extremes of environments cannot be well controlled. The difficulty is that human response to these challenges will vary between subjects and, unless the experimenter undertakes a full medical screening, the degree of variability will not be known. However, such screenings, except for classical laboratory experimentation, are costly and have no real value. At the very best the experimenter should consult with appropriate specialists to understand the degree of variability.

P8 Human Variation

A final source of variability, which is also generally overlooked, depends on what might be induced by instrumenting subjects or equipment. The experimenter will wish to gather some objective data at some stage, but the presence of objective probes on the subject's body may alter how the subject responds to the experimental design, or how he behaves. An obvious instance is where the data gathering equipment is mounted on the subject's body. In this case the subject may alter his clothing, load carriage equipment of other items of personal equipment. Again, there is little an experimenter can do in this case except to undertake some limited pilot studies to understand the impact of the instrumentation on subject behavior.

Principle 9.

Defense experiments conducted during collective training and operational test and evaluation require additional experiment design considerations

Principle 9 shows that experimenting during training exercises and operational test and evaluation (OT&E) events, where considerable infrastructure is provided, represents cost-effective opportunities only if appropriate and special design considerations can be devised to meet the four requirements for valid experiments. This is an area where organizations can get important leverage from their programs (science and technology (S&T); research and development (R&D); concept, demonstration and experimentation (CD&E); procurement; OT&E; operations; and training) when exploiting, for example, a M-E-M paradigm. Operational assessment using troops and simulators are especially useful early in the capability development cycle.

Opportunities to conduct such experimentation may be found in operations as well as in exercises and in OT&E events. The drive to conduct experimentation activities during operations and exercises is almost entirely due to the difficulty of acquiring the resources (equipment, estate, human) to undertake experiments of any significant size. Arguably, the equipment programs that require most support from experimentation are those intended to enhance **collective** rather than team or individual effectiveness. Most nations generally do not have units and formations available to dedicate to experimentation where collective groups of personnel are required. Therefore exploiting routine training exercises and other collective events should be given serious consideration.

Exploiting collective training (exercises) has a range of benefits as well as disadvantages and a variety of factors must be taken into account in both planning and execution. The principal one is that training always has primacy and the experimenter has little control over events, thus the skill is in understanding the constraints that the exercise opportunity will present and knowing how to work within them. Exploiting exercises for the purposes of experimentation is most achievable during the prototype validation phase of an experimentation campaign when functional prototypes exist. Although exercises and operations do not allow execution of elaborate experiment designs (because it would impede training and impact operational readiness), scientific methodology and the four experiment validity requirements can be applied to experiments embedded in real-world exercises.

Experimentation during exercises, OT&E, and operations naturally provides the strongest venue for meeting the fourth experiment validity requirement, *i.e.*, the ability to **relate** results to actual operations. While operational necessity restricts the ability to meet the first three experiment validity requirements, the experimenter can ameliorate the limitations to some degree.

Principle 9. Defense experiments conducted during collective training and operational test and evaluation require additional experiment design considerations

The reader is encouraged to apply and adapt this Principle. However, the examples herein are based on the specific perspective and experience of different lead-nation authors with contributions from other participants. Therefore they may require supplementary effort to relate them to national perspectives.

9.1 Introduction

Opportunities to conduct experimentation may be found in operations as well as in exercises and during operational tests and evaluations. The drive to conduct experimental activities during operations, exercises, tests and evaluations is almost entirely due to the difficulty of acquiring the resources (equipment, estate, human) to undertake experiments of any significant size. Operational assessments, in particular, provide an opportunity for conducting experimentation early in the testing and acquisition cycle by employing substantial technical and expert staff support using simulators. Arguably, the equipment programs that require most support from experimentation are those intended to enhance **collective** rather than team or individual effectiveness, and thus collective groups of personnel (which may comprise command teams with higher and lower controllers) are required to undertake that experimentation. It is a simple fact of life in the early 21st Century that most nations generally do not have units and formations available to dedicate to experimentation, except for the most limited-scale activities. Therefore exploiting routine training exercises and other collective events should be given serious consideration.

Exploiting collective training (exercises) has a range of benefits as well as disadvantages and a variety of factors must be taken into account in both planning and execution. The principal one is that training always has primacy and the experimenter has little control over events, thus the skill is in understanding the constraints that the exercise opportunity will present and knowing how to work within them. Exploiting exercises for the purposes of experimentation is most achievable during the prototype validation phase of experimentation when functional prototypes exist. Although exercises and operations do not allow execution of elaborate experiment designs (because it would impede training and impact operational readiness), the scientific methodology and the four experiment validity requirements can be applied to experimentation embedded in real-world exercises.

Experimentation during exercises and operations naturally provides the strongest venue for meeting the fourth experiment validity requirement, ability to relate results to actual operations. While operational necessity restricts the ability to meet the first three

experiment validity requirements, the experimenter can ameliorate the limitations to some degree. Prototype testing prior to the exercise or operation enhances the chance to use the experimental capability and to ensure that it will function during the exercise trials (Requirement 1). Additionally, the prototype engineers should be on hand to train and assist the operators in the use of the system. Establishing a pre-exercise definition of expected performance and comparing the prototype's actual performance during the exercise to its expected performance provides the necessary ability to detect change (Requirement 2).

While the ability to isolate the observed change to the experimental prototype is the most problematic in embedded experimentation (Requirement 3), experimenters can achieve some level of satisfaction here, also. When examining different capabilities during a single exercise, the experimenter should conduct different prototype trials at different times so the effects of one prototype do not influence the effects of the other. It is prudent to have an experienced exercise "observer-controller" view the prototype trial to assess the extent that any observed results were the results of the experimental capability instead of unintended causes. Additionally, showing that the rigorous experiment data accumulated during the concept development phase of the prototype is still relevant to the exercise conditions also supports Requirement 3 assessments. Finally, a M-E-M paradigm that was successfully calibrated to the operational exercise results would allow follow-on sensitivity analysis to demonstrate that inclusion and exclusion of the experimental capability accounted for decisive simulation differences.

The potential to include experimentation within operational test and evaluation (OT&E) programs is high. This is so in part because many of the components of OT&E events are the same as their counterparts in experiments. Unlike exercises and operations, OT&E events can support detailed data collection and in some cases can support elaborate designs. Consequently, OT&E generally supports Requirements 2 and 3 well. Requirement 1 may be met where new systems can be included within OT&E programs. Such systems must be nearly ready for operations with potential for immediate transition. Although Requirement 4 may not always be met, the potential is very high when the OT&E is conducted in field trials with operational scenarios, staff and equipment. While the language, techniques and resources are quite similar, the basic philosophical approach to designing an experiment is unique and must be adhered to for a successful experiment to occur. Practical experiments may be conducted, often as excursions for the test scenario, without significant changes to OT&E events.

9.2 Experimenting during Training Exercises

9.2.1 Introduction

This section, which deals with the issues arising from attempting to achieve parts of an integrated analysis and experimentation campaign during collective training exercises, is aimed at informing the potential experimenter of the benefits and pitfalls of exercise exploitation, so that one can construct a campaign taking all possibilities into account.

Exercise exploitation is often a practical necessity but depending on the individual circumstances, views vary enormously about the feasibility of achieving anything meaningful by so doing. As an example, even the experimental force (EXFOR) in the US Army Task Force XXI (TFXXI) Advanced Warfighting Experiment (AWE) of 1997 had to meet their training objectives concurrent with testing the TFXXI hypotheses. In the UK, the MoD has been actively exploiting standard collective training rotations in constructive, human-in-the-loop and live environments for some years, and the experimenters have gradually been allowed to introduce elements, which actually have an impact on the training process. But does this damage training, and can any genuinely useful experimentation be achieved within the constraints of typical training exercises and objectives?

The perceived benefits of experimenting during training will be covered next, followed by a short dissertation on the nature of training exercises, and consequently what constraints are likely to be placed on the key activities covered by this guide. This section is based largely on the UK's experience of exploiting Army training exercises over the past five years, but the lessons learned should have wider application.

9.2.2 Benefits of Experimentation during Training

9.2.2.1 Introduction

As indicated previously, the drive to perform experimental activities during training is almost entirely due to the difficulty of acquiring the resources (equipment, estate, human) to undertake experiments of any significant size. Arguably, the equipment programs that require most support from experimentation are those intended to enhance **collective** rather than **team** or **individual** effectiveness. Thus collective groups of personnel (which may comprise command teams with higher and lower controllers) are required to undertake that experimentation. This in turn means the use of either large-scale real estate or major simulator training systems. Except for delivering the training for which they were designed and procured, these capital facilities often have little spare capacity for other activities, such as experimentation.

It is a simple fact of life in the early 21st Century that most nations generally do not have units and formations available to dedicate to experimentation, except for the most limited scale activities. Therefore exploiting routine training exercises is a tempting alternative. The following sections go on to discuss the key characteristics of collective training and then the constraints that these impose on the experimenter. However, it is worth noting that there are reasons for exploiting training other than the paucity of dedicated opportunities for experimentation.

9.2.2.2 Engagement of Experimental Subjects

Military personnel of all ranks take collective training very seriously. It is often assessed and can have a direct effect on their subsequent careers. Hopefully any staff detailed off to take part in a dedicated, or bespoke experiment will also take it seriously and give it their all, but it is difficult to achieve the degree of engagement or immersion routinely

found in major exercises. This can have both positive and negative impact. The positive side is self-explanatory; the teams of players will be trying their utmost to do well and will be deeply immersed in the simulated action. Moreover, training exercises involving a live opposing force (OPFOR) are normally very competitive. In addition, if equipment is on test, or some simulated or surrogate future equipment is being experimented with, its acceptance or effective use by the subjects yields a much more conclusive result than in a standalone test. This is precisely because the main objective of the subjects is not the test itself—in a training exercise, they are trying their utmost to win, usually by any available means. However, this is a two-edged sword. If, for whatever reason, despite a good pre-training program, the equipment is not liked, or not accepted, it will probably be ignored (or in the case of dismounted soldier systems even thrown away!) for exactly the same reason. If it is perceived to help the subjects win, it will be used; if it isn't, it won't. Thus the use of new equipment surrogates on training exercises has a tendency to produce black-or-white results.

A related point is that participants in training exercises are normally exercised (in the broadest sense) to the full. They will work long hours, often in highly stressful conditions, especially in live field training. It is difficult, and some would argue unethical, to attempt experiments during dedicated training activities.

9.2.2.3 Use of Training Infrastructure

Training infrastructure in this context covers training staffs and facilities; planning effort; and exercise support during execution and after-action review (AAR). The use of training staffs and facilities does not, of course, imply the exploitation of training exercises themselves. However, the staffs and facilities in question are normally heavily utilized during the training year and simply may not be available at any other time. The main benefit worthy of further discussion is the general reduction in staff effort (and cost) on the part of the experimenter if training exercises are used.

This includes the detailed exercise planning, including scenario writing. However, there are limitations of using training scenarios and these will be covered later. The effort of ensuring that the right military personnel (experimental subjects and exercise control and data collection staff) are in the right place at the right time should not be underestimated, and if a large proportion of this planning effort is already being done for training exercise, this can be enormously helpful. In addition, training will normally culminate in an AAR. Although this will mainly bring out training points, the AAR is a useful source of player feedback, especially if the training staff allow some discussion of whatever experimental doctrine, structures or equipment are being experimented with. Running an effective AAR is not straightforward and if the training staff are proficient at it (and not all are) and they understand the needs of the experimenter, using the training AAR as a source of player feedback can be very effective.

9.2.2.4 Sample Sizes

Defense experimenters usually need to rely on events that occur frequently in the context of a small number of vignettes or missions if they are intended for

demonstrating statistical significance. However, training rotations using similar, if not identical scenarios are not uncommon in some nations. Given that the overall design, planning, execution and analysis cycle for major experiments can be in the order of two or more years, there is scope (time and potential) for exploiting these repeated rotations; this can be done in one of two ways.

First, they may be used as a baseline. The experimental force (EXFOR) can then run whatever excursion is required in its own rotation, and spend less of the available time running baseline cases. This has the benefit that the baselining can be done unobtrusively and economically, and (say) two years worth of exercises will provide some idea of the variability of the key parameters. This in itself can assist the analysis of the experimental phase with the EXFOR. In the UK a good example of this was the BIG PICTURE 1 experiment, which was performed in the US Army Simulation Network (SIMNET) facility in Germany in 1997. The aim of the experiment was to test the hypothesis that digitizing a company-level force would improve operational effectiveness. Control Group data was collected from standard training rotations over a two-year period, resulting in a distribution of 12 event-versus-time graphs for company-level attacks, which could be compared with a smaller set derived from the experimental week with the digitized EXFOR (who had no training objectives to fulfill). Figure 39 shows the event timing results of this experiment with three phases as follows: reconnaissance (Recce), command and control (C2), and assault. The x-axis shows the key events and the y-axis the time taken to achieve them. Each hatched line represents an attack from a standard training rotation and each thin solid one an attack by the digitized EXFOR. The thick lines represent the means of the two populations. Statistical analysis backed up the conclusion that can be drawn from visual inspection of the mean lines: that there was a significant improvement in the speed of the C2 phase due to digitizing the force. Other work, also exploiting data captured from the training rotations, was able to reject various alternative hypotheses as to the cause of this apparent improvement. The large quantity of data collectable from the training rotations also indicated the large natural variance in such event timings.

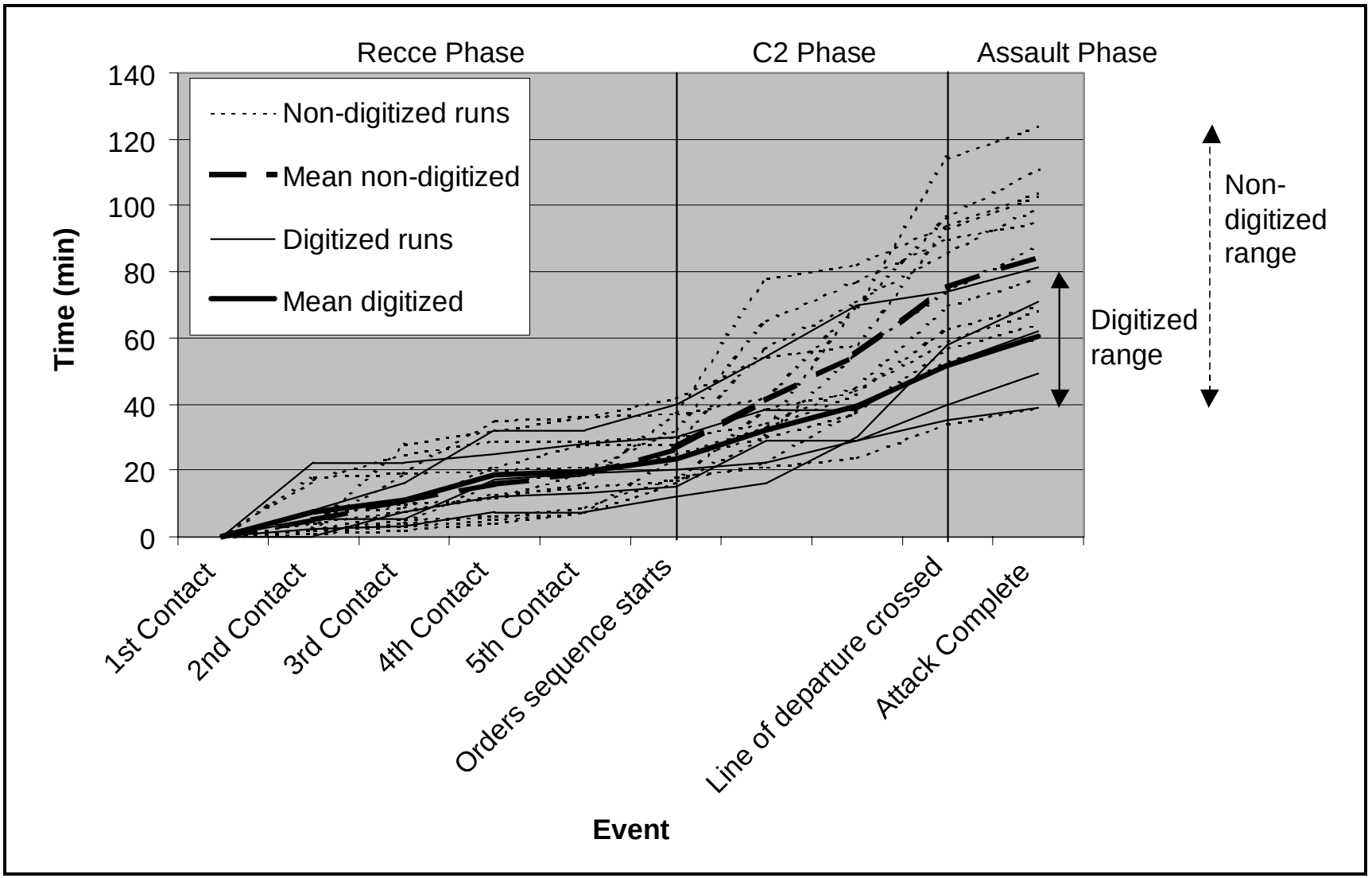


Figure 39 Event timing results from the UK digitization experiment in SIMNET environment, 1997. Solid lines from digitized EXFOR show better results than using standard training rotations, *i.e.*, non-digitized.

Second, if some level of intrusion into the standard exercises has been authorized, exploiting a series of similar exercises can provide a reasonable sample size with an experimental excursion ("reasonable" in this case being more than one!). This is particularly powerful if each exercise is split into a standard set of missions and a balanced trial design can be used so that the excursion is spread evenly across them. Table 4 below gives an example, where four exercising units each perform four missions. It is thus possible to achieve two sets of data (**control** and **treatment**), which have an equal number of missions of each type and are unbiased toward any particular unit.

Mission	Unit	A	B	C	D
1		■			■
2		■	■		
3			■	■	
4				■	■

■ = experimental case

□ = baseline case

Table 4 Balanced trialing during training rotations

9.2.3 Nature of Collective Training Exercises

9.2.3.1 General

Collective training exercises are intended to enable structured groups of personnel to learn and practice (train or practise in UK English) collective skills. They often include an element of assessment (sometimes to the extent that it is difficult to determine when the actual learning takes place!). Whatever extraneous activity takes place in an exercise, it is fundamental to note that meeting training objectives will almost always take primacy. In other words, everything else must fit in around training.

9.2.3.2 Design

Exercises are designed to stimulate various training points in support of the training objectives. That may or may not satisfy the needs of an experiment; if it doesn't, there is not much the experimenter can do about it except choose a different exercise. There are sometimes other factors to be taken into consideration. In particular, exercises are often designed to ensure that all participating elements get a fair share of the action. It would certainly be a waste of time and effort for some of them if that were not the case, but such artificialities can potentially serve to confound the findings of any experiment.

9.2.3.3 Scenarios and Settings

Training exercises obviously tend to be run at fixed training establishments and the real estate at those establishments and the precise training objectives have a major effect on scenarios and settings. Even command and staff training (CAST), using some form of real-time wargame, is sometimes constrained to use scenarios that take place in the area local to the training facility, so that "live" reconnaissance and "intelligence preparation of the battlespace" (IPB) can be performed as part of the exercise. This is in effect another constraint for experimenters and one that they can do little about. Most nations formulate a range of standard scenarios for their own force planning and operational analysis work; these are normally mutually exclusive from standard training scenarios. Thus the relationship between an experiment taking place on the back of an exercise and OA using some entirely different setting needs to be thought about carefully.

9.2.3.4 Exercise Control Intervention

Exercise Control (EXCON) organizations intervene in the execution of exercises for many reasons. Safety concerns are perhaps the most regular one, but it is often simply because a particular mission is not going to plan and either the training force or the OPFOR are apparently winning too easily (maybe only in a particular region of the battle). This can result in elements of the OPFOR being restrained on the one hand or "resurrected" (more than is usual) on the other. If an experiment is piggybacking on an exercise when that happens, this can produce real problems. It may be that a sequence of like exercises is being used to produce a consistent sample and that in some of the individual exercises, a manipulation is being made, such as the provision of a new item of equipment. The purpose would normally of course be improving operational effectiveness. However, if the result of achieving that in practice was that EXCON reacted to enhance the OPFOR in some way, it would be very difficult to measure the improvement.

Steps can be taken to avoid this; the main being to ask that OPFOR only be enlarged (in a simulation) or resurrected (in a LIVEX) **after** the standard OPFOR has been defeated or when reaching some agreed threshold. This enables good data collection up to that point and usually meets the needs of the training process too.

9.2.3.5 Training Progression

Any unit or formation will undergo a range of different training exercises during its training cycle. This progression generally moves from purely command and staff training through collective simulation to full live simulation and live fire training. This results in a potential trap for the unwary experimenter. Despite the fact that a particular exercise might appear to be a good match to the experimenter's needs, if it is too early in the unit or formation's training cycle, the collective skills may still be too low to provide a valid experimental environment. For example, there is little benefit in experimenting with some new command and control software if the users are barely proficient in standard C2 procedures and staff work. There is no golden rule to

determine which exercises may be used and which not, but it is recommended that the potentially participating units and their chain of command be consulted in depth as to an exercise's suitability.

9.2.4 Constraints

9.2.4.1 General

From the previous section it is clear that training exercises impose a range of diverse constraints upon the experimenter. Overall the following questions have to be asked when an exercise exploitation event is being planned:

1. Am I likely to get clearance to perform any experimental activity during a training exercise?
2. Can a useful experimental activity be fitted around the exercise in question?
3. Can all of the required data be collected?
4. Can the experimental team attend or preferably actively engage in after-action reviews (AAR) to elicit further information from the training audience?
5. Is it possible to intrude into the exercise in some way without significantly impacting on the training objectives?
6. Can we actually enhance the training by performing experimental activity?
7. Does it matter that **training** scenarios, force structures and settings are being used?

9.2.4.2 Understanding the Training Environment

Assessment issues. The mere intention to collect data from training exercises is seen as controversial in some quarters. This is mainly because formal assessments during training can have a profound effect on officers' careers. Conversely some armed forces do not do training assessment as a matter of policy. Thus a scientific team collecting data on (say) command processes in headquarters can be seen (wrongly) as adding an extra layer of assessment, and an unwanted one from the perspective of the trainers and training audience alike. This is very much a question of perception and discipline when reporting such work. When applying to do the work, always use terms such as "data collection," "process analysis," *etc.*, rather than "assessment."

Commenting on units. One of the perennial problems of exploiting exercises is performing the required analytical work without explicitly making comparative comments about different individuals or units. Experimenters and practitioners will inevitably see some variation in process, command style and performance across a set of broadly similar exercises, but it is not their role and place to comment on this explicitly. If they do, and units or individuals are recognizable, they will probably find that no further invitations to attend exercises run by the same Command or formation will ever be forthcoming. When practitioners are dealing with a small number of units, they simply refer to unit as "unit A, unit B, unit C,*...*" in their report(s), and if possible, they do some simple shuffling of the chronological order in which they exercised (*i.e.*, so "A" is not the first one and "D" is not the last) to avoid obvious identification of the individuals involved.

"White coats and clipboards." It is no exaggeration to say that some in the military have something of a fixation about scientists striding around their HQ in the middle of an important exercise, asking inappropriate questions at inopportune moments. This has certainly happened on occasions in the past and practitioners need to gain the trust of the training organization and training audience if this impression is to be avoided. In essence, good practitioners must display appropriate professionalism and be aware of the environment they will be working in, *i.e.*, go properly prepared; ensure that some members of the team have been to similar exercises before; if possible have a dry run. Whatever has been agreed with the chain of command in the run-up, clear the ground rules locally with the training unit CO. If possible add a serving officer or two to the experimenter team, either to be part of the team or to act as uniformed liaison (see Figure 40). It is also important for personnel from a particular agency not to create a bad name for that agency and prejudice future opportunities.

Sometimes, practitioners will be caught out by the environment (Figure 41). Whether on land, at sea, or in the air, military training takes place in sometimes hostile environments. However, if practitioners follow the advice above, this will simply be treated as a little light relief by the military authorities, rather than the straw that breaks the camel's back and gets people thrown off the exercise.

Benefits for training. When discussing what may or may not be allowable on an exercise, be sure to emphasize the possibilities for actually enhancing the training, rather than focusing wholly on what the negative impact might be. Potential benefits fall into a number of categories. First, the fact that a greater emphasis than usual is being placed on data analysis and collection, might allow you to provide the trainers with more data and other objective feedback to support the AARs than would be normally be possible. The AAR is a very important part of the training process and anything practitioners can do to improve it is usually welcomed. Second, the effects of intrusively adding surrogate new equipments to the exercise. This is usually the most controversial aspect of exercise exploitation. However, experience with the British Army has shown that if planned and executed sympathetically, such additions can actually enhance training. For example, coalition operations with the US will often bring allied forces into contact with equipment concepts that will not be integrated in their own service for a few years. Good examples at the time of writing are JSTARS (ASTOR), Tactical UAV's[36] and various digital CIS. Thus enabling the forces under training to experiment with future equipments can actually prepare them better for coalition operations in the near term. Also, it may be that the unit undertaking the training may have been instructed by their parent Service or Command to experiment (in a loose sense) with new doctrine or procedures in preparation for some new type of equipment (for example, attack helicopters). Frequently, the best that can be achieved by doing this with in-service equipment is often far from a credible representation of the future capability. Thus proposals for providing them with even a small number of credible surrogates or simulations of the future capability will usually go down very well in these circumstances.

[36] Or TUAV, tactical unmanned air vehicle.

Figure 40 "Assault combat data collectors" and military minder—go properly prepared!

Figure 41 Practitioners "will sometimes get caught out by the environment."

Experimenting with the radical during training. Notwithstanding the comments above, it is normally not possible to exploit an exercise intrusively with **radically** new concepts, which take the training audience far away from the current *modus operandi* that they are supposed to be training for. For example, many of the longer-term themes for Network Enabled Capability (NEC) or Network Centric Warfare (NCW) initiatives across the nations are concerned with bringing genuine jointness down to the lowest tactical levels of command. Undoubtedly these initiatives will require considerable

experimentation to enable them to be taken forward in an effective and coherent manner. However, (possibly with some exceptions) such experimentation is unlikely to be undertaken on the back of training, as the concepts are just too different from current practice. There is a related risk even with not-so-radical concepts, namely that trainees who are being assessed as part of the training process will themselves be wary of anything different that might detract from their own assessment. Minimizing this risk can only be done as part of the briefing and negotiation with the training audience in the period leading up to the exercise.

9.2.5 Summary

Training exercises can offer an excellent environment for some types of experimentation. To make the best use of them, it is essential to understand both their benefits and constraints. A brief summary is as follows:

Benefits

1. Availability of experimental subjects in large numbers
2. High level of engagement of experimental subjects
3. Use of training infrastructure
4. Moderate sample sizes, for repeated exercise series
5. Ability to use repeated exercises as a control group, or baseline
6. High rating in terms of relating any detected change to real operations.

Constraints

1. Design
2. Training has primacy. Can a genuine experimental design be fitted around training?
3. Scenarios and settings designed for training purposes
4. Interventions by Exercise Control for training reasons
5. Training progression: Exploitation of an exercise too early in a unit's training cycle can yield poor results.
6. Intrusion: Limited opportunities to make intrusive changes to the exercise or collect data intrusively
7. Commenting on units: Can results be published without breaching the anonymity of the training audience?

Several of the threats to valid experimentation described in Principles 2 and 3 apply particularly to the exploitation of training exercises. It is hard to generalize about how these may be overcome due to the enormous variety of training exercise types, but the main trick is to understand the constraints and work within them. If it is not possible to fit an experiment into a training exercise without significant changes to the exercise design, then exercise exploitation is probably not the best way forward.

9.3 Differences and Similarities between Experimentation and Operational Test and Evaluation

9.3.1 Introduction

The discussion in this section deals with the issues related to conducting portions of an integrated analysis and experimentation campaign during OT&E events. One of the key considerations is that much of what is required to conduct OT&E (technical staff, equipment, and procedures) is also required to run an experiment. The differences and similarities between experiments and tests will be examined herein in order to provide guidance on the best approaches for designing experiments within OT&E events.

9.3.2 Benefits of Experimentation during OT&E

The factors driving organizations to perform experimental activities during OT&E events are the same as for training events. Most nations do not have units and formations available to dedicate to experimentation, except for the most limited scale activities. Therefore exploiting OT&E (as well as training) is an option to be considered.

OT&E events are important components in the acquisition and maintenance phases of equipment life cycle management programs. They are well supported by the technical/engineering community and valued by the operational community as a component of the operational readiness process. The operational community will therefore generally be engaged in OT&E events and the potential to include them in experiments as well can be very good.

An important benefit to experimenters is the OT&E infrastructure, which includes engineering/technical staffs and facilities; planning support; test support during execution and evaluation support for the AAR. The benefit from the use of OT&E staffs and facilities is realized because of the strong overlap between the two processes. This overlap is shown in Figure 42. An important benefit to the OT&E community is that the prototypes from experiments may soon be operational systems. In such circumstances, there is a significant advantage to be obtained by the inclusion of OT&E staffs in the experimentation on these systems. It is worth noting that the development of new OT&E procedures and facilities has been stimulated and improved through OT&E involvement in experimentation. So the two communities gain by working together, OT&E gain in new/novel apparatus and methods, and experimenters gain in trial infrastructure and the associated knowledge, the know-how embedded.

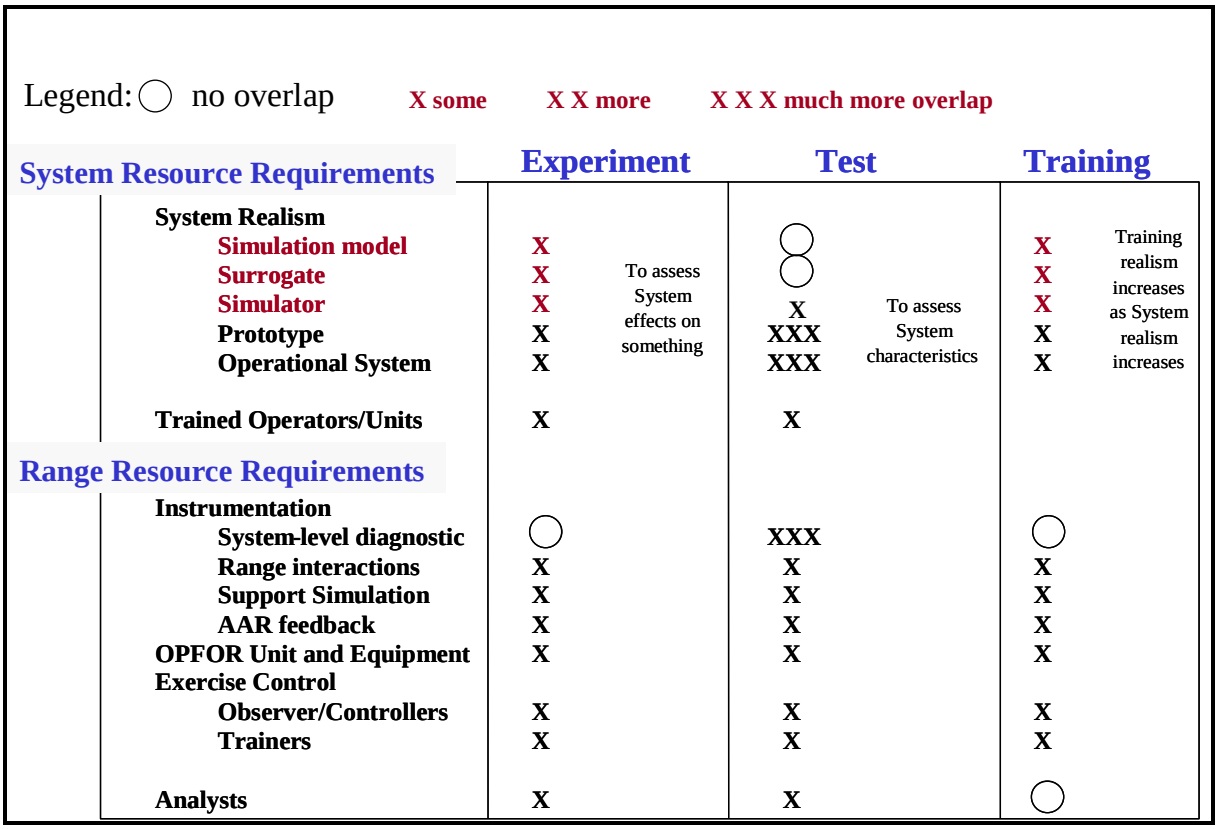


Figure 42 Comparison: similarities and differences between experiments, tests and training

9.3.3 Experiments *versus* Tests: The Differences and Similarities

OT&E is generally for the test and evaluation of new and in-service systems in support of operational readiness evaluations. The events are designed to quantify various aspects of equipment performance or are conducted to determine if a standard for performance is being met. This environment may or may not satisfy the needs of a particular campaign. OT&E scenarios are linked to establishing a performance standard and most nations formulate a range of standard scenarios for their requirements. Therefore, the feasibility for conducting an experiment on the back of a test using entirely different settings needs to be thought about carefully. Events like the Joint Warrior Interoperability Demonstration (JWID), currently Coalition Warrior Interoperability Demonstration (CWID) program, where new systems are under test, may be more flexible.

Understanding the differences and similarities between tests and experiments is important if the experimenter is to utilize OT&E events successfully. A helpful step for establishing such understanding is to look at terminology. Assuming that the terms **problem** and **requirement** can be used interchangeably, then the following expressions serve to distinguish between training, demonstrations, tests and experiments. One would say:

1. in training we "**practice** to meet the requirement,"
2. in demonstrations we "**show** how to meet the requirement,"

3. in tests we "**determine if this meets** the requirement," and
4. in experiments we "**determine the best way** to meet the requirement."

These expressions are used in Figure 43 with the example of a new sensor, A, and detections, B, applied to the four different types of events for perspective.

Sorting Through Terminology

A = New Sensor
B = Detections

Event	Goal Stimulating Event	Purpose of Event
Training	**Practice on A to get B.**	Operation to **assist** entity in acquiring ability to do A.
Demonstration	**Show how A works to produce B.**	Operation to **show/explain** how A works.
Test	**Determine if A works** (produces B). •How effective is **A**? •Can operator/unit do **A**?	Operation to **confirm the quality of A.**
Experiment	**Determine if A solves B.** •Is A related to **B**? •How much does **A** affect **B**? •Did something else produce **B**?	Operation to **discover a causal relationship between B and something else, A.**

Figure 43 Comparison: terminology for training, demonstration, tests and experimentation

Examining tests and experiments only, the differences in these activities can be characterized by the questions one typically would ask for each type of event. A **test** is an operation to assess or quantify the **quality or presence of something**. It asks questions like:

1. Does this work?
2. How well does this work?
3. Under what conditions does this work?
4. Does this work with that?

These questions all relate to defining a threshold of performance. An experiment is an operation to assess a causal or quantifiable relationship. It asks questions like:

1. What is important in this type of operation?
2. What is the impact of this on that?
3. What is the best thing to do in this situation?
4. What will fix this problem?

5. Where and when are the best times to use this?
6. Why does this work?
7. How does this work?
8. Did something other than **A** produce **B**?

The theme throughout this section on differences and similarities might be captured in the single statement **"different questions, but similar design and execution."** This is depicted in the example shown in Figure 44. A test for a new sensor determines if it meets the threshold for performance of detecting 14 targets. An experiment for a new sensor determines answers for questions like the effect of target type on detection. It is a clear difference in purpose, but the design and execution can be essentially the same.

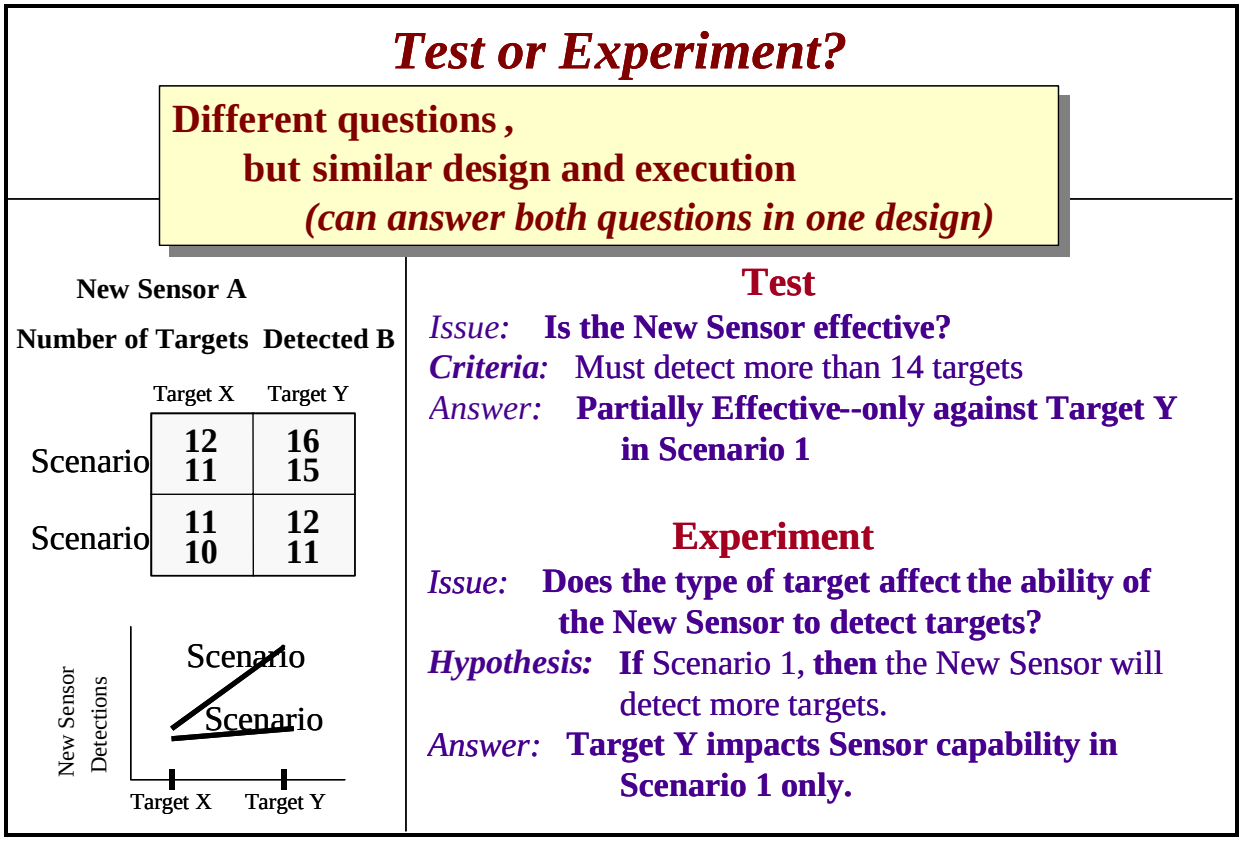


Figure 44 Contrasting tests and experiments

This leads to the two final questions for this section; **Can one Test during an Experiment?** and **Can one Experiment during a Test?** In principle, the answer is yes to both. The real answer of course is in the details. To consider if one can test during an experiment, the different types of experiments should be examined. Figure 45 shows the capabilities of the four types of experiments to support tests.

Opportunities for testing a system emulated in an experiment using a wargame will be limited, but can contribute under certain circumstances. The system's characteristics can only be examined in general terms, but it is possible to look at the impact of the

proposed capability in an operational scenario. Experiments using constructive simulation or human-in-the-loop simulators have the best potential to support tests. They are particularly useful in tests for providing sufficient repetitions on a system and for collecting quantitative data on a system. Human-in-the-loop simulators can also provide diagnostic data. While experiments involving field system prototypes have the best potential to support tests to assess system characteristics, they have quite a few limitations. The tests will be limited to the experiment scenario and conditions and sufficient repetitions will be a problem.

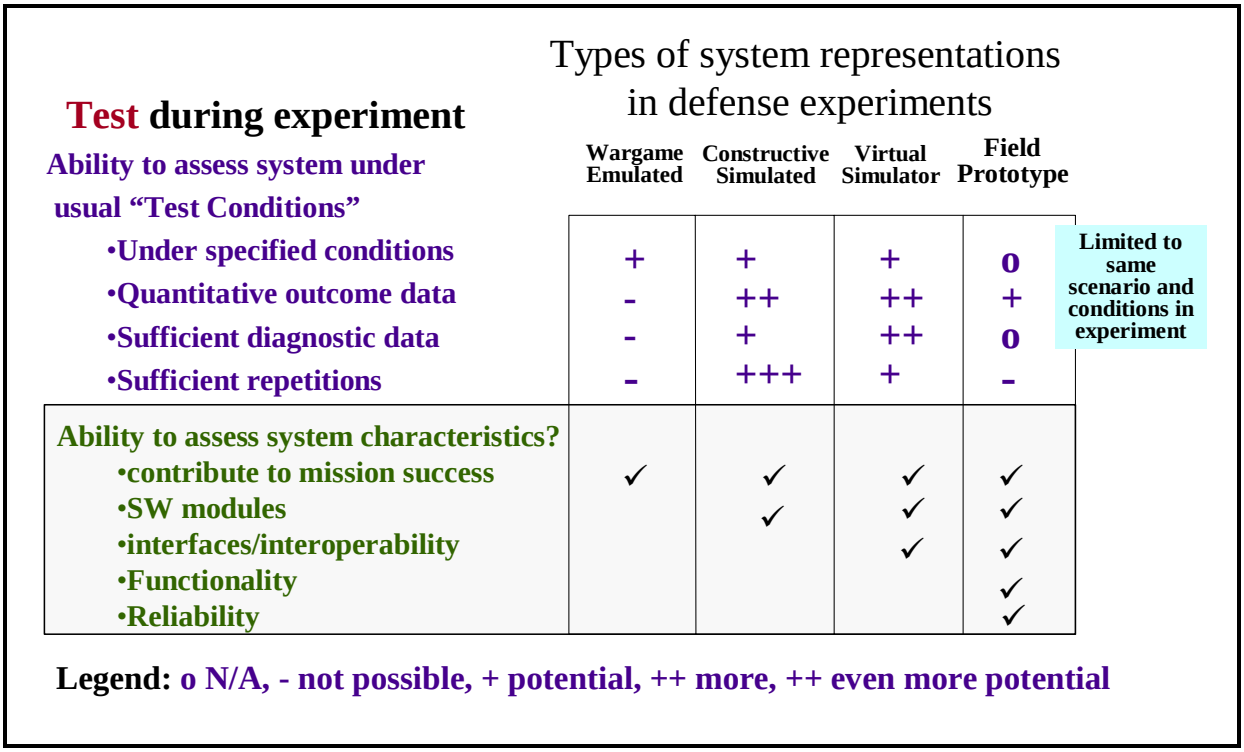


Figure 45 Can one test during an experiment?

The alternative question, can one experiment during a test, can be examined in terms of the different types of tests. Three categories are shown in Figure 46 and are evaluated in terms of features that can be manipulated for experimentation. The first category—Constructive Simulation—requires special attention. While field tests and human-in-the-loop simulations are well suited to support OT&E, constructive simulation is probably only suited to testing a system concept. This type of test should typically be useful for experiments on different system characteristics and for experiments examining different scenarios. Changing doctrine, TTPs or organizations is typically difficult in constructive simulations.

Field tests and tests using human-in-the-loop simulators are considered to be the more common options for OT&E. Tests using human-in-the-loop simulators will have less flexibility for experiments on system characteristics and by extension tests with real systems or prototypes will have little or no flexibility for such experiments. Human-in-the-loop simulators are suitable for experiments examining a system in different

scenarios and, given the human component, are most suitable for experiments on different doctrine, TTPs or organizations. Experiments conducted within field tests using real systems or prototypes are limited by the test range and forces assigned to the test and there may be little or no flexibility to experiment with different scenarios. The field venue is, however, also good for experiments on doctrine, TTPs and organization.

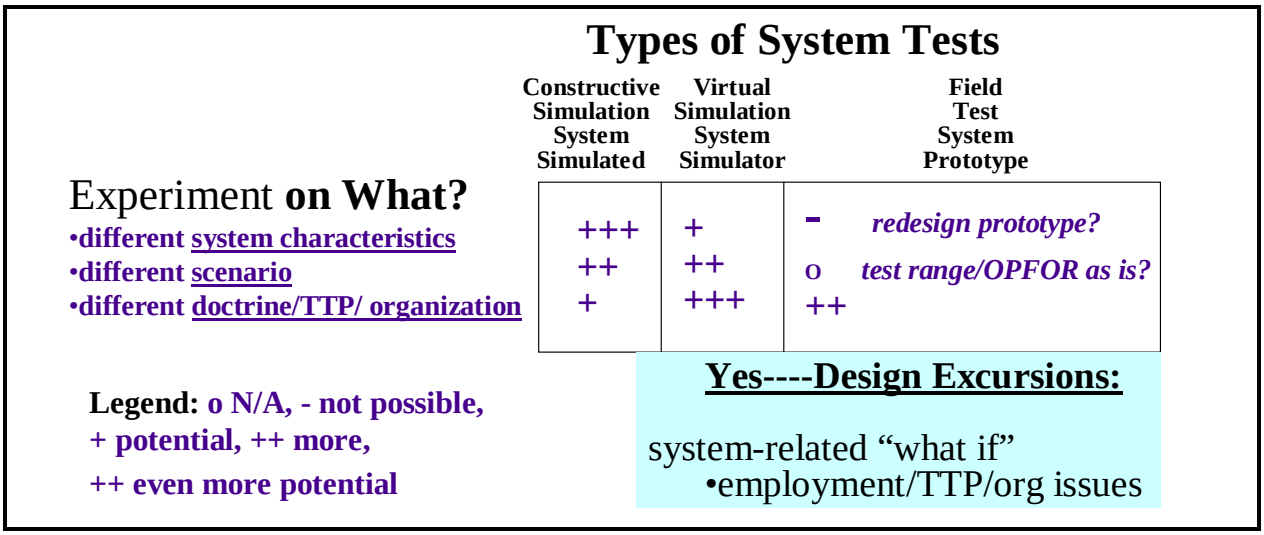


Figure 46 Can one experiment during tests?

Summarizing the differences and similarities between test and experiments, the resource requirements needed to examine a system are generally more demanding for tests. The range requirements for field tests and experiments are usually similar. It might be a good policy to develop future ranges to accommodate testing, training, and experimentation. Clearly system testing can be done in field experiments, especially if the experiment is near-term in nature. There is, however, generally little control over conditions/scenario. Furthermore, there are fewer repetitions and less system-level diagnostics, hence less quantifiable data. On the plus side, the experiment can provide a test with data on system functionality and interoperability. As for experimenting during a system test, again this is possible if the experiment is near-term. The experiment is generally limited to excursions on the system under test. It is possible to examine innovative doctrine, tactics, and organization issues as well as other system interactions. Another plus is that the range resources are in place for the test.

9.3.4 Constraints

9.3.4.1 General

Similarly to training events, it is clear that OT&E events impose a range of constraints upon the experimenter. Overall the following questions have to be asked when the exploitation of a test event is being planned:

1. Am I likely to get clearance to perform any experimental activity during an OT&E event?
2. Can a useful experimental activity be fitted around the event in question?
3. Can all of the required data be collected?
4. Can the experimental team attend or preferably actively engage in after-action reviews (AAR) to elicit further information from the event audience?
5. Is it possible to intrude into the event in some way without significantly impacting on the objectives?
6. Can we actually enhance the OT&E by performing experimental activity?
7. Does it matter that *readiness* scenarios and settings are being used?

9.3.4.2 Understanding the OT&E Environment

Assessment Issues. Data collection is a standard and very important feature of OT&E events. Generally, OT&E data collection requirements will exceed the requirements of the experimenter making these events attractive collateral events for experiments. The assessment process is somewhat different and the experiment planner must ensure that the MoPs and MoEs required to examine the hypothesis are supportable. The OT&E data collection plan cannot be assumed to suffice.

Reporting Issues. One of the perennial problems of exploiting tests is performing the required analytical work without explicitly making comparative comments about different individuals or units. There will inevitably be observations on some variation in process, command style and performance across a set of broadly similar events. It is not appropriate to comment on this explicitly (see training Section 9.2.4.2).

Personnel Issues. The comments in training Section 9.2.4.2 are equally appropriate in OT&E events, except that scientists tend to blend in with the engineering and technical staff fairly well. Still, it is good advice to “be aware of the environment one will be working in and go properly prepared.”

9.3.5 Summary

OT&E events can offer opportunities for some types of experimentation. To make the best use of these events, it is important to understand both their benefits and constraints. A brief summary is provided:

Benefits

1. Availability of operational staff and platforms
2. High level of engagement of technical community
3. Use of OT&E infrastructure
4. Moderate sample sizes, for repeated test series
5. Ability to use repeated tests as a control group, or baseline
6. Strong potential for relating any detected change to real operations.

Constraints

1. Design
2. OT&E has priority and the experiment may not interfere with test objectives
3. Scenarios and settings designed for OT&E purposes
4. Limited opportunities to make intrusive changes to the test or collected data
5. Can results be published without breaching the anonymity of the test audience?

If it is not possible to fit an experiment into an OT&E event without significant changes to the test design, then it is probably not appropriate to try.

Principle 10.

Appropriate exploitation of modeling and simulation is critical to successful experimentation

It is estimated that as much as 80% of experiments employ M&S in some fashion.

Human-in-the-loop simulations, constructive simulations and analytic wargames offer an immersive and safe environment in which to explore operational activities and have a range of other advantages over live simulation (as used in Field Experiments) such as: increased control, ease of data collection, the ability to simulate events and capabilities impossible in the live environment, and the capacity for personnel to experience a representation of the future. M&S can either be created specifically for the purpose of experimentation, or alternatively for other purposes such as training.

However, the all-pervasiveness of M&S is not without its problems. Costs are often high; there is usually a wide range of potentially applicable M&S to select from; and the question of validity is never far away from the experimenter's list of priority issues. Therefore the **appropriate** use of M&S is vitally important for successful experimentation, and that is the subject of this Principle.

Principle 10. Appropriate exploitation of modeling and simulation is critical to successful experimentation

The reader is encouraged to apply and adapt this Principle. However, the examples herein are based on the specific perspective and experience of different lead-nation authors with contributions from other participants. Therefore they may require supplementary effort to relate them to national perspectives.

10.1 Introduction

We have reached the stage that modeling and simulation (M&S) is intrinsic to conducting most defense experiments. There is now a wide range of M&S techniques (both real-time and faster-than-real-time) available and this makes the innovative use of M&S cost effective for many experimentation applications. However, there are some significant issues associated with selecting both the types of M&S to be used and the specific elements of the experiment federation.

10.2 Fidelity *versus* Adequacy

For many years, as rapidly increasing computing power led to many new modeling possibilities, there was a generally held view that greater fidelity, or accuracy, was always better. Indeed, many took the term "validity" to be almost synonymous with fidelity and detail. The modern view is that validity actually means "fitness for purpose," with the **purpose** being to execute the desired experimental design. This means that we should consider the main measure of merit for M&S to be adequacy to support our experimentation, not fidelity of battlespace representation. The experimental design should effectively define what level of fidelity is adequate.

10.3 Excessive Fidelity or Detail

Cost usually rises with fidelity or detail, so clearly getting this aspect of M&S definition wrong can add considerably to the experiment's price tag. However, that is not the only drawback. In "The Lanchester[37] Legacy" [Bowen and McNaught 1996: Vol. III, Ch. 9], the authors wrote: "It has long been understood by operational researchers that, in dealing with complicated situations, simple models that provide useful insights are very often to be preferred to models that get so close to the real world that the mysteries of the world they intend to unravel are repeated in the model and remain mysteries." In other words, the main point of modeling is to rationalize the complexity of real life by simplifying it. This applies just as much to the M&S we use to support defense

[37] F.W. Lanchester was one of the pioneers of military operational research.

experimentation as it does to M&S used in operational research or in "experiments using constructive simulations." We can therefore imply an axiom that M&S should be as simple as possible while remaining adequate for the task in hand. This does not of course mean that all M&S should be simple in an absolute sense; "as simple as possible" will in some cases still be very complex: it just should not be over-complex.

The main manifestation of over-complexity in practice is more (or more complex) internal relationships and interactions than are necessary for the model's intended use. This leads to:

1. more supporting data being required than is really necessary,
2. a greater requirement for validation and testing,
3. a greater chance that the analysts will not understand elements of the M&S or be able to interpret the experiment's results correctly,
4. a greater chance that the results of the experiment will be, at least in part, an artifact of the M&S, and
5. a greater chance that unnecessary variability will be injected into the experiment, thereby threatening experiment validity Requirement 2.

10.4 Validation

10.4.1 Taking a Balanced View

As already described, a common view in the past has been that more M&S fidelity than actually required is fine, but less than required would be invalid. The argument presented above is that the former is not always true, but what about the latter? Let us postulate that for a particular experiment there is some acknowledged shortfall in the fidelity of available (or perhaps affordable) M&S. The first point to note is that simply acknowledging that fact is a good start. There are often non-M&S-based workarounds that will mean that insufficiently detailed M&S can, in fact, be made acceptable. For example, the addition of a human response cell that takes its input from the simulated action, with clearly laid out rules governing their actions, could fill a gap in M&S coverage. Also, earlier Principles in this document have described the various threats to experiment validity and it must always be borne in mind that there is no such thing as a perfect experiment. Therefore the modeling shortcomings might not actually be the weakest link in the validity of the experiment as a whole, so it is essential to take a balanced, holistic view of validity.

10.4.2 Validating M&S

There are many standard methods of validating and verifying models and simulations and sometimes these can seem rather procedural. This document will not go into detail about any one of them, but will instead cover some basic principles that should always be borne in mind. Despite the arguments presented above about adequacy, fitness for purpose and the importance of taking a holistic approach, the fact remains that models and simulations are abstractions of the real world and therefore must be shown to be

reasonable abstractions. In other words, the question we should be asking is: **Is the M&S a fair reflection of the real world, inasmuch as this matters for the purpose of conducting my experiment?** The arguments presented earlier suggest that we should not overly concern ourselves with those potential shortfalls that don't really matter to us (or at least are by no means the greatest threat to the validity of our experiment). However, sooner or later we will need to address those aspects that really are important to us. We therefore need to ask:

1. What are those aspects and how do I know one if I see one?
2. In the case of simulating future military operations, what does "real world" actually mean?
3. And, having established that, how do we confirm that our M&S is a reasonable reflection of the real world?

These questions are now addressed in turn.

10.4.2.1 How do I Know What's Important?

Sadly there are no hard and fast rules here; it's really a matter of judgment. However, the best place to start is the hypotheses you are testing and the cause-and-effect relationships you are trying to establish. Consider the case where a computer generated forces (CGF) package is being used to support experimentation using human-in-the-loop simulators. One experiment might be to determine if changing an ISTAR sensor mix enables a HQ to become aware of more potential strike targets. In this case the CGF would really just be creating a land battle backdrop to be viewed by virtual airborne sensor simulators. Important aspects would be entity density, laydown (disposition) of forces, and signatures. If GMTI radar was being considered, then the ratio of moving-to-static vehicles would be important, and perhaps some subtleties about how groups of vehicles move in formation: for example, in a bounding advance, all of the vehicles in a platoon would never move together.

On the other hand, consider the same CGF being used to support an experiment using human-in-the-loop simulators to compare new armored fighting vehicle (AFV) concepts. Macro force laydown becomes much less important here, but micro-details of vehicle formations and everything associated with direct fire engagement, previously almost entirely irrelevant, now become predominant. This would include:

1. vulnerability to different munitions, probably as a function of aspect angle;
2. movement characteristics over different types of terrain;
3. actions on contact;
4. target detection and identification capability related to those of the manned simulators; and
5. weapon effects.

Operational test and evaluation provides a well-tried and tested methodology for asking the right kinds of questions and formulating expressions analogous to what the OT&E community terms "critical operational issues" (COIs). COIs, tailored to an experiment's hypotheses, draw attention to the nature of interactions between elements (systems, processes, humans, *etc*.) of an experiment. For example, using a human-in-the-loop

synthetic environment to support doctrine refinement for a new armed reconnaissance helicopter (ARH) capability, might warrant a COI along the lines of "Can the ARH be threatened appropriately?" This COI drives effort to ensure that the simulated operational environment suitably stresses the ARH systems and crews. The OT&E methodology advocates development of just a handful of critical issues (per experiment), each of which creates the foundation for development of MoPs and MoEs on which to base assessment of whether the issue has been satisfied (whether the M&S representation is "fit for purpose"). For the same reason that it's not possible to fully satisfy all four validity requirements in a single experiment, compromises for some COIs will often be necessary. This should not be seen as opening the door on criticism of the form: "the experiment is therefore invalid". It simply provides a basis on which to trace outcomes that should be considered less reliable than others. So, by considering the M&S requirements as they relate to the experiment in hand, it is possible to make rational, and mostly objective, judgments about which areas of modeling do need to be properly referenced to the real world. So far so good, but what is the real world exactly?

10.4.2.2 What Is the Real World Exactly?

If we were in the business of re-creating historical battles, we could define the real world quite well. We usually know about the equipments and the geographical areas in some detail and if necessary could even re-create quite an accurate meteorological picture. There has also been considerable research across the nations into human effectiveness in combat and how it compares with exercises and range firings, and if appropriate, the relevant degradation factors could be embodied into the CGF or constructive elements of our M&S.

But what happens in the normal situation we find ourselves in, when we are experimenting into possible futures? We can only rely on those enduring aspects of the current or past worlds that can reasonably be extrapolated into the future. These are mainly the laws of physics and any well understood human combat effectiveness norms that exist. But even these are not sacrosanct. There is usually a trade-off to be made between external validity (*i.e.*, reference to the real world) and internal validity. For example, one may require a simulation that deliberately accelerates events (or at least doesn't deliberately slow them to established combat norms) so that sufficient action can be got through in a limited experimental period (perhaps to obtain the required sample of events for statistical analysis). Now, as long as it can be shown that such acceleration will not affect, or bias, the results of the experiment (for example, by inducing excessive workload), then such an approach could be entirely reasonable.

Returning to the earlier example of a CGF being used in concert with human-in-the-loop AFV simulators, another interesting case-in-point is visual detection modeling in a CGF. There are now various well-established methods of modeling visual target detection, mostly based on sound empirical data. However, the visuals in a simulator might not be sufficiently accurate to reflect these models particularly well. Obviously the ideal solution would be to calibrate the simulator visuals carefully with the available empirical

data. However, this would be a very expensive and analytically difficult thing to achieve and may well be considered impractical. Therefore, in order to ensure a fair fight, it may be more important for the CGF visual detection modeling to match the (partially flawed) simulator visuals than to be the best-known reflection of the real world. The underlying principle here is that all modeling or experimentation is an abstraction of the real world and it is more important that the level of abstraction is **consistent**, rather than elements of it being the **best**.

Another aspect of the real world is the definition of the future operational or equipment concepts that are to be experimented with. In our abstract version of the real world we must clearly represent these as intended by the designer. The representation of operational concepts is normally best validated subjectively by the concept designers. When doing that, clear visualization, so that the concept designers can clearly see what is (and, equally, what is not) being represented in the M&S, is paramount. When considering the validation of more tangible concepts, such as platforms or networks, architectural frameworks (as described later in Section 10.6) are an excellent means of describing concepts precisely and providing a baseline against which modeled representations of them can be compared.

10.4.2.3 How do You Confirm that a M&S Reflects the Real World Adequately?

Even if one is fortunate enough to have a range of test data against which a model or simulation can be compared, the key is to break the model or simulation down into manageable chunks and compare their behavior with relevant real-world referents. Some of these chunks may well be purely theoretical in nature (*e.g.*, the radar range equation) and so can be compared directly with results predicted by the underpinning theory. Others, for example simulated combat, may benefit from correlational studies between simulated combat outcomes and the combat outcomes of matched realistic field training exercises. Still others, for example CGF behaviors, must be assessed by appropriate military subject matter experts, although flexible and objective data analysis and collection tools are becoming available to support them in this task. Thus we should ensure that our M&S has both subjective and objective credibility in all-important respects. Much of this credibility will be relevant to many different applications of the M&S, and so it is essential to record all validation steps in some sort of logbook, so the work does not have to be repeated unnecessarily. The drawback of logbooks, however, is that they can lead the reader into a misconception that validity is an absolute attribute, whereas, as has already been described, it can be strongly dependent upon application.

10.5 M&S Definition

It is a key principle that the definition of the M&S to be used in an experiment should be derived from the experimental design, and not the other way around. However, rarely will practitioners have the luxury of completing their experimental design, then moving through a user requirements definition process, and subsequently a system requirements definition process in sequence. Usually a concurrent process is necessary,

with the processes beginning in the order given above. A spiral development process can then take place, as shown in the diagram below.

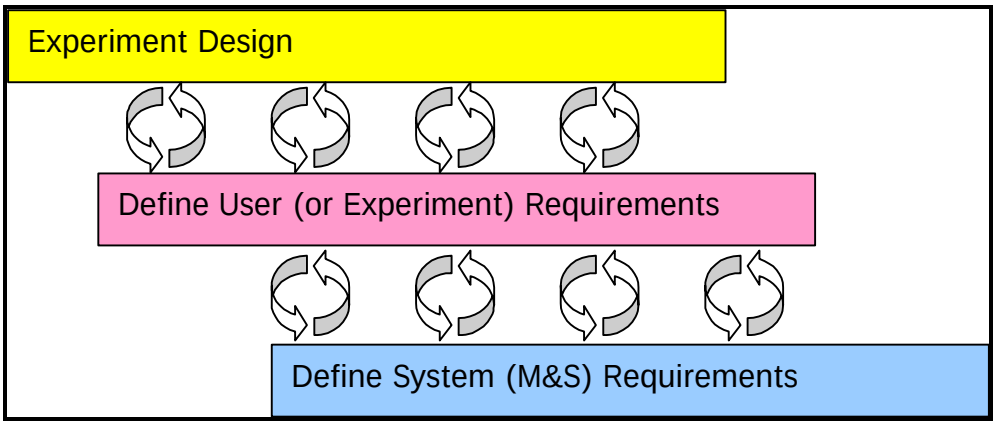


Figure 47 Cyclic-concurrent process from design to M&S requirements

The experimenter is usually limited by the range of M&S practically, or affordably, available to him. Therefore, the development may actually be the development of a federation rather than a specific model or simulation, with the aim of making the best use of those simulations that are obtainable. Either way, development processes are often rushed and it is not uncommon for some of the desired (or even required) functionality to be missing by the time the M&S is needed for experimentation. Therefore right from the outset, the experimenter should be prepared for M&S federations to have some shortfalls and be prepared to implement workarounds to cover them.

10.5.1 Recognized Methods

10.5.1.1 FEDEP

The US Defense Modeling and Simulation Office (DMSO) has developed the Federation Development and Execution Process (FEDEP), which has transitioned into an IEEE recommended practice (IEEE Standard 1516.3). The FEDEP is a detailed set of processes to assist with the design, development and implementation of High Level Architecture (HLA) federations. It deals in essence with application and problem domains and begins with the definition of the federation objectives. From the experimenter's viewpoint these are obviously directly related to the experimental design.

The scope of the FEDEP is restricted in the following ways:

1. it does not cover the complete lifecycle of a SE, since it just focuses on the federation development part;
2. the main emphasis of re-use is at the federate level, rather than at the federation and component level; and
3. it is focused on the implementation of HLA federations and does not support other interoperable technology, *e.g.*, DIS.

10.5.1.2 SEDEP

EUCLID (European Co-operation Long Term In Defence) RTP 11.13 was a major European research initiative to improve and promote the utilization of Synthetic Environments (SEs) in Europe. The aim of EUCLID RTP 11.13 is to overcome the obstacles that prevent SEs being exploited in Europe by developing a SE Development and Exploitation Process (SEDEP) and a SE Development Environment (SEDE) based on an integrated set of commercial off-the-shelf (COTS) products and prototype software tools. The objective of developing a process underpinned with a software toolsuite is to reduce the cost and timescale of specifying, creating and utilizing SEs for defense based applications, *e.g.*, simulation-based acquisition and collective training. Another key output of the program is a prototype "pan European" Repository that will provide a basis for the management, storage and retrieval of information relevant to SE development, execution and analysis activities.

The SEDEP provides extensions and enhancements to the original DMSO FEDEP to satisfy the wider needs of the SE community, for example definition of "Steps" dedicated to analyzing top-level user needs and evaluating results from SE experiments. Some of these enhancements have contributed toward the evolution of the FEDEP into the IEEE 1516.3 FEDEP Standard that was issued in March 2003. The purpose of the SEDEP is to:

1. encourage use of SE technology to benefit different application domains;
2. provide guidance for developers and users to plan and perform the different activities necessary to produce the required products and results;
3. promote good practice for developing SEs on time and within budget;
4. facilitate re-use of products (federation, federates, components) and results; and
5. establish a process that can be underpinned with a software toolsuite aimed at reducing the cost and timescale of specifying, creating and utilizing SEs.

The SEDEP is relevant to all military or civil applications of SEs and covers all aspects of its specification, development and operation. It is applicable to creating and utilizing small SEs, involving a few networked simulations running on a Local Area Network (LAN), through to large SEs, running on a Wide Area Network (WAN) across national borders. Although the SEDEP uses terms from the High Level Architecture, *e.g.*, "Federation", the process can be tailored to support other interoperability technologies, such as Distributed Interactive Simulation (DIS).

The long-term focus is for the SEDEP and FEDEP to merge into one process by taking the best aspects from each process. The convergence process has already started since the IEEE 1516.3 FEDEP Standard incorporates ideas promoted by the SEDEP v1.0: for example, the definition of a step dedicated to analyzing and evaluating the results from an SE-based experiment.

10.5.2 *Ad-hoc* Methods

For smaller tasks, such general purpose and well-documented processes as FEDEP and SEDEP can sometimes seem a little cumbersome. It is, of course, possible to create any

number of *ad-hoc* processes to replace them and sometimes that seems like the most pragmatic approach. However, the philosophy behind FEDEP and SEDEP is sound and provides a logical and auditable means of defining M&S based on experiment (or other) requirements and confirming that the federation or systems remain fit for their intended purpose (*i.e.*, are valid) once they are implemented. Therefore, it is important that the key tenets of these processes are followed even if the toolsets are not used or the "letter of the law" is not being strictly adhered to.

10.6 Modeling the Process to be Experimented with

Experiments and observational studies (where a concept is subjected to objective observation, but without manipulation) are intrinsically connected to the idea of hypotheses. The hypothesis is simply a plausible proposition about either causal or associative relationships. Thus in a general sense there is always implicitly a model of the process being experimented with by virtue of there being one or more hypotheses.

However, it is possible, and in most cases desirable, to model the process in advance in a much more tangible way, regardless of whether a strict model-exercise-model philosophy is being followed. In particular, architectural frameworks such as Zachman [Zachman 1987] and DoDAF [DoDAF 2004] represent an excellent and increasingly popular means to describe military problems and potential candidate solutions in a variety of different ways.

The IEEE STD 610.12 defines "architecture" as "the structure of components, their relationships, and the principles and guidelines governing their design and evolution over time." An architecture description is a representation, as of a current or future point in time, of a defined "domain" in terms of its component parts, what those parts do, how the parts relate to each other, and the rules and constraints under which the parts function.

An architecture framework provides guidance on describing architectures and in this way provides the governance on how an architecture should be constructed. Several such frameworks exist, each with their own particular applications. The Zachman framework [Zachman 1987] provides a matrix, linking focus (what, who, where, *etc*.) to perspective (owner, designer, builder, *etc*.), with each cell containing a specified representation of the enterprise at a particular level. The matrix thus provides a ready-reckoner that allows the architect to ask the question: "Have I captured all that is relevant to the enterprise?"

The C4ISR Architecture Framework (DoDAF) is intended to ensure that the architecture descriptions developed by defense commands, services, and agencies are interoperable between and among each organization's operational, systems, and technical architecture views, and are comparable and can be integrated across joint and combined organizational boundaries through a series of operational (OV), system (SV) and technical (TV) views.

Constructs such as these can be used in a number of ways in conjunction with experimentation. Primarily they can be used to describe the processes being experimented with in terms of information nodes, data flows and a range of different inter-relationships. This helps the experimenter formulate better hypotheses and develop MoEs. This type of modeling is also excellent for the definition of candidate solutions and experimental treatments in precise terms. Some of the software tools that enable the creation of architectural products using standard frameworks also allow dynamic process modeling to be performed using the completed architectures. When using a model-exercise-model paradigm it is often more straightforward for the "model" element of the process to be of this type rather than a complex, constructive combat simulation.

10.7 Summary

Modeling and simulation of many types has become pivotal in defense experimentation, to the extent that many defense experiments simply would not be practical without it. However, the means of selecting the types of M&S to be employed and validating that they are fit for the purpose of supporting a particular experiment are complex issues. Although a number of standard and well-documented methods exist, it is important to maintain a sense of perspective at all times: the **spirit** of the laws of validation is often more important to follow than the **letter** of the law. A key message is that the adequacy of M&S does not always increase with fidelity, and often, simple non-combat simulations can be the best solution, particularly when using a model-exercise-model paradigm.

Principle 11.

An effective experimentation control regime is essential to successful experimentation

Principle 11 affirms that without an effective experiment control regime, an experiment will likely fail to deliver valuable results because control must be applied from inception to execution (concept development, experiment design, resource arrangements, experiment execution and data collection, analysis, and reporting).

Experimentation is intrinsically a controlled activity, although the degree of possible and required control varies from case to case. The experiment design should be explicit in describing which variables must be controlled and which ones can be allowed to remain uncontrolled though usually recorded. It should also describe the control regimes to be put in place to ensure that this occurs in practice. The identification of intervening variables and learning effects must be well understood. However, simply outlining the required measures in the experimental design document is not sufficient. The experiment director and his team must actively seek to impose the required controls throughout the planning and execution phases of the experiment.

For example, joint and operational level experiments typically focus on new organizations, processes, and new technologies to deliver joint capabilities. These are often large experiments requiring 12 to 18 months of planning and up to several hundred people to execute. Coalition experiments may be distributed involving groups working at different sites employing a computer network to communicate and collaborate on the subject of the event. Distributed events are very complex to manage, thus proactive management is required throughout the planning and development process to ensure that there is a single unified agenda among all of the participating organizations. Control during the planning phase is particularly challenging in multinational experiments. Closer to the event itself, rigorous briefings must be carried out to ensure that all of the control staff are familiar with the control regime and, importantly, **why it is necessary**, so that they understand the experiment designer's intent. The detailed plan of who is to do what, and when, must accurately reflect the needs of the experimental design.

During the execution phase, the roles of the experiment director and his team are crucial to keeping the event on track and under the necessary level of control. Events within the experiment scenario can sometimes drift beyond the limits defined by a particular experimental treatment and this must be identified and corrected as soon as possible if the validity of the experiment is not to be compromised.

In summary, defining the experiment controls is primarily a scientific activity to be undertaken during the design phase. Implementing those controls is a complex management activity, which often requires military operational authorities and needs to be undertaken during both the planning and execution phases.

Principle 11. An effective experimentation control regime is essential to successful experimentation

The reader is encouraged to apply and adapt this Principle. However, the examples herein are based on the specific perspective and experience of different lead-nation authors with contributions from other participants. Therefore they may require supplementary effort to relate them to national perspectives.

11.1 Introduction

Experimentation is intrinsically a controlled activity, although the degree of possible control varies from case to case. The experimental design should be explicit in describing which variables must be controlled and which ones can be allowed to remain uncontrolled. It should also describe the control regimes to be put in place to ensure that this occurs in practice. However, simply outlining the required measures in the experimental design document is not sufficient. The experiment director and his team must actively seek to impose the required controls throughout the planning and execution phases of the experiment.

For example, joint and operational level experiments typically focus on new staff organizations, operational planning processes, and new technologies to support an operational or joint headquarters. These are often large experiments requiring 12 to 18 months of planning and up to several hundred people to execute. Coalition experiments may involve groups working at different sites employing a computer network to communicate and collaborate on the subject of the event. These experiments are distributed events and are very complex to manage. Proactive management is required throughout the planning and development process to ensure that there is a single unified agenda among all of the participating organizations. Control during the planning phase is particularly challenging in multinational experiments [ABCA 2004; AUSCANNZUKUS MONIME 1993]. Closer to the event itself, rigorous briefings must be carried out to ensure that all of the control staff are familiar with the control regime and, importantly, why it is necessary, so that they understand the experiment designer's intent. The detailed plan of who is to do what, and when, must accurately reflect the needs of the experimental design.

During the execution phase, the roles of the experiment director and his team are crucial to keeping the event on track and under the necessary level of control. Events within the experiment scenario can sometimes drift beyond the limits defined by a particular experimental treatment and this must be identified and corrected as soon as possible if the validity of the experiment is not to be compromised.

In conclusion, defining the experimental controls is primarily a **scientific activity** to be undertaken during the design phase. Implementing those controls is a complex

management activity, which needs to be undertaken during the planning, execution and even the analysis phases. Therefore, a large investment in design and planning involving all of the experiment stakeholders is required. Design and planning are often confused, but they are identifiably different functions. Viewing them as different is an important principle in the control of the experiment's development. The experiment design is an articulation of what the designer wants to achieve in order to meet the experiment's objectives. It can be said to be the preparation for the analysis phase. Planning, conversely, is the practical implementation of all aspects of the experiment in advance of the execution phase, to enable the design to be realized. It can thus be said to be the preparation for execution. Control aspects of design, planning, execution and analysis are now discussed in turn. The difficulty of achieving the required controls depends on the size and complexity of the experiment or campaign in question. This Principle focuses on the most difficult, large and complex end of the scale, but the general principles can be applied in a more modest form for smaller events.

11.2 Experiment Design

The experimental design process is a logical journey from the questions to be answered, or hypotheses to be tested, to the detailed definition of the experiment. Thus the experimental design is the cornerstone of the control regime throughout the life of the experiment, since it sets out in broad terms what needs to be done. From the treatments and trials it will derive the detailed day-to-day program and also provide the foundation for the data analysis and collection plan, which has an audit trail back to the design. It is a fundamental principle that nothing about the experiment should have a life outside of its design.

Success in designing experiments is rooted in early stakeholder engagement to establish objectives and intent. An integrated analysis and experimentation campaign plan goes a long way toward providing the framework for detailed stakeholder guidance. Furthermore, nothing allows for the control of variables during experiment design more than early, firm decisionmaking. The longer decisions on scenario, participation, funding, technical environment, and study issues are allowed to linger, the more options the experiment designers must keep open and the harder it is to control the variables that can affect the outcome of the experiment.

The tendency, especially in high-visibility experiments, to layer multiple concepts and capabilities into an experiment over time must be resisted. While refinements can continue to be made throughout the planning of an experiment, there must be a point in the design process when the major components of an individual experiment are locked down. This has been called the "good idea cutoff point" (GICP). This colloquialism has given many the false notion that, once the GICP is reached, no changes can be made to the experiment. This misperception hampers the detailed planning and decisionmaking that must go on when preparing an experiment. Changes to the design can be made after the GICP, perhaps as a result of practical considerations arising during planning, but these should be moderate amendments to a

"chilled" design rather than fundamental alterations. In fact, the design must be stabilized in order to conduct the experiment.

Too often, the range of questions that are suggested to be addressed within an individual experiment far exceeds the analysis and data collection resources available or the designers' ability to avoid mutually confounding factors. There is also sometimes a tendency for the objectives of an experiment, especially one of high visibility, to serve more purposes than just to frame the analytic effort. Objectives can be written to encompass political intent, public affairs intent, and training intent. As a result of a wide range of competing intents and demands, the analysis effort can become spread "a mile wide, but an inch deep."

Instead, the design of an individual experiment should be carefully focused on those areas and study issues for which that experiment was specifically intended. It is the responsibility of the experiment director to ensure the proper orientation to cause the study team to drill down into a focused set of questions (Figure 48). This process of bounding the issue set and focusing the analytical effort contributes to the control of variables in an experiment.

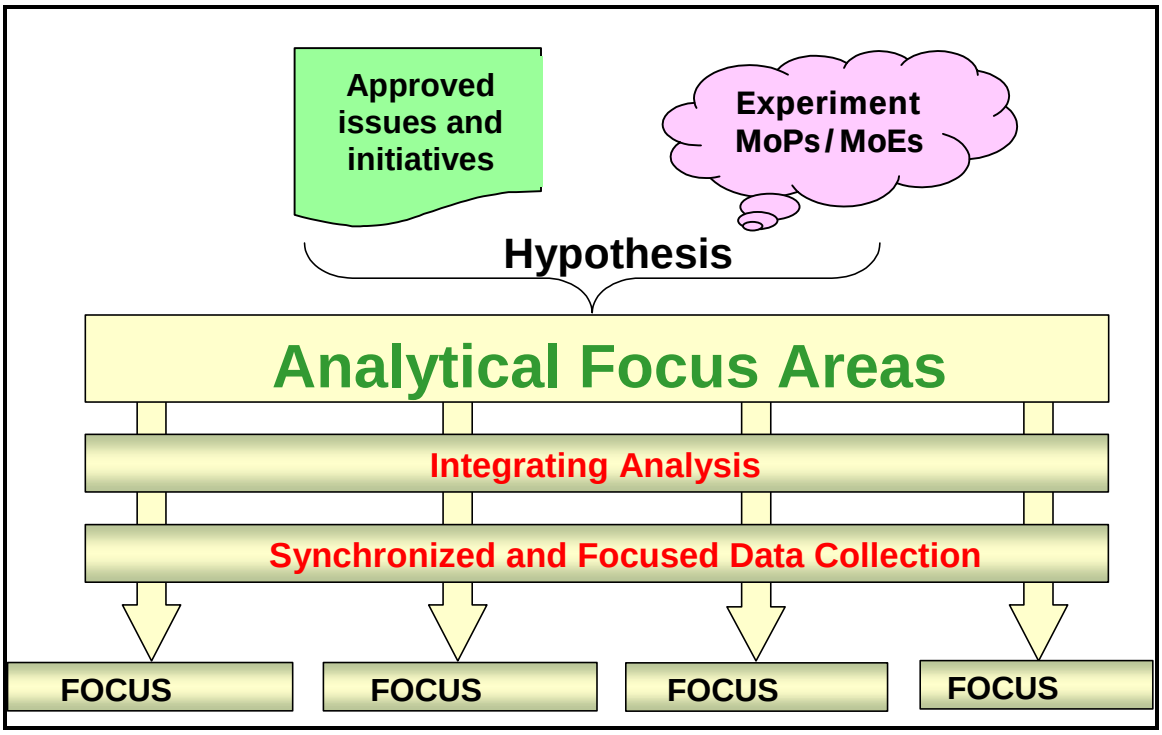


Figure 48 Focusing the analysis effort

Focused experiment design also requires the support of the concept developers or other experts in the subject matter being experimented with. Analysts can never be as well versed in all aspects of each concept as the SMEs are. The SMEs, therefore, become critical resources to the process of the analyst understanding the subject matter or concept, and understanding the linkages among issues. Care must be taken in this relationship to ensure that the analyst's design retains an unbiased examination of the subject matter.

11.2.1 Concepts

Those characteristics of the concepts or capabilities, or other subject matter, that are the focus of the experiment must be clearly defined early and in sufficient detail for the experiment. All other aspects of experiment planning and design flow from the definition of the subject matter being experimented with. Analysts cannot complete the development of the experiment design without an understanding of this subject matter. Similarly, scenario developers cannot ensure the experiment scenario provides the appropriate context within which to analyze the subject matter, nor technical developers ensure that all aspects of the experiment environment appropriately replicate the subject matter at the required level of fidelity.

11.2.2 Scenarios

The scenario defines the operational context for the experiment. As such, it has a major impact on the design of the experiment and its results. Therefore, the determination of the base scenario to be used for an experiment must be made early and must be made with a careful consideration of the experiment purpose and objectives. Specific areas that require early decision (because they often require long lead times to implement in the experiment environment) are the terrain box and level of detail required, the base force structure and equipment list, and the level to which units must be represented. The experiment director or his representative should be a part of the team that determines the scenario and refines it for use in the experiment, and advise on how nuances in the scenario will affect the analytic results.

Scenarios, including the input data for the tools the scenario is represented in, must be consistent across the integrated analysis and experimentation campaign plan so that results can be effectively compared across individual campaign elements. They should be consistent with relevant national defense planning guidance and must be approved by the authorities managing the campaign plan. Individual participants in a campaign should not be allowed to create their own scenarios for their contribution. Additionally, a spectrum of scenarios must be employed in a coordinated manner to illuminate results across an array of situations and missions. Experiment scenarios must also enable the accurate representation of concepts and capabilities outside the domain of the experimenting organization (*i.e.*, Army experiments must accurately reflect appropriate joint capabilities as they apply to a particular event).

11.3 Experiment Planning

11.3.1 Planning Reviews and Conferences

The planning of major defense experiments requires a management team, which takes the decisions required to settle high-level issues, has oversight on the activities of the various teams, and ensures that experiment planning and organization develop toward the objectives in a timely manner. A series of reviews throughout the planning period is usually necessary to ensure that the process of preparing for the experiment is remaining on track. For larger experiments, *e.g.*, joint or coalition ones, it is common to

employ conferences for this purpose, organized and run by the management team. In addition to any conferences to develop the concepts or capabilities being experimented with, typically three experimentation planning conferences should be used.

The *Initial Planning Conference* deals with planning the experiment in full detail. Large experiments will require that planners subdivide into groups by task or specialty. The groups may include concept development, scenario development, experiment design, technical systems development, and experiment assessment or analysis. There may be groups dedicated to special issues like network support or a particular concept that requires a dedicated working group to evolve.

The *Main Planning Conference* is where all work required for conducting the experiment is confirmed. At this point progress will be reported for all significant activities. Concepts, concepts of operations (CONOPS) and TTPs should be complete and available to all organizers. All technologies should be well progressed and deadlines for testing and rehearsals should be confirmed. Any problems or shortfalls should be identified and the experiment leaders should be prepared to curtail any elements that are not progressing properly. This may lead to changes in the experiment design. No major deficiencies should go unresolved beyond this point in the planning process.

The objective of the *Final Planning Conference* should be to confirm that all of the preparations are progressing properly, that the CONOPS and TTPs have been distributed and no major issues have arisen, that technical development is on schedule and that all equipment and software is expected to be ready. The planning for rehearsals and training events should be discussed and confirmed at this conference.

11.3.2 Planning Activities

There are three main components that make up the planning activities: the participants, the technical architecture, and the data collection planning. Each must be specifically developed to support the objectives of the experiment. All three require long lead times to ensure that they are prepared, tested, and rehearsed properly.

11.3.2.1 Participants

The experiment participants consist of the players, the experiment controllers, the analysts, and the technical support team. All must be thoroughly trained on the subject matter (*e.g.*, concepts or capabilities) that is the focus of the experiment. Controllers, analysts, and the technical support team must also completely understand the focus of the study effort embedded within the experiment. Participants must not only understand the construct of the subject matter, but also the current status of its evolution. The subject matter must be stabilized well before the participants begin training. Stabilization is required to allow all participants the time to achieve a higher degree of understanding so that military capabilities are implemented effectively during the experiment. Continued changes only cause confusion and negatively impact the analytic rigor desired. It is also beneficial to the consistency of an analysis and experimentation campaign plan if the majority of participants remained constant over the course of the campaign. This would gradually reduce training and rehearsal time, as

well as develop a higher order of knowledge and understanding about the state and direction of the campaign plan, which would result in better feedback from the experiments.

11.3.2.2 Technical Architecture

The technical architecture is the global combination and denotation of the following elements used to support an experiment: communications systems, computer support systems, networks, workflow plan, databases, models and simulations. The two most significant challenges for the technical team are defining the requirements from the concept developers and experiment designers and then integrating the various systems into a workable experiment environment. The experiment concept will frequently call for a system to emulate a future capability. The designers may identify or develop a simulation or tool to provide this capability. Typically, the technical team will also be implementing a suite of systems or tools to provide a mock-up for the operation of a headquarters. Tools like JSAF may also be used to simulate Red, Blue and Neutral forces and this information would be provided to the tools supporting the headquarters through C4I interfaces. The technical team will typically have the largest task load in the experiment. Therefore, this team should be provided as much time as possible to conduct their work. The technical team requires a clear plan and objectives as early as possible from the experiment designers. As discussed earlier, the subject matter must be stabilized early and in sufficient detail for technical development to proceed. Then, SMEs must work closely with technical developers to share their detailed vision of the subject matter so that it can be represented properly. The technical team should identify risk areas in their program early and communicate cut-off dates to designers, bringing tools or other components to the experiment on or after the *main planning conference*.

Finally, for each experiment, there must be a well-constructed process for integrating, testing and verifying the components of the technical architecture. This process must be well managed using a step-by-step approach building from testing individual elements through to the performance of the entire system in scenario vignettes. Concept developers and other SMEs must participate in the testing process to verify that the results being generated within the technical architecture adequately represent the current state of the subject matter under investigation. This must be done and documented properly as well so that analysts understand the strengths and weaknesses of the technical architecture and can bound their analyses appropriately.

11.3.2.3 Data Collection

The planning of the data collection follows from the analysis plan, which is driven by the hypothesis (or objective) and by the design of the experiment. Automatic data collection systems, surveys, and observers are all approaches that can and should be employed. The data collection plan must strike a balance between efficiency and completeness. The plan should specify reasonable requirements that will not overburden the participants and collectors (or collection systems) and yet will not create undue risk for the analysis.

The planning process should include frequent consultation with the SMEs and, as discussed earlier, the design and analysis teams must work very closely together. The data collection planning starts with the preparation of the analysis objectives, followed by detailed examination of the CONOPS, TTPs and the technical architecture. The analysis team should therefore interact frequently with the technical team as well.

The data collection plan should be prepared in time for testing during rehearsals. The final plan is not required until shortly before execution. It is important however, to ensure the entire team reviews the plan and that all the analysis objectives can be supported from the data. Planning an experiment is a long process and final checks are important to ensure good coordination.

Given that many defense experiments attempt to examine concepts that are not developed sufficiently to be assessed in a purely quantitative manner, combined with the fact that many aspects of warfare cannot be modeled explicitly, most experiments commission observers of one type or another to collect data. It is often difficult to acquire qualified observers. Furthermore, the crew of observers is different in just about every experiment and the increased acuity from consistent observers who build a personal knowledge base of subject matter understanding over the course of many experiments is never gained. It is sometimes more effective to employ analysts, designers and SMEs, supplemented by a core pool of contractors, in an effort to maintain consistency over the course of a campaign.

11.4 Experiment Execution

The experiment management team usually transforms into the control staff during execution. The controllers' roles are to ensure that the experiment is progressing according to schedule or to be on top of the situation if it is not. The controllers observe the players, collect their input daily and work closely with the analysts in monitoring the progress of the experiment. The controllers provide feedback to the experiment director and implement changes as required to ensure the event achieves the experiment objectives. In performing this function, the controllers should be addressing cause-and-effect with the analysts and the experiment director as changes to the experiment are developed. In doing so, the controllers must deal with military judgment (observations from the players) and scientific objectivity (inputs from the analysts).

The experiment controllers and analysts must develop an effective battle rhythm in order to observe, communicate and direct effectively. In large experiments, and especially in distributed experiments, daily meetings are required with clearly defined reporting formats and responsibilities. There are many administrative and exercise control functions involved in the execution of an experiment, but they will not be addressed here in detail. For complex experiments it is worth considering the use of a networked exercise management tool. Such tools can contain background documents such as the data collection plan and also maintain up-to-date information on the progress of the scenario or master events list and any recent exercise control pronouncements from the experiment director. For experiments running continuously

(*i.e.*, 24 hrs a day) such tools can be particularly useful for situating each new shift of controllers and data collectors.

The actual execution of an experiment normally entails at most 20% of the overall time allocation in an experiment (design/planning = 40%; post-analysis and reporting = 40%). During execution, the bulk of effort is put into the accurate, complete collection of the data required by the experiment design. The experiment director must ensure that collection activities meet the requirements of the experiment design during various periods of execution. Data collection capabilities must be integrated across all available sources (observers, analysts, simulations, instruments, prototypes and platforms).

A detailed data collection plan will have been produced during the planning phase. However, for a large and complex experiment, putting the plan into practice requires far more than just reading it. If at all possible, the data collection team should exploit rehearsals and other pre-experiment tests to verify the data collection regime. Analysis "pilots" should be considered if possible to confirm the data collection requirements and data analysis methods.

11.5 Experiment Analysis

The analysis or assessment team for an experiment should ideally be derived at least partly from the experimental design team, and they should work closely with the subject matter and technology teams. During the course of an experiment, analysts compare observations and results and begin to integrate their views of what is being learned from the experiment. As sufficient data are collected, analysts begin to form preliminary insights. These preliminary insights are not based on completed analysis, but they are of sufficient fidelity to spark more focused discussion about the experiment results. Depending on the type of experiment and the desires of the experiment director, preliminary insights can be used to frame periodic discussions among senior participants in the experiment. These discussions should be small, closed forums due to the preliminary nature of the results; however, they are useful in adding professional military judgment to the initial results being contemplated by the study team. They also assist the leaders of the experiment in providing focused reasonable feedback to other senior leaders regarding what appears to be coming out of the experiment.

Preliminary insights can form the basis for the development of the first report from an experiment, the *"initial insights"* or *"first impressions"* report. This report, while based on the results of the completed experiment, is not the result of completed post-experiment analysis. It can take the form of either a short document or a scripted briefing. It can be either delivered hard copy or briefed to senior leaders and then followed with a document. The goal is to publish this report while the experiment is still fresh in the minds of the leaders and while it is still being discussed to frame the ongoing professional debate on the results.

However, the temptation to announce some startling finding (especially one that it is believed the experiment sponsor will like) should be resisted at all costs, because it is

quite likely that when the analysis is complete, that finding will at best need to be modified, and at worst, changed altogether. Thus, first impressions should generally be conservative; this is an important control consideration.

Principle 12.

A successful experiment depends upon a comprehensive data analysis and collection plan

Principle 12 upholds the importance of adequate data analysis and collection planning since this directly affects the knowledge that can be gained from an experiment or series of experiments. Proper instrumentation and metrics lead to richer data and better observation of expected and unexpected behaviors and responses to control variables, which may clarify the causal relationship between two variables. Without appropriate data collection, analysis becomes futile. The nature of the experiment and the models at play dictate the data analysis and collection requirements. This is especially important with complex systems where one needs to define what needs to be seen in order to devise instruments to observe the behavior or phenomenon of interest.

Data collection is designed to support the experiment analysis objectives which rely on a conceptual model that underlies the experiment. The model should support the investigation of the hypothesis, which is determined in the experimentation design phase. The hypothesis and the model are used to define the measures of performance (MoP) and/or measures of effectiveness (MoE). The data analysis offers the opportunity to revisit the conceptual model developed for the experiment and determines cause-and-effect relationships. For causal hypotheses, controls are necessary to eliminate plausible rival explanations and these need to be considered in the data analysis and collection plan.

The data analysis and collection plan is an essential part of an experiment. The data management architecture of the experiment must be understood and should be exploited in a good data collection plan. The plan ensures appropriate and valid data are generated and that the key issues of the experiment are addressed. When determining analytical techniques to use, an estimate for the number of observations must be considered to allow for the possibility of statistically significant findings from the experiment. It is essential to prioritize and ensure there are sufficient observations for all objectives, MoPs, and MoEs requiring analysis. It is possible that other relevant measures will be discovered during the analysis process.

All analysts should be involved early in the development of the plan. The experiment design and analysis staff should work closely together or be part of an integrated team. Meetings, conferences (or teleconferences), and workshops should all be employed. Early planning for these events is important.

Principle 12. A successful experiment depends upon a comprehensive data analysis and collection plan

The reader is encouraged to apply and adapt this Principle. However, the examples herein are based on the specific perspective and experience of different lead-nation authors with contributions from other participants. Therefore they may require supplementary effort to relate them to national perspectives.

12.1 Introduction

Data collection is designed to support the experiment analysis, which develops from the conceptual model underlying the experiment. The model should support the hypothesis, which is determined early in the experimentation design phase. The hypothesis and the model are used to define the measures of performance (MoP) and/or measures of effectiveness (MoE). The data analysis offers the opportunity to revisit the conceptual model developed for the experiment to determine cause-and-effect relationships. For causal hypotheses, controls are necessary to eliminate plausible rival explanations and these need to be considered in the data analysis and collection plan.

The data analysis and collection plan is an essential part of an experiment. The information or data management architecture of the experiment must be understood and should be exploited in a good data collection plan. The plan ensures appropriate and valid data are generated and that the key issues of the experiment are addressed. When determining analytical techniques to use, an estimate for the number of observations should be considered. It is essential to prioritize and ensure there are enough data points for all objectives, MoPs and MoEs that require analysis. All analysts should be involved early in the development of the plan. The experiment design staff and the experiment analysis staff should work closely or be part of an integrated team. Meetings, conferences (or teleconferences), and workshops should all be employed. Early planning for these events is important.

12.2 Data Collection

Data analysis and collection plans are an essential element of experimentation. They ensure appropriate and valid data are generated, obtained and organized and that the key issues of the experiment are addressed. Data analysis offers the opportunity to revisit the conceptual model developed for the experiment.

All analysts need to be involved early in the planning process and well integrated in the other teams setting up the experiment. Staff continuity for experiments, and even for an entire campaign, is beneficial.

When determining analytical techniques to use, the number of estimated observations needs to be considered. The technique will probably be different if only a few observations are likely to be made. It is essential to prioritize and ensure there are enough data collection points on all objectives, MoPs and MoEs that require analysis. Since simple correlations can neither prove nor disprove a causal hypothesis, it is essential for causal inference that data be systematically collected on the potential rival explanations for the hypothesized causal relationship so that controls may be introduced in the analysis and these rival explanations (*e.g.*, training levels) ruled out or otherwise accounted for.

12.2.1 Basic Data Collection Process

A significant part of the experiment consists of gathering data and information. Interpreting the information into findings and combining them with already known information to obtain new insights tends to be challenging. Once it is determined what needs to be measured, a decision is required to identify the data necessary and to analyze it using analysis techniques (*i.e.*, statistics such as analysis of variance, chi-square tests, *etc*.).

There exist various types of collection mechanisms used in experiments. The following ones will be discussed:

1. questionnaires/surveys,
2. automated collection systems,
3. observers, and
4. scheduled meetings.

12.2.1.1 Questionnaires/Surveys

Questionnaires (also referred to as surveys) are often used in data collection. They offer a convenient way to collect data in an easy-to-review format. This section provides information about how to design questionnaires and factors to keep in mind.

Questionnaires can be used to gather numerous types of information. The participants' background can be obtained through this means. This can be done before the start of the experiment. The participants can be questioned about aspects of the experiment such as their perceptions about the systems and processes under test, asked for their views on other systems and processes supporting the experiment, and be allowed to recommended improvements. Others involved with the experiment (*e.g.*, experiment director, controller, senior concept developers, analysts) can also be provided with questionnaires. They can often provide different views about the systems and processes involved in the experiment as well as their overall perceptions on the execution of the event. Some questions should be included to allow comments about areas not specifically asked in the survey. It is a good idea to pre-test the questions to verify that there is no misunderstanding. This will ensure quality data by verifying that the wording is unambiguous and that there is no misinterpretation of the questions.

Likert scales[38] are useful because they structure the data and require less effort to analyze. The answers can be machine read, however, it is recommended to provide a comment space for each question. This allows the participant to clarify responses. For example, if the only choices to a question were yes and no, the comment box could be used to indicate "not applicable" if the question wasn't relevant to the respondent. Although having comments implies more data to go through, so a longer time to analyze, getting the participants to write lots of comments ensures their answers are interpreted properly and their point of view clearly understood.

Filling out questionnaires takes time. It is thus suggested that time be allocated during the experiment to complete the surveys. This will help ensure the participants actually answer all the questions without being rushed. Questionnaire loading should also be considered. There is concern that too many questions will be asked, that the questions might not be relevant, or that they could appear to be repetitive. Two factors about the relevance of questions, which should not be ignored, include asking questions only to the appropriate participants and ensuring questions are asked at the right stage of the experiment. This will help ensure the participants do not get frustrated with the surveys and that the quality of their answers does not decrease.

Some of the data collection, especially on the human factors side, can be obtained by having the participants answer probes throughout the execution. These are typically questions asked at specific times throughout the day and can be used to assess their situation awareness. One important point to keep in mind is to ensure the probes don't cause interference with the experiment (or at least minimize it). The timing of the probe questions should be considered because they can inadvertently be set up for a time where intense discussions are taking place. If the probes are asked then, the execution is interrupted and starts up after a small break. This can affect the way the discussions are going. If the probes are delayed, this can interfere with the overall results. A trade-off needs to be agreed upon.

An important new technology for providing and collecting questionnaires and surveys are web-based tools. Tools such as JDCAT (JBC Data Collection and Analysis Tool) have significantly reduced the work for analysts and made the chore somewhat more convenient for the experiment audience. The collation of all of the data in a central database has made the data management immensely easier and far more accurate. Tools like JDCAT, for example, also include analysis tools and can output data to MS Excel and MS Word.

12.2.1.2 Automated Collection Systems

With information systems becoming more crucial, automated collection of data is now more important. If automated collection is to be used, expertise in the system is required. Knowledge of the data and variables to be collected is also essential. It is

[38] This refers to a widely used questionnaire format named for its developer, Rensis Likert. Respondents of questionnaires are asked to choose from several responses in a range such as 'strongly agree', 'agree', 'undecided', 'disagree', and 'strongly disagree'. Each response receives a number rating (*e.g.*, 1-5). The five-point Likert scale is most common. http://www.cirem.co.uk/definitions.html#l

important to determine what clock each system that is used to collect data is synchronized to in order to facilitate analysis. Using Zulu time can help with this process, especially when different time zones are involved. An example of a data harvesting system, data logger, is REPEAT (REpeatable Performance Evaluation and Analysis Tool). REPEAT collects all incoming messages into GCCS, writes them to a file and provides analysis tools to investigate these files. Some platforms (such as warships or larger aircraft) can be fully instrumented with video recorders, CRT display recorders, and data taps on computerized systems. The danger here is data overload and experiment designers must exercise judgment in the data requirements for their MoPs and MoEs.

12.2.1.3 Observers

Observers play an important role in the experiment by capturing the interactions between participants. For instance, they take notes about what is going on, crucial events taking place, notable behaviors and other such activities. To perform their tasks efficiently, it is recommended that the observers be SMEs in the area covered by the experiment or at least have skills as an observer. They have to obtain the information they require without being intrusive to the experiment play. There should be time-set aside during the experiment to allow the observers to review their notes and possibly code the information they gathered. Otherwise, it is likely that the observers will not have been used to their full potential as their insights may not have been taken into account.

Observer-type personnel can also be used to provide a chronological narrative of the events that occurred. This involves noting significant activities that happen throughout each day. The time of occurrence, description of the event and people involved should be logged. This provides documentation about what happened during the experiment and can be used to explain why certain results occurred. It has to be determined before the start of the experiment who will record the chronological narrative and how they will go about performing this task. Interviews with the participants can help add to the narrative since some events were possibly not observed directly by the person in charge of observing this activity.

The number of observers available is not always sufficient, especially when they are required in numerous locations. As well, they are only able to gather a certain amount of information. Thus deciding what data to capture becomes an issue. Audio and video recordings can be useful as well as saving electronic chats, which provides a written record of discussions. However, going through the recordings requires extensive time and effort. This fact must be taken into account when deciding whether or not to use any type of recording.

Greybeards and Senior Concept Developers (SCDs) are other types of personnel that can be employed. As the names imply, they should have extensive expertise in an overall concept or process being addressed during the experiment or enough experience to provide ideas and opinions. They can provide insights about concepts, processes and procedures that the direct participants might not have considered.

Normally being at the rank of commodore/brigadier-general or above, their background and experience is invaluable in these areas.

Greybeards and SCDs can provide input at the national and multinational level. At the national level, they can address how the country could potentially use the concept at hand in the future. In a multinational environment, collaborating with Greybeards and SCDs in other nations can help identify issues affecting the way coalitions would employ concepts. SCDs can provide views on what parts of the concept worked well, what parts needed improvement, and make recommendations. One of the main outputs SCDs can provide are papers on the topic at hand. Asking open-ended questions in a survey improves the collection of SCDs' inputs.

12.2.1.4 ***Scheduled Meetings***

Throughout the execution of the experiment, a good structure for analysis involves scheduling sessions to see where things stand. Daily hot washes, azimuth checks and after-action reviews (AAR) are such sessions. Regular get-togethers like these are required to ensure everything is on track and allow discussion for any changes or modifications that might be necessary. Regular meetings are required which should be tailored to the venue. These sessions require considerable preparation time in gathering thoughts, writing presentations, *etc*. The experiment battle rhythm needs to include review and preparation time. The extent of the work involved needs to be fully understood beforehand and clearly indicated. These sessions are extremely valuable but take up considerable resources, which need to be properly planned. A description of all sessions, meetings, and presentations that will be required throughout the course of the experiment should be provided well in advance to ensure appropriate resources are available.

12.2.2 Data Collection in Distributed Events

Coordinating an experiment involves many challenges. One of them is getting everyone involved to be aware of the status of the experiment and work together on issues. Working in a multinational experiment, and often a national experiment, means that personnel are not always co-located. To ensure no one is out-of-the-loop and provides input into the experimentation process, coordination is essential. There are various methods to promote feedback and interaction with all interested parties. The following ones are briefly discussed:

1. teleconferences,
2. distributed collaboration tool,
3. planning conferences, and
4. analysts' workshop.

12.2.2.1 Teleconferences

Teleconferences provide the opportunity to keep in touch on a regular basis and allow verbal discussions instead of only written ones. It is easier for some to express their

ideas, opinions and points of view orally rather than in writing. As well, teleconferences offer the chance to give feedback within a short timeframe. It also forces those involved to follow up on issues on a regular basis, in preparation for the teleconferences. The interval depends on the experiment but weekly teleconferences can be useful in large multinational experiments.

12.2.2.2 Distributed Collaboration Tool

In the time leading up to the experiment, a way to share information is to use a distributed collaboration tool (*e.g.*, Groove, Info Work Space, Defense Collaboration Tool Suite (DCTS)). This allows analysts to place files for others to view and comment on, edit common files, post messages for all to view and perform other related activities to ease the collaboration process.

12.2.2.3 Planning Conferences

To ensure the experiment is on track and that the various groups (design, analysis...) are in sync, periodic planning conferences should take place. These allow presentations and face-to-face discussions. Another benefit is the human factors aspect. Meeting people often helps develop good working relationships. Once the relationship is established, collaboration becomes easier. Trust seems to grow with such meetings.

Beside having sessions where multiple groups are present, planning conferences offer the opportunity for analysts to discuss issues in breakout sessions. They can thus further their assessment plan for the experiment.

Larger experiments typically have a concept development conference (CDC), an initial planning conference (IPC), a main planning conference (MPC) as well as a final planning conference (FPC). There may also be other conferences such as a pre-CDC. They are often scheduled a month and a half to two months apart to allow all involved to follow-up on the issues and action items identified.

12.2.2.4 Analysts' Workshop

Once the experiment is over, personnel usually return to their normal place of business and work from there. This is particularly true in multinational experiments. This means that the analysis is done by different groups and not necessarily in the same way. It is practical for the analysts involved to get together a few weeks afterwards to discuss their results. Conducting an analysts' workshop has many benefits. The analysts can compare their results and discuss any differences as well as explain certain points that were observed.

It is useful to distribute the workload (computer supported and collaborative) during the analyst workshop. Holding it at an off-site neutral location ensures that none of the analysts are distracted by normal-work issues, and that no one interferes with the analysts' discussions or tries to sway them in a certain way. Enough time has to be set aside for each to bring up their respective analysis conclusions and any additional points they have. This is an opportunity for all to discuss what they saw and what results were obtained from the experiment. If a multinational report is to be written, the workshop

can be an opportunity to co-evolve the final draft report while all are present. Different ways of dividing the analyst work for a multinational report are a challenge. JFCOM's J9 has experience in this area and the reports from Multinational Limited Objective Experiment II (MN LOE 2) and Multinational Experiment 3 (MNE 3) provide useful examples.

12.2.3 Metrics in a Socio-cognitive Environment

This section provides an overview of the discussion on measuring performance and effectiveness in the cognitive domain contained within Appendix A of the CCRP COBP for Experimentation [Alberts and Hayes 2002]. The goal of military experimentation and modeling is to develop an understanding of how specific transformation initiatives relate to improvements in performance and effectiveness. In the past, military experimentation has focused mostly around the assessment of technology, using traditional measurement approaches from the engineering sciences. Yet, the emergence of concepts such as Network Centric Warfare, Information Age Warfare, and Effects-Based Operations have placed increasing importance on how well this technology performs within a more complex socio-cognitive setting. Evidence [Alberts and Hayes 2002; NATO 2002] from the past few years suggests that these traditional measurement approaches have yielded little more than anecdotal insight into how technology and human performance combine in either productive or unproductive ways. Such data, unfortunately, does not provide a quantitative foundation for assessing the return-on-investment of various transformation initiatives. As a result, experimenters have begun to search for more appropriate, quantitative methods of data collection that can be usefully applied to addressing performance in the cognitive realm.

To this end, it is useful to address the methods for quantifying performance in the social sciences. Such methods have matured over the past several decades to provide socio-cognitive research with the same types of statistical analysis and modeling tools used in the physical sciences. A number of behavioral observation methods exist and can address a variety of dimensions of performance involving humans, organizations and technology. These methods were designed to capture subject matter expert judgments of task performance in the form of a standardized set of quantitative measures.

Two of these methods, which are relevant to military experimentation, include Behavior Observation Scales (BOS) and Behavioral-Anchored Rating Scales (BARS). These methods can be used as direct observation tools during an experiment, or applied via structured interviews after an experiment to quantify important dimensions of performance in the cognitive domain. Behavior Observation Scales are basically a checklist of observable, critical behaviors that correlate with acceptable task performance. The Behavioral-Anchored Rating Scale is used to assess the degree to which a task dimension is performed, typically on a 3-, 5-, or 7-point scale that extends from unacceptable performance to outstanding performance. At each point along the scale, the levels of performance are "anchored" by detailed descriptions of what type of behaviors might be seen by an observer in a real-world task setting. Compared to the

BOS, a well-developed BARS provides more utility for conducting an in-depth analysis of cognitive performance observed during an experiment.

While BARS instruments are useful for assessing certain types of observable task performance, their utility is limited for assessing outcome measures in the cognitive domain. One method for organizing the collection of cognitive performance metrics involves the development of a critical event framework. Similar in purpose to BARS, this technique focuses the attention of data collectors on relevant events within the experiment. However, in this case, the critical incidents are often obtained through post-experiment interviews with the participants rather than being observed in real-time during the experiment.

12.2.4 Statistical Data Analysis

Principles 2 and 3 present the issues associated with the design of good experiments and show that the application of experiment control is essential to the successful examination of causal hypotheses. The onus is then on the data analysis and collection. Appropriate statistical data analysis is required to verify the causal hypothesis. Quoting Case Study 1 author, “Controls are key to causal analysis: good experiments provide them [controls] before the fact in the design; good analyses do them after the fact in the statistical data analysis. To ignore them is to deny the possibility of cogent causal analysis, for potential rival explanations for your findings will abound.”[39]

The causal interpretation of a simple (or partial) correlation depends upon the presence of a compatible causal hypothesis and the absence of a plausible rival hypothesis to explain the correlation on other grounds [Cook and Campbell 1979; Dagnelie 2003; Shadish *et al.* 2002]. What is always important when attempting to make causal attributions is the elimination of plausible rival explanations. Some evidence of causality can be obtained by controlling for confounding variables and ruling out plausible rival hypotheses. The third experiment validity requirement in Section 3.3 explained this further.

Data are typically analyzed using the general linear model (GLM) [Johnson 2000]. All special cases of the GLM are correlational [Kerlinger 1986] where the relations between variables are modeled. Techniques are available for controlling confounding variables. For example, confounding variables can be statistically controlled by collecting data on the key confounding extraneous variables and including those variables in the GLM. Similarly, the relationship between selected confounding and independent variables can be eliminated using matching or quota sampling approaches. Statistical control is usually preferred over individual matching. Analysis of Variance (ANOVA) and multiple regression and correlation (MRC) are both "special cases" of the GLM, as these are approaches to statistical analysis. For examination of non-linear relationships in the data, various transformation techniques are available, *e.g.*, log-linear transformation [Law and Kelton 2000; Shorack and Wellner 1986].

[39] Private communication, June 2004.

When looking at causality, experimenters should always address the three necessary conditions for cause-and-effect. The first required condition is that the two variables must be related. The second condition is that proper time order must be established (*i.e.* if changes in variable **A** cause changes in variable **B**, then **A** must occur before **B**). The third necessary condition is that an observed relationship must not be due to a confounding extraneous variable (*i.e.* the lack of alternative explanation condition). There must not remain any plausible alternative explanation for the observed relationship if one is to draw a causal conclusion.

Different techniques, approaches and methods can be used for analysis. Many fields of research deal with related issues previously mentioned. *Evaluation: A Systematic Approach* [Rossi and Freeman 1985] is an example in the field of social science. The book discusses evaluation research (or evaluation), which it defines as "the systematic application of social research procedures in assessing the conceptualization and design, implementation, and utility of social intervention programs." The approaches discussed can however be applied to other areas. Two of the main chapters deal with randomized design and nonrandomized design for impact assessment. Some topics discussed include: data collection strategies for randomized experiments, analysis of simple randomized experiments, complex randomized experiments, quasi-experiments, the use of generic controls in assessing impact, the use of statistical controls in assessing impact and supplementary use of statistical controls. The section on statistical controls points out that statistical control is an excellent procedure to apply when control variable measures can be entered that reflect competing explanations. While most people can understand the logic behind statistical controls, proper employment of the techniques is essential.

Statistics books can also provide information on the various techniques that can be used for analysis. Correlation and covariance are discussed in many publications (*e.g. Introduction to Probability and Statistics for Engineers and Scientists* [Ross 1987] and *Statistics* [McClave and Dietrich II 1991]). Other techniques such as regression and analysis of variance can also be found in such reference documents or in more advanced statistical publications. The mathematical side will not be discussed in this section. Readers are encouraged to consult other books for further information.

Foundations of Behavioral Research [Kerlinger 1986] is a book about scientific behavioral research. This book has many excellent sections for the experiment designer. The main purpose is to help understand the basic nature of the scientific approach to problem solving; technical and methodological problems being discussed. It features topics including scientific research; the relations between research problems and the design and methodology to solve them; and the concept of set, relation and variance as well as statistics and measurements. The various chapters address numerous areas such as hypotheses, variables, conceptual and mathematical foundations, probability, randomness, sampling, statistics, statistical inference, analysis of variance, designs of research, types of research, measurement, methods of observation and data collection, multiple regression, multivariate analysis, factor analysis and analysis of covariance structures.

12.3 Summary

It is evident to most experimenters that the data analysis and collection plan is an essential part of an experiment. Therefore, it must be emphasized that plenty of time is required to collaborate and develop this plan. The information or data management architecture of the experiment must be understood and exploited properly. A good plan ensures appropriate and valid data are generated and that the key issues of the experiment are addressed. It bears repeating that early planning is essential to the production of a well-coordinated effective plan.

Principle 13.

Defense experiment design must consider relevant ethical, environmental, political, multinational, and security issues

Principle 13 offers experimenters the basis for navigating the troubled waters of all human-in-the-loop defense experiments by addressing the relevant ethical, environmental, political, multinational, and security issues. We have to maneuver through high-classification cells to ethical, health and safety issues for any experiment involving human subjects and human data collectors, in which cases proper clearances and signed agreements must be obtained ahead of time. With the advent of the Combined Federated Battle Lab network (CFBLNet) these issues are exacerbated by differences between participating countries and consequently percolate at the international levels.

Other considerations include the potential environmental impact of any field experiment, the security level of the experiment, and the political and multinational issues.

Principle 13. Defense experiment design must consider relevant ethical, environmental, political, multinational, and security issues

The reader is encouraged to apply and adapt this Principle. However, the examples herein are based on the specific perspective and experience of different lead-nation authors with contributions from other participants. Therefore they may require supplementary effort to relate them to national perspectives.

Given the nature of TTCP and the current availability of the CFBL, the ability to conduct experiments across the laboratories of the participating countries becomes realistically achievable. Principle 13 describes possible issues for initiating the planning of human-in-the-loop experiments, particularly multinational ones. Distinctive regulations, security rules and national practices should not be underestimated and proper preparation must be put in place. We expect that such international activities will require several months of preparation. Reservation of resources and people may require booking 18-24 months in advance.

13.1 Political and Multinational Considerations

It is quite likely that by the time a multinational experimentation concept gets into the hands of various nations' practitioners, most, if not all of the political issues will already have been ironed out. However, the practitioner cannot afford to be too complacent because he may become involved in early conceptual work and planning, and differences in approach by the various participating nations may emerge during the design and planning processes.

The types of problems that can arise include:

1. **Synchronization of different nations.** Organizing major defense experiments is difficult even in a single nation. It is even more so for multinational experiments. These events take much more time to organize and require a higher degree of proactive organization and synchronization than single nation events. If the event is a real time distributed experiment, the issue of multiple time zones makes the execution phase more problematic than would otherwise be the case.
2. **Treaties and conventions.** Different nations have signed up to (or not) various international treaties or conventions (*e.g.*, anti-personnel landmines), which may preclude their involvement in some types of experimentation, or even inhibit agreement on scenarios.
3. **Scenario sensitivities.** Most nations have their own standard scenario sets, which are not usually shared. Although these are only intended to be indicative of the range of operations that their national foreign policy would imply, there are usually considerable sensitivities surrounding these. Thus scenarios for coalition experiments are normally something of a diplomatic compromise and may not be ideal from the perspective of some of the participating nations.
4. **Standards of experimental design and analytical rigor.** Nations should have a common lexicon of experimental terms and mutually recognized analytical techniques. However, this

cannot be guaranteed and individuals from each nation may have widely differing views on what it is reasonable to expect from coalition experiments generally, or for the particular experiment in question.

The four subparagraphs above indicate that there are many potential areas for disagreement in the concept development, experiment design and planning phases of a coalition experiment. These are borne of differences in foreign policy, culture and other national influences. The most important point to bear in mind is that attempting to hide such disagreements, for example by using vague or diplomatic forms of words that the differing nations can sign up to, is never a good policy. As plans become firmer and more detailed, the differences will inevitably resurface, and if taken through to the actual execution, can be a recipe for disaster. Ultimately it is better for a nation to withdraw, or even for the whole experiment to be completely recast, than for suppressed (significant) differences in opinion to endure to the later stages of planning and into execution.

13.2 Environmental Considerations

Wherever there is live activity, there will be some level of environmental impact. In particular, great care must be taken regarding proximity to historical or cultural sites. As well as legal and multinational environment issues, environmental constraints generally will have an impact on the scope of any live experiment or exercise. It is essential that results be interpreted in the light of all environmentally imposed artificialities. The test and training communities have been working with environmental issues for years and there is no reason for the experimentation community to deviate from the various protocols that already exist. However, there are some particular issues for coalition experimentation, and these can be outlined as follows.

1. Most coalition experimentation is likely to be in the form of coalition Command Post Exercises (CPXs), driven either by a straightforward master scenario events list or some form of analytic wargame or distributed simulation. In such events there are barely any environmental considerations, except potentially where operations being wargamed could have significant environmental impact and thus cause political sensitivity.
2. Environmental legislation varies widely between nations, so at its simplest this means that the participating nations must abide by the legislation in force at the experimental venue (if there is only one).
3. Assuming that the participating nations are coming together to a particular geographical location for the experiment, there can still be contention regarding environmental aspects, notwithstanding a common understanding of local laws. In particular, if a visiting nation has more stringent environmental laws than the host nation, they may feel uncomfortable participating in an experiment or exercise in which systems are being used by other nations (*e.g.*, the host) that would be banned by their own.
4. Another possibility is when a distributed live exercise or experiment is to take place, perhaps within the context of a “virtual map” so that the exercise play is simulated to be in the same geographical area. This concept has the benefit that individual nations’ live forces do not have to move across the globe to take part, and thus it might be expected that at least some coalition experimentation might take this form in the future. In this case, the effectiveness or tasking of individual nations’ forces could be heavily dependent upon local environmental considerations, for

example the size of training areas. This is clearly not a showstopper, but must be taken into account during design and analysis.

13.3 Security Considerations

Even within single-nation experiments, security issues can give rise to real practical problems. In particular, the rise of secure digital C4I and sensitive ISTAR sources (which are often themselves at the centre of the experimental purpose) has resulted in security considerations becoming much more prominent in the design and execution of defense experiments than hitherto.

The main areas of concern are:

1. model and simulation (input) data,
2. management and storage of captured data from C4I systems and simulations,
3. human logged data (questionnaires, interviews, observations),
4. scenarios,
5. analysis results,
6. particular high-classification cells or equipments within an experiment,
7. coalition data sharing and connectivity issues, and
8. secure networks.

As a general rule, the lower the security classification of these elements, the lower the cost and risk of the experiment and thus experiments should be run at the lowest classification level possible. This is not to say, of course, that undue efforts should be made to make everything unclassified or artificially low in classification. As previously discussed, all experiments are compromises, and the experimenter needs to decide where the benefits of (for example) higher classification or fidelity representations of equipments and scenarios outweigh the benefits of using lower classification (and hence cost/risk) analogues. The point is that this should be a conscious decision based on benefit and cost.

13.4 Ethics in Experimentation

Any experiment which involves human subjects and human data collectors could potentially pose both ethical and safety issues. By recruiting subjects to undertake an experiment, or by exposing the data collector to a potentially hazardous military environment the experimenter is expecting them to operate outside their normal working practices. Should an accident befall any of the subjects or data collectors, and it can be shown that this was due to the negligence of the experimenter then they, their employers and, potentially the government could be liable to compensation claims, or in extreme cases to prosecution. In this section the necessity and mechanics for gaining ethical clearance for experimentation, and the health and safety requirements

are reviewed. It is acknowledged that detailed procedures differ between TTCP nations but the advice provided here is as generic as possible.

13.4.1 The Importance of Ethics in Experimentation

In recent years there have been increasing efforts both nationally and internationally to ensure that research involving human subjects, and increasingly data collectors, meets acceptable ethical standards. The issue of human experimentation, safety and ethics has become an increasingly important consideration in all countries over the past 20 years, in part due to the increasingly litigious tendencies of western society. In this section it is intended to review the general ethical standards and safeguards that should be employed to ensure subject and experimenter safety; the concept of formal ethical scrutiny, or approval; and the decision points that must be made in determining when such ethical approval should be obtained. Ethics is a complex field, but in professional contexts its fundamental concerns are to:

1. respect the autonomy of individuals,
2. avoid causing harm,
3. treat people fairly,
4. act with integrity, and
5. use resources as beneficially as possible.

In addition, the research or experimentation must have true scientific value and worth. The research must be carried out with integrity with researchers demonstrating that they are genuinely striving to achieve the objectives of sound research by ensuring valid methodology, objective research processes and well-grounded findings. Research which lacks integrity is generally considered to be ethically unacceptable as it not only misrepresents what it claims to be but also misuses resources.

13.4.2 General Ethical Principles for Research and Experimentation

The set of principles[40] to be addressed includes:

1. **Respect for human dignity**: protecting the multifaceted interests of the participant, *i.e.*, the bodily, psychological and cultural interests;
2. **Respect for free and informed consent**: recognizing that potential participants must be given full and complete information before they can give their consent;
3. **Respect for vulnerable persons**: providing care and special protection against abuse, exploitation, or discrimination;
4. **Respect for privacy and confidentiality**: protecting the access, control and dissemination of personal participant information to avoid anguish;
5. **Respect for justice and inclusiveness**;
6. **Being fair in reviewing research protocols**;

40 Modified from the Tri-service Policy Statement issued jointly by the Medical Research Council of Canada, the Natural Sciences and Engineering Research Council of Canada and the Social Sciences and Humanities Research Council of Canada (1998). Cited in [DRDC HREC 2002].

7. **Not unfairly burdening anyone as a research participant in respect to harm**;
8. **Not discriminating against any participant who could benefit from the research on the basis of gender, culture, *etc*.**;
9. **Not involving in research, participants from populations that will not benefit from it**;
10. **Balancing risks and benefits**: establishing that the foreseeable risks to the participant do not outweigh the anticipated benefits of the research;
11. **Minimizing risks**: employing methods for assuring participants that risk of harm is minimized, including the employment of the smallest number of participants to achieve scientific validity, minimal time involvement, and acceptable experimental design and data analysis to achieve scientific validity;
12. **Maximizing benefits**: establishing that the research benefits the participants, the relevant Armed Forces or government, and society as a whole.

13.4.3 Required Ethical Standards for Research

As a result of these general principles, it is possible to derive some general ethical standards [DRDC HREC 2002], which should be followed when there is human participation in a research project or experiment.

1. The research project must contribute significantly to an approved program and have reasonable prospects of yielding important results.
2. The results obtained from the research project are not known to be obtainable by other means of study than through human participation.
3. The number of human participants used to achieve the required results will be kept as small as reasonable to reduce the risk of harm.
4. The research project will be conducted so as to avoid all unnecessary physical and/or mental discomfort, suffering or injury, or the misuse of the participant's time.
5. No experiment will be initiated if, in the light of current knowledge, there is any reason to believe that death or a disabling injury is likely to occur. Sufficient knowledge must be acquired from reliable sources to give assurance of reasonable safety prior to any consideration of human participation; such sources include animal models, laboratory experiments by others and information from the scientific literature.
6. The degree of risk that is to be taken should never exceed that which is commensurate with the urgency or importance of the program to which the study is related, and shall be reduced as small as practical.
7. Proper preparations must be made, and adequate facilities provided, to protect the human participant against all foreseeable possibilities of injury, disability or death.
8. Only persons having the requisite scientific, technical and/or medical qualifications shall conduct the research project. All persons who participate in the study will apply the highest degrees of skill and care during all stages of the study.
9. Reasonable steps shall be taken to ensure that none of the participants has any physical or mental conditions or previous exposures, which may make participation more hazardous than it would be for a normal healthy person, unless such a condition is a prerequisite for the particular study.

10. The Principal Investigator, each member of the investigative team and the Medical Officer shall be prepared to terminate the participant's involvement at any stage if anyone has reason to believe that continuation is likely to result in injury, disability or death.
11. There shall be no greater intrusion into the privacy of the participant than is absolutely necessary for the conduct of the study.
12. Normally, no experiment will be initiated if there is any reason to believe that such experiments have been done previously and that no important new knowledge is likely to be obtained. However, replications may be made in exceptional circumstances, *e.g.*, repeating previous work where the purpose is to define boundary conditions for future extensions of a research project, or where the results of previous research are required but are not available to the sponsoring organization, *e.g.*, because of corporate restrictions on data dissemination.

13.4.4 Ethical Clearance and the Ethics Committee

A common, and in some cases nationally mandated, means of ensuring that the above standards are met is through a formal process of ethical clearance by an accredited ethics committee. Experimental institutions or sponsoring authorities (*e.g.*, governments or defense departments) will usually have their own independent ethics committees. Such committees normally accept research or experimental protocols in a standard form for their review. An experienced academic will generally chair the committee who will be supported by permanent or co-opted experts that increasingly will also include legal experts, medical doctors, military representatives and union representatives. Each committee member will normally have clearly defined terms of reference. Such committees meet on a regular basis and at these meetings will review all experimental protocols and will often take reports from experimenters on completed studies. These reports act as an audit trail for previously approved studies and provide a means of reporting any incidents that arose during the experiment (personnel withdrawals, injuries incurred, expected compensation claims, *etc.*).

13.4.5 The Requirement for Ethical Clearance for Human Experimentation

There are no definitive guidelines particularly that are common to all TTCP nations, which clearly state when formal ethical clearance by an ethics committee would be required for defense experimentation. Each experiment must be judged on its own merits and a decision made during the planning phases. There are, however, a number of considerations that can help the experimenter in deciding if clearance is necessary. It is generally acknowledged that decisions relating to ethics are complex and it is impossible to provide generic guidance that will cover all possible issues. The decision whether a particular research proposal requires scrutiny and clearance by the ethics committee normally rests with the line management who will be held accountable. An important consideration is the need (or otherwise) for whatever protection will be put in place as a result of ethical approval being given. For example, in the UK, the MoD has a no-fault compensation scheme whereby an experiment participant can seek compensation for injury or other harm. Although a claim against the scheme would not prevent individuals making a claim for negligence in a court of law, it does allow compensation to be paid without the need to go to court as a necessary first step. This offers the benefit of a simpler procedure for the claimant and provides the experimenter

with expert support in a potentially very difficult area. No-fault compensation arrangements only apply if a MoD ethics committee has approved the experiment and the full range of approved procedures are followed.

When reviewing the acceptability of research on human participants, experimenters and line management, in the order shown below, should take the following steps:

1. Make comparisons with any parallel work and identify any principles of good practice that might apply.
2. Establish whether there is an existing generic protocol within which the work can be conducted. If so no further approval is required providing the work is carried out within the framework of the protocol. However, ethics committees would normally wish to be aware of all work carried out under such protocols.
3. Consider whether participation in the research would be significantly more stressful and/or hazardous than the normal duties of the subject *or* could be regarded as "intrusive." Here intrusive is defined as: "That which involves interference with the subject (psychological intrusion, including intrusion on privacy, or physical invasion)." If the work would fall into either of these categories, then it should be referred to an ethics committee for formal review.
4. Consider whether the nature of the work or any experimental mishap could potentially raise public concern or lead to legal action. If so then the work should be referred to an ethics committee for formal review.

Ethical considerations do not cease once the experiment has ended. Ethics committees now examine immediate post-study treatment and the long term monitoring of subjects who have been involved in certain types of experimentation. An example of the increasing importance of post-experimental treatment of subjects comes from the United States. A university ethics committee rejected a study of the effects of hypnotics on sleep because suitable arrangements for taking subjects home after the experiment had not been put in place. The experimenter had not assessed the risk that was associated with the subjects driving themselves home when there was a possibility that the circulating levels of the hypnotic could impair cognitive and physical performance. In addition, ethics committees may take an active role in ensuring the safe keeping of subject records as required by data protection legislations, where they exist.

13.4.6 Considerations for Defense Experiments

It is strongly recommended that the "required ethical standards for research" outlined in Section 13.4.3 be followed for all defense experiments. They are good practice regardless of the legal situation in any particular nation (which may mandate them to varying degrees). However, when considering the need for formal ethical clearance, there are some aspects of human-in-the-loop defense experimentation that set it apart from more conventional human sciences experimentation.

First, defense experiments often have more in common with training exercises or OT&E (as described earlier in this document) than they do with traditional human experimentation. The latter tends to deal with the investigation of human responses, both physiological and psychological, to various stimuli. OT&E, on the other hand, tends to deal with testing equipment, including its human operators, in realistic operational

conditions. Thus in OT&E, the humans are an essential part of the activity but are not the central focus. Perhaps for that reason, OT&E does not generally require ethical clearance in most nations. Defense experiments fall somewhere in between the two; the emphasis is on the socio-technical system under investigation, in which the humans play a more central role than in OT&E, but less so than in classical human experimentation. This is a further indication that ethical clearance for defense experimentation is something of a grey area.

Second, because defense experiments usually require whole units or HQs, the participants are *de facto* not truly volunteers; they are supplied by their chain of command as part of their normal service duties. In traditional human sciences experimentation, it is normally considered unethical to exploit non-volunteer subjects, but this is one of those areas where defense experimentation using human-in-the-loop simulation has more similarity to training and OT&E. However, if it is decided (through whatever process is operating in the relevant country) that there is a need for ethical scrutiny, then almost by definition, the participants must be volunteers. They should consequently be asked to consent in writing and it must be quite acceptable to the experimenters if they decline, once the details of the experiment have been explained to them. Clearly then, there is a very strong incentive to design defense experiments that involve minimal risk, certainly no greater than would be experienced in normal training.

13.4.7 Multinational Experimentation

If formal ethical scrutiny is considered to be appropriate, the situation is further complicated if the experiment in question is multinational. Individual nations' no-fault compensation schemes normally apply only to their own citizens. When these citizens are participating in an experiment in another country, especially when under the direction of that country's military commanders or experiment director, the situation regarding no-fault compensation is unclear. If one or more collaborating nations have taken different views on whether formal ethical scrutiny is required for the event, then the position is even more ambiguous.

Unfortunately no clear advice can be provided on this issue at this stage, but it could potentially be a very significant issue for future coalition experimentation, especially if at least one participating nation has decided to apply their own formal ethical scrutiny procedures to the activity, but not all have.

13.5 The Importance of Health and Safety in Experimentation

There is an obligation for defense experiments to comply with relevant national Health & Safety legislation and to provide working conditions that would ensure, as far as reasonably practicable, a healthy and safe working environment. Health & Safety criteria apply to research in addition to ethical criteria. Again, individual nations have their own sets of procedures and laws. However, in general, the following should be addressed.

The experimenter[41] must consult the subjects or their representative[42] on matters relating to health and safety at work, including:

1. any change which may substantially affect health and safety at work, *e.g.*, in procedures, equipment or ways of working;
2. arrangements for getting competent people to help participants/subjects and experimenters satisfy health and safety laws;
3. the information given on the likely risks and dangers arising from the experiment, measures to reduce or get rid of these risks and how to deal with a risk or danger;
4. the planning of health and safety; and
5. the health and safety consequences of introducing new technology.

In general, the experimenter's duty of care to the subjects includes:

1. making the workplace (experimental venue) safe and without risks to health;
2. ensuring plant and machinery are safe and that safe systems of work are set and followed;
3. providing adequate welfare facilities;
4. giving information, instruction, training and supervision necessary for health and safety.

In particular, the experimenter must:

1. assess the risks to subjects' health and safety;
2. make arrangements for implementing the health and safety measures identified as being necessary by the assessment;
3. record the significant findings of the risk assessment and the arrangements for health and safety measures;
4. draw up a health and safety policy statement, including the health and safety organization and arrangements in force;
5. appoint someone competent to assist with health and safety responsibilities, and consult subject or subjects' safety representatives about this appointment.

The experimenter must also:

1. set up emergency procedures;
2. provide adequate first-aid facilities;
3. make sure that the experimental venue satisfies health, safety and welfare requirements;
4. make sure that work equipment is suitable for its intended use, so far as health and safety is concerned, and that it is properly maintained and used;
5. ensure that appropriate safety signs are provided and maintained; and
6. report certain injuries, diseases and dangerous occurrences to the appropriate health and safety-enforcing authority.

[41] The wording of the UK Health and Safety at Work Act uses the term employer, however in the context of this guide the word experimenter is applicable.

[42] In this context representative could be the commanding officer, senior NCO or union representative depending upon the subject population.

However, the experiment subjects, normally also have health and safety (H & S) responsibilities:

1. taking reasonable care of own health and safety and that of others who may be affected by what the subject does or does not do;
2. co-operating on health and safety;
3. correctly using work items provided by the experimenter, including personal protective equipment, in accordance with training or instructions; and
4. not interfering with or misusing anything provided for health, safety or welfare.

This list is far from exhaustive but gives a clear impression of the health and safety responsibilities that both the experimenter and the subject must observe. Individual nations have their own requirements for H&S plans and risk assessments, which are normally mandated.

Principle 14.

Frequent communication with stakeholders is critical to successful experimentation

Principle 14 crowns GUIDEx by advocating that every integrated analysis and experimentation campaign needs a champion. Otherwise, whatsoever outstanding the results of a particular campaign or experiment might be, it may fail to have a real impact on operational systems and their future capabilities. Experimenters must identify the key stakeholders of the client organization, especially those who can implement recommendations from the experiment.

An essential element of any campaign is the maintenance of an effective and frequent dialogue with stakeholders. This permits a clear understanding of the question and issues to be addressed and maintains contact with any changes to priorities. While maintaining the integrity of the experiment, the experimenter should invite stakeholders to attend the experiment. This approach helps key stakeholders take ownership of the experiment and its products.

It is recommended that key stakeholders be personally briefed on the results from the experiment in addition to being sent a short but focused report.

Principle 14. Frequent communication with stakeholders is critical to successful experimentation

The reader is encouraged to apply and adapt this Principle. However, the examples herein are based on the specific perspective and experience of different lead-nation authors with contributions from other participants. Therefore they may require supplementary effort to relate them to national perspectives.

In Principle 14 we discuss the importance of engaging in continuous dialogue with stakeholders. While a single experiment is used as the example, the best practice recommended can equally be applied to other activities in an integrated analysis and experimentation campaign.

14.1 Introduction

Good communication is central to achieving a successful outcome; and yet it is possible to find an experiment, or integrated analysis and experimentation campaign, which does not have a rational plan for communicating with key stakeholders[43]. A communications plan must consider how the different stages in running an experiment may require different approaches to good communication; stages such as determining the right set of questions and issues to be addressed, maintaining the confidence of key stakeholders that the potential changes to their priorities are being considered, ensuring all stakeholders have appropriate access during the experiment and making sure that they understand the output from the experiment and the evidence that supports any subsequent exploitation. This guidance is drawn from the best practice of a number of nations.

14.1.1 Why a Communications Plan

The final product of any defense experiment must be the evidence that the right questions have been addressed and that the evidence required for its findings be exploited effectively. There are many examples where this aim has been met, and other examples where it has not been met, *i.e.*, where no connections between the initial question and the final product were found. An example of the latter was seen in a maritime program, which was examining methods of reducing fouling of ships' hulls. The program or campaign supplier (the experimenter in this example) did not address possible anti-fouling measures, but researched and subsequently reported on the life cycle of crustaceans. This was not revealed until the report was delivered, where it also transpired that they had not held regular meetings with the sponsor. Had a suitable

[43] Stakeholders are defined as persons who have a vested interest in the product from the experiment or program.

communications plan been put in place, this misunderstanding would not have occurred.

A communications plan, which need not be extensive or formally binding, should address the following points:

1. **A timetable of regular meetings with key stakeholders.** This should be no more than standard business practice but the plan should record a set of mutually agreed dates, times and venues for progress meetings. This should not prevent *ad hoc* meetings of opportunity to be held.
2. **An agreed timetable for briefing all stakeholders and visits to the experiment.** This should give dates, times and venues and identify those who are attending each brief and visit. Ideally this should also give a broad impression of the brief (*e.g.*, open forum) and briefing material that will be available on the day (*e.g.*, PowerPoint slides). Visitors should be provided with a set of the briefing material.
3. **An outline of the product from the experiment, and method of dissemination of the final report(s).** This should address both hot debriefs and final reports. This should also give an outline of the structure of the final report(s).
4. **Names and contact details of lead experimenters and all stakeholders.** This helps record the main users of the research and helps with subsequent exploitation of the product.

14.2 Determining the Right Set of Questions and Issues

A key prerequisite to an integrated analysis and experimentation campaign is the identification of the origins of the question(s) to be addressed and identification of key stakeholders. One difficulty is that the obvious stakeholder is often not the person that originally posed the question. Therefore an initial step must be to chase down the origins of the question, and from that define the key stakeholders who need to be influenced. However, the question may arise from many sources and it may not always be possible to directly engage or even identify the original source. For example the question may have arisen from a strategic plan which states that “there is a need to enhance interoperability with our allies to a level which will allow us to undertake concurrent medium scale operations.” This will reflect a political imperative, and whoever is responsible for the strategic plan may have appointed intermediaries whose task is to implement this directive. In this case, these are all key stakeholders, and it is essential to determine their relationships and how they work together. Intermediaries will have formed their own understanding of the question being posed and may have defined a campaign with which to implement their directive. This campaign will identify the suppliers, the timescales, other stakeholders and possible exploiters of the information and the finance available to undertake the campaign.

14.2.1 Early Engagement with Key Stakeholders

The initial meeting is important as it invariably sets the tone of the professional relationship for the rest of the campaign. The experimenter should expect a full brief on the work required, and the required outputs. The supplier may consider briefing on previous work that they have undertaken. There are a number of important questions

that must be answered at this meeting as they provide the framework for a common understanding of the problem space. These questions are shown in the list below, in what is by no means exhaustive:

1. **Who are the key stakeholders?** The identity of key stakeholders and their relationships should be explained with an appreciation of the roles that they fulfill within the organization.
2. **What was the original question, and in what context was it posed?** As described above the original question may have arisen elsewhere within the organization and been interpreted by intermediaries. It is important to understand the process of interpretation and to understand the context in which the original question was posed as this will help identify drivers and constraints.
3. **Do the key stakeholders know what they want, and are the outputs compatible?** If the question arose from a sufficiently high level, the output requirements may be quite different from those of the stakeholders that the experimenters meet. It is also important to understand the differences and how they interrelate. It may be that such questioning highlights differences in expectation that exist between key stakeholders. This question should also identify if there is a clearly defined strategy and understanding for briefing the results back up the chain of command.
4. **Is there a required format for the deliverable?** Have the key stakeholders considered how they wish the experimentation findings to be communicated? Is there a preferred distribution list for the report?
5. **Who are the other stakeholders?** Who else has a clear interest in the findings? For example the information arising from the interoperability question above may be used for setting policy, but also for doctrine purposes. A list of all stakeholders should be compiled and agreement reached for the campaign (or experiment) supplier to visit all stakeholders and brief them on the integrated analysis and experimentation campaign (or experimentation approach).
6. **Are there other users of the information?** It is worth identifying other potential recipients of the information. For example, could it be used as evidence to support formal procurement decision points?
7. **What is the exploitation route for the output?** Do the key stakeholders have a clear internal process for exploiting the product? How will it be exploited? Who will be the recipient? Is it appreciated that experimenter's briefings be a valid exploitation route?
8. **What is the time pressure?** When is the information required? Ideally a timetable for each stakeholder's requirements should be established.
9. **Are there alternative campaign (or experiment) designs that will answer the question?** Although stakeholders may favor a particular course of action, this is an opportunity to understand their thinking and propose alternative approaches.

The effort that is placed in gathering this background information will not only establish a good working rapport with key stakeholders, but will also help define the deliverables and the communications plan. The key stakeholders should now have ownership of the campaign and be its champions.

Following this meeting the final campaign plan can be prepared. A communications plan should be one component of this campaign plan. Best practice has shown that there should be a high level of stakeholder involvement in the derivation of the plan. Regular circulation of drafts of elements of the plan helps ensure ownership and reduce the chance for any ambiguities or misunderstandings. Circulations can be on an *ad hoc*

basis but the feedback can be discussed within the framework of the progress meetings.

14.3 Communications in the Run up to the Experiment

14.3.1 Progress Meetings

Although this will be a particularly busy period, it is essential that regular dialogue be maintained with the stakeholder community prior to the experiment. For example, it is common practice to hold regular progress meetings to which all stakeholders are invited. These meetings are excellent opportunities to ensure that the structure of the campaign continues to be appropriate for the question posed. For example, within the lifetime of the campaign, which could be in excess of a year, stakeholder priorities may change. By maintaining this regular dialogue, changes in priorities can be quickly identified and accommodated. The meetings should also be used to provide detailed briefs on progress, and possibly the most important brief will be on the experiment design.

14.3.2 Briefing the Experiment Design

Prior to the briefing, it is advisable to circulate a copy of the campaign plan, related documents[44] and the experiment design as this allows stakeholders to understand the objective and to help phrase questions and revisions. The briefing should be given to the key stakeholders and, if appropriate, the commanding officer of the troops who are taking part in the experiment. Potential attendees should be notified of the date and venue for the meeting as early as possible, and be afforded the chance to invite others who may have an interest in the problem. The presentation should be open-ended, thereby allowing the audience to participate in discussion and to understand and take ownership of revisions to the experimental design. Such briefings should empower the stakeholders in their respective roles. It is essential that minutes are kept of the meeting and that changes to the experimental design are recorded and are circulated to all stakeholders to ensure agreement.

14.4 Communications During the Experiment

In most cases, major interaction with stakeholders occurs during the visitor day. However, the benefit gained from visitor days does not always justify the effort expended. In general, the day is carefully stage-managed to ensure that each visitor gains an impression of the experiment design and its context within the campaign, views any hardware, has the opportunity to talk to the experimentation subjects and may be given an early appreciation of the data. However, the opportunity for detailed discussions with stakeholders is usually limited and they may leave the experiment

[44] For example, CONOPS, TTPs, proposed equipment lists, training plans.

without a full understanding of key issues, or an idea of how the product will finally be communicated.

Experience has shown that the ideal size for individual briefing meetings is about six. By having such a small audience it is possible to provide a more in-depth brief and the opportunity is afforded to "talk through" each product and raise awareness of the issues surrounding the campaign. The visitors should be carefully selected so that the experimenter is briefing common areas of interest. Visitors should be encouraged to view the entire experimentation process, and be invited to observe and interact with the subjects in a way that does not interfere with the experiment. This approach removes the artificiality of the stage-managed visitor day and gives a far clearer picture of the major issues surrounding the experiment. It also allows in-depth discussions in the margins of the experiment.

Additional attendance of stakeholders with a direct involvement in the campaign implementation, outside the specific visitor day, improves communication by providing more opportunities to brief them at regular intervals. This allows the experimenter to discuss issues with them as they arise, and provide an instant resolution.

14.4.1 Visitor Day Guidelines

Visitor days are important events and represent the best way to get across to stakeholders the key first impressions, the successes that have been achieved, and the views of the experiment's subjects. They will normally be at a time shortly before the end of the activity. Careful planning of the following issues is advised:

1. the layout of the experiment set-up;
2. the potential intrusion into the experiment activity itself;
3. the devising of an interesting and persuasive program;
4. the potential to undermine the final results to be presented later; and
5. the opportunities for visitor discussions with the experiment's subjects.

14.4.1.1 Layout

A typical C2 experiment set-up can often look to the uninitiated like a confusing plethora of networked computers, many showing mapping data of various sorts; some showing ground truth; and some showing subject perceptions, *e.g.*, a common operational picture. There is often a great deal of new material for the visitors to take in. Simple actions such as putting up posters in the vicinity of particular cells, thus indicating what they do, and labeling screens, perhaps using a simple color code to distinguish main types of display, can go a long way to helping the visitors appreciate what is being demonstrated to them.

14.4.1.2 Intrusion

The potential for intrusion can take many forms. The obvious one is the impact of senior officers simply looking over the shoulders of the subjects when they are attempting to carry out their duties. This can have a profound effect on their behavior!

Second, the presence of a large number of people streaming through an experiment area (an HQ for example) can also cause considerable disruption to the experiment play. It is disingenuous first to ask the military subjects to take the experiment as seriously as they would a training exercise, and then to disrupt them. The ideal solution here, if the time budget for the experiment will allow, is to put aside a period for the visitor day after all the key experiment work has been completed. By this time, the subjects will be at the zenith of their training level, any bugs in the set-up should have been well and truly ironed out, and there is no chance of compromising the main experiment.

A live video feed to a separate room has been found to be a useful means of briefing visitors in an immersive fashion without them having to spend too much time looking over the shoulders of either subjects or EXCON staff. Thus this can be an entirely non-intrusive means of viewing experimental activity.

14.4.1.3 Visitor Day Program

A one or two hour presentation on the whole campaign, followed by a quick look at the experiment set-up, does not constitute an interesting or persuasive visitor day. The visitors should go away with a clear idea of what went well (and what didn't, and why); the first impressions of the analytical team and some views from the subjects. Overall they need to be confident that the experiment will, when properly analyzed, achieve what it set out to. A tried and tested visitor program structure included:

1. Provide a brief presentation on the campaign as a whole and the nature of the specific experiment. Just enough to put the day in context.
2. In particular, avoid having your visitors seated, listening to long briefings in the afternoon. Sending them to sleep is not good practice!
3. Demonstrate some of the experiment equipment "off line," with a hands-on session if possible (with a helpful guide, of course).
4. Try to avoid explicitly demonstrating potentially tedious parts of the experiment activity, *e.g.*, "now we'll see how in the baseline case it takes ten minutes for the data to pass through the network."
5. Walk-through of the experiment set-up, preferably when something exciting is going on (though note the "intrusion" paragraph).
6. Discuss with the subjects.
7. Present on first impressions and general discussion.

14.4.2 Potential to Undermine Subsequent Results

There is usually considerable pressure to provide some sort of "first impressions brief" at visitor days. This should not be resisted simply because the analysis obviously hasn't happened yet, but it is important to keep the scope genuinely to highly caveated first impressions. The temptation to announce some startling finding (especially one that the visitor will be known to like) should be resisted at all costs, because it is quite likely that when the analysis is complete, that finding will at best need to be modified, and at

worst, changed altogether. All that can be said with reasonable certainty on a visitor day in the latter part of any experiment is the extent to which:

1. The experimental equipment worked.
2. The subjects were engaged and were able to use the experimental equipment.
3. The experiment ran to plan.
4. The requisite data were collected.

If the answer to these questions is mostly "yes", then that in itself is a convincing indication of an experiment well run.

14.4.3 Talking with the Experiment's Subjects

Allowing the visitors to discuss the experiment directly with the military subjects can yield many benefits, but also has some pitfalls. In no case should the visitor be allowed to interact with the military subjects prior to the military subjects reporting their opinions on questionnaires used in data analysis. When interaction is not intrusive, attempting overtly to influence what the subjects will say could be very counterproductive. The best advice here is to stay alert throughout the experiment to get an idea what the subjects would say to a visitor and then make a late decision on whether a free walkabout should be a part of the visitor day program.

Whatever approach is taken, it remains essential that stakeholders are exposed to the experiment and be given the opportunity to gain an early understanding of issues and possible results.

14.5 Dissemination of the Results

In a perfect world the key stakeholders would wish to receive a one-paragraph, or one-page report, which contains the findings. On the other hand, the wider stakeholder community would prefer a short report that outlines the study and contains the findings. Finally, the experimenter would like to produce a comprehensive report! The trick in the dissemination of the results is to accommodate all these requirements.

A well-written report will contain a one-page abstract, an executive summary and a full report. The traditional approach to dissemination of results has been to produce a paper that is sent to key stakeholders, with or without a presentation. While this has obvious merits, the general experience is that the approach produces "shelf-ware."[45] It should be remembered that these are busy people who will wish to gain quick appreciation of the key issues and findings, in order to exploit the information.

Prior to writing the report, it is advisable to discuss the structure and communication of the product with the key stakeholders. Serious consideration should be given to providing a hot de-brief as soon after the completion of the experiment as is feasibly possible. This should concentrate on key issues and be supported by a document of no

[45] A term that means the report is produced but does not have any influence on the decisionmaker.

more than 6 pages that is handed to all attendees. The de-brief audience should be the key stakeholders, line managers from the parent establishment and the main experimenters. Where possible it should be held at a key stakeholder's offices.

The meeting, which should be chaired by a key stakeholder, should take the form of an open forum at which audience participation is encouraged. A possible agenda for the presentation could be:

1. short overview of the experiment's question,
2. design,
3. implementation,
4. results,
5. conclusions, and
6. possible exploitation path.

The meeting should be minuted and the notes used to help formulate the final product.

Although a final product based on a slide presentation has been used widely in many countries, as it has the attraction of providing information in a digestible form, it is considered inadequate for the presentation of detailed data. In addition, the experiment can only be described in fairly superficial terms and, without strict configuration control, the slides can be used out of context.

Better practice is for a short report supported by a CD-ROM, which contains releasable data. The short report should address an outline of the methodology with the key findings related to the original question being carefully described and the estimated impact of the proposed change. The CD-ROM should contain the more detailed report with both the raw and processed data. There is also a trend to provide video recordings of experiments to further amplify the report. This is an extremely powerful communications tool.

14.5.1 Stakeholder Feedback

Many institutions issue customer satisfaction forms which elicit stakeholders' views on the final product. Although these forms provide some guidance on stakeholder satisfaction, they are occasionally not completed and returned or have the barest of comments.

A far better approach is to continue the dialogue with the key stakeholders to determine how the work has been received and, more importantly, how it is being exploited. This is clearly essential when the particular experiment is one activity in a wider integrated analysis and experimentation campaign.

14.5.2 Key Stakeholders and the Question

In the opening paragraphs it was identified that the key stakeholders may not be the source of the original question. An important post-experiment activity is to be able to help the key stakeholders produce their brief for the source of the original question.

While it would probably be unethical to write the brief, the experimenter should offer to assist in identifying key findings and help in the interpretation of the findings. He should also volunteer, if appropriate, to attend any briefing or provide a brief in person.

Part III
Introduction to, Exemplar Findings from, and Précis on GUIDEx Case Studies

Introduction to and Findings from the GUIDEx Case Studies

Too often, defense acquisition programs have not delivered the promised force capabilities. However, appropriate experimentation has demonstrated its value in supporting the transition from concepts to capabilities that truly meet national (and sometimes international) objectives. The following high-level results from GUIDEx Case Studies (CSs) reveal some of the reasons, substance and importance to operations of the findings unveiled and supported by valid experimentation.

These Case Studies expand on the techniques addressed previously in GUIDEx. Each Case Study includes a discussion related to GUIDEx 14 Principles, the 4 Requirements and 21 Threats to experiment validity. Relationships between the Case Studies and the GUIDEx Principles are summarized in table format in the Case Study summaries. To help the reader to appreciate the depth and breadth of the Case Studies, a summary of each is given below.

1. *Testing Causal Hypotheses on Effective Warfighting*: During a Persian Gulf air/sea scenario, in the common operational picture (COP) experimental treatment condition, all parties—higher echelon and lower echelon—had both the national intelligence supported big picture and the local tactical picture. This combination was experimentally proven to be superior technology for such operations, resulting in greater shared situation awareness and better bottom line combat effectiveness (see conclusion validity and details in GUIDEx CS1).

   In CS1 series of experiments on the COP prototype ('90-'91), the scenario was modeled on "Operation Praying Mantis," which occurred in the Persian Gulf during 1988. In it, US ships and protected oil platforms were under assault by Iranian "Boghammers" (small fast attack craft) and other ships. The perpetrators had to be identified and retaliated against. US Defense Secretary Frank Carlucci from the Pentagon optimally terminated one operation, since he was provided a near real-time tactical picture of the operation by the transmission of the JOTS picture over a wide area network from the ship in the Gulf to the Pentagon NMCC. The COP prototype technology tested in the experiment was an expansion of this JOTS technology additionally providing for the simultaneous transmission of a national intelligence based "big picture" view from the Pentagon down to the ships operating in the Gulf. Thus, in the COP experimental treatment condition, all parties, higher echelon and lower echelon, had both the big picture and the local tactical picture. This was experimentally proved to be superior technology for such operations.

   While assuming that a valid causal interpretation of any correlation between two sets of observations depends upon the presence of a compatible causal hypothesis and the absence of a plausible rival hypothesis to explain the correlation on other grounds, this study employs Yule's Covariance Theorem to prove that a controlled experiment provides an unequivocal test of a causal hypothesis. It then recounts six cases of controlled experiments using human-in-the-loop simulation, including three replications of the original GCCS COP prototype experiments with experienced military officers engaged in realistic crisis scenarios, which demonstrated significant improvements in combat effectiveness. These experiments are, therefore, consistent with P1, P2, and P3, and the causal hypothesis is strongly supported by the experimental evidence. Since there was a three-year hiatus between the completion of this series of experiments and the onset of official technology adoption and engineering, the project would have benefited from earlier and more effective communication with the decisionmaker (P14).

2. *UK Battlegroup Level UAV Effectiveness*: This experiment supported a major UK UAV acquisition program in demonstrating the huge information gathering potential of UAVs at the tactical level,

compared to existing ISTAR assets. However, equally important, it showed that if integration into the supported HQs is not achieved effectively, then the resulting information overload can have a hugely detrimental effect on mission success.

This CS illustrates how one can both make the most out of scarce resources and maintain internal validity by piggybacking experimentation activities onto collective training exercises using properly tailored design (P9). In this particular instance, the six-battalion weeks invested for training purposes provided sufficient observable events to deliver strong statistical power to the causal experimental hypothesis. This CS also shows how simple M&S can be used in conjunction with live action to achieve some of the benefits of both experiments using human-in-the-loop simulation and field experiments.

3. *UK NITEworks ISTAR Experiment*: The UK, like other nations, is presently investing heavily in ISTAR sensors and systems. However, it is widely recognized that effective information requirements management (IRM) is vital to the efficient use of those systems. This experiment investigated both technological and procedural means of improving IRM. It showed conclusively that a collaborative working environment with appropriate working practices would have a major beneficial effect on IRM effectiveness. This assisted the development of ISTAR management priorities in the UK.

   This was a classic experiment design of a defense experiment, resulting in fruitful application of Principles 2 & 3. It would have had more external validity (P3) by the addition of the M-E-M paradigm (P7) that would have added a workflow model that portrayed RFI timeline flows in typical operations to compare with the non-collaborative experiment. The fact that number of RFIs in the scenario was based on historical information is a rudimentary pre-experiment model that allowed for some generalization of experiment conclusions. In this Case Study a possible avenue to further increase external validity would be to conduct a follow-on experiment in a different venue (P7): a field exercise large enough for collaboration in a larger HQ.

4. *Pacific Littoral ISR UAV Experiment (PLIX)*: This CS provides insights difficult to capture without experimentation; the strong hypothesis of identifying and tracking **all targets** proved not to be attainable even though sensor coverage was nominally complete, pointing to integration requirements for an effective ISR architecture.

   The PLIX Case Study is a particularly good example of the importance of P4 and P5, in particular the use of an iterative campaign mitigating the outcomes of a single experiment. For example, in this case, the conditions under which the hypothesis could hold were increasingly better understood. The outcomes have been used to develop a subsequent experiment in the campaign and useful insights were acquired and contributed to the understanding of the contribution of the UAV to the Recognized Maritime Picture (RMP) and its requirements for a multi-node ISR architecture. The individual experiment could have been improved by more attention to P3, to use the capability, and better understand the relationship between the independent and intervening variables, MoEs and their assessments. Contrary to P7, the exclusive reliance on live experiments may have limited return on investment since basic integration issues, dependencies and other requirements could have been identified in a controlled environment.

5. *An Integrated Analysis and Experimentation Campaign: Army 21 / Restructuring the Army 1995-99*: This campaign demonstrated the importance of detailed problem definition and an iterative approach based on wargaming, field trials and analytical studies. The warfighting concept under test (A21) was found to fail under realistic environmental constraints. However, the results led to an alternative concept, which is the basis for current Australian Army force development.

   This CS showed the advantage of early communication with the customers to develop a commonly agreed and understood definition of the problem. Subsequently substantial modeling activities were pursued to better address the experimentation challenges at stake. The critical objectives of the campaign, initially identified as hypotheses, served their purpose well.

6. *The Peregrine Series: A Campaign Approach to Doctrine and TTP Development*: This on-going campaign of experiments and studies is directly contributing to the development of the doctrine for employment of the Australian Army's new Armed Reconnaissance Helicopters and demonstrates how experimentation can be used to inform capability development questions at unit level and below.

   This CS provides illustration of the advantages of a campaign (P4-6) over a single experiment or a short series of events, and demonstrates how a less controlled, exploratory experiment can be used with a number of more focused events to build validity (P7). However, the Case Study also demonstrates one of the difficulties in developing a campaign plan, with the problem of gaining clear direction, guidance and commitment from the stakeholder in advance (P14), preventing the development of a long-term detailed plan. In order to overcome this difficulty, a number of exploratory events were required to assist with problem definition, but at the expense (in terms of time and resources) of more focused activities (P5).

7. *Multinational Experiment Three (MNE 3)*: Despite the complexity of the MNE 3 effects-based planning (EBP) experiment and the findings that the concept and supporting tools require further development, the event demonstrated the potential for EBP to make a coalition task force a more effective instrument of power. It also showed the benefits for collaboration to produce the best ideas from a collective thought process in a coalition which included a civilian interagency component.

   This CS offered strong external validity (P2-3 r4[46]) by its use of an operational scenario and database, and its use of operational personnel from various nations. However this CS also demonstrated how emphasis on external validity makes it difficult to achieve internal validity. MNE 3 demonstrated problems meeting (P2-3 r1) having all the users understanding and using the process from Day One, and problems in meeting (P2-3 r3), the ability to determine the reason of observed improvements. This demonstrates that in designing an experiment one needs to find balance among the four requirements of experiment validity, especially internal and external validity.

8. *Improved Instruments Increase Campaign Values*: While improved experimentation instruments provided the opportunity to generalize some results, they also increased the validity of campaign's results and knowledge generation synthesized for future information management systems as illustrated by the MONIME[47] campaign.

   This CS exploited all the methods of knowledge generation of a campaign as described in GUIDEx that were made available to the international collaboration. Success of the MONIME campaign was due to proper problem definition (P4); an iterative process to reach an agreement between analysts and management (P5); integration of the three scientific methods of knowledge discovery and synthesis (P6); exploitation of all the methods available from national resources supported by adequate experiment design to increase analysis robustness (P1-3, 7); techniques to counter human variability (P8); special considerations in exploiting collective training (P9), adequate exploitation of M&S (P10), impressive (exhaustive) data analysis and collection plans (P12); and most importantly a continuous (P14) review of progress with the customer.

[46] The notation r with a number refers to one of the four GUIDEx requirements for valid experiments, *e.g.*, r4 for generalizability or strong external validity.

[47] Designed by AUSCANNZUKUS management of organic and non-organic information in a maritime environment (MONIME) ad-hoc working group.

The following table is an indicator of the coverage of the different analytical methods employed by the Case Studies.

	Defense Experiments using:				Real Ops
	Constructive Simulations	Analytic Wargames	HITL Simulation	Live Simulation	
CS1: Testing Causal Hypotheses (US)	✓		✓ (x6)		Data from Operation Praying Mantis
CS2: BG UAV Experiment (UK)			✓	✓	
CS3: NITEworks ISTAR Experiment (UK)			✓		
CS4: Pacific Littoral ISR (CA)				✓	Real ops incident encountered during field experiment
CS5: Restructuring the Army (AU)	✓	✓		✓	✓ Review of historical ops
CS6: Armed Reconnaissance Helicopter (AU)	✓	✓	✓		
CS7: MNE 3 (Multinational)		✓ Wargaming in a general sense, but not with a computer wargame	✓ M&S federation[48] was running in the background		
CS8: Instruments (Multinational)	✓	✓	✓	✓	✓ AUSCANNZUKUS issues

Table 5 Environments or venues exploited for the Case Studies (ticked when used)

[48] Federation of 3 simulations running: JSAF, JOANA (Germany), and ALLIANCE (France).

Case Study 1. Testing Causal Hypotheses on Effective Warfighting

1.1 Overview

Warfare, an ancient and complex phenomenon, allows political units to impose their will on others. Through the years a bevy of hypotheses have been advanced regarding the factors making for effective warfighting. More troops, more firepower, superior doctrine, better training and superior C4I are all capabilities hypothesized to make for more effective warfighting. Warfighting, itself, is adjudged more effective when enemy combat losses appreciably exceed own force losses. How does one go about testing the many hypotheses on the causes of warfighting effectiveness against observational evidence? One can prove that controlled experiments provide unequivocal tests of such causal hypotheses, and further argue that they provide the only feasible conclusive test; otherwise the observed results are open to rival explanation in terms of causation by some of the uncontrolled factors. We first examine hypothesis testing with observations on a single group and then move to the method of using simple correlational data for two groups. This requires us to confront the open ended issue of controls and control variables in testing causal hypotheses which in turn leads us to consider the most conclusive testing method, controlled experimentation. We then demonstrate the feasibility and utility of this method by providing examples of substantial results from six controlled experiments on the causes of warfighting effectiveness: two on the effects of alleged superior doctrine, *viz*. use of contingency planning, and four on the effects of alleged superior C4I, *viz.* use of the common operational picture (COP) and use of a prototyped planning aid. The experimental method advocated here for effectiveness testing of proposed defense capabilities and technologies is similar to the procedure of randomized clinical trials employed in the health sciences to determine whether or not the use of potential new healing drugs causes improved health.

1.1.1 Observations on a Single Group

For centuries contingency planning has been recommended as a superior military practice. Does use of contingency planning by a battlefield commander, in fact, cause improved warfighting effectiveness? We could observe the use of contingency planning (x) in the operational setting of a battle or military exercise and see if the Blue commander destroyed more Red platforms than he lost (y). But even were this to occur, to claim that x caused y remains open to the obvious rival explanation that some other variable (c) or combination of variables occurring simultaneously with the battle or exercise, *e.g.*, use of advanced weaponry, was the true cause of Blue's success. To draw such an inference from the observations is to be guilty of the classic *post hoc ergo propter hoc* fallacy[49]. Only if we could isolate the cause would such a procedure provide

[49] The *post hoc ergo propter hoc* (after this therefore because of this) fallacy is based upon the mistaken notion that simply because one thing happens after another, the first event was a cause of the second event. Post hoc reasoning is the basis for many superstitions and erroneous beliefs. http://skepdic.com/posthoc.html

a convincing test of our hypothesis. As an improvement on the above, we could intervene in an exercise to delay the onset of contingency planning, measure y, then insert x and remeasure y at the end of the exercise. But even here, any improvement in y could be criticized as having arisen not from x but from other new events surrounding x that occurred simultaneously with x during the later phase of the exercise, *e.g.*, Blue troops have learned more about Red causing them to perform better in the second phase. So while always instructive and often productive of valuable insights and new hypotheses, neither single shot case studies nor before and after measures on a single group provides a definitive test of a causal hypothesis [Shadish *et al.* 2002]. To rid ourselves of such rival explanations for our findings, we need a comparison group to ascertain what would have happened if contingency planning were not used, *i.e.*, we must deal with the counterfactual conditional.

1.1.2 Simple Correlation with Two Groups

As a test of our causal hypothesis, we could compare the combat wins (y) of a group of commanders who employed contingency planning (x) with that of another group of commanders who did not (~x). Passive observational evidence for and against the contingency planning/warfighting effectiveness hypothesis could be gleaned from the history of past battles or military exercises. We can array notional findings of research into two dozen such battles in a fourfold table as shown in Figure 49 below and calculate the correlation between the variables, ϕ_{xy}. Here x is the independent variable, usually a defense capability; and y is the dependent variable, usually a warfighting effectiveness measure.

	~x	x	Σ
y	4 (.17)	8 (.33)	12 (.50)
~y	8 (.33)	4 (.17)	12 (.50)
Σ	12 (.50)	12 (.50)	24 (1.00)

Figure 49 Use of contingency planning (x) by combat outcome (y)

Equation (1) is the simple *Phi* coefficient of correlation[50]

$$\phi_{xy} = \frac{P_{xy} - P_x P_y}{\sqrt{P_x Q_x P_y Q_y}}$$

which results in the following for our example: $\phi_{xy} = \frac{.33 - .50 \cdot .50}{\sqrt{.50 \cdot .50 \cdot .50 \cdot .50}} = .32$

ϕ_{xy} is simply a measure of association or correlation between two dichotomous variables where the numerator is the difference (.08) between the empirically observed association of x and y (.33) and what would logically be expected for their joint occurrence assuming statistical independence of x and y, where half the battles involved Blue use of contingency planning and half were Blue wins, (.50 X .50 = .25). As shown in Equation (1), this degree of association is assessed relative to the denominator which measures the total variability in x and y: ($\sqrt{.50 \cdot .50 \cdot .50 \cdot .50} = .25$). Random association would yield ϕ_{xy} = 0. Here we have a simple measure of David Hume's covariation of observations. Clearly, we observe a tendency for use of contingency planning to be disproportionately associated with, *i.e.*, correlated with, successful combat outcomes (ϕ_{xy} = .32). Indeed, two thirds of the battles surveyed were either contingency planning wins or non-contingency planning losses. So we find that use of contingency planning covaries with, *i.e.*, is correlated with, combat success, and that we cannot reject our causal hypothesis with these data.

1.1.3 The Role of Controls in Ruling Out Rival Explanations of Observations

Suppose, however, someone advances a rival hypothesis to explain our findings, suggesting that it is not the use of contingency planning, *per se*, but superior military training that caused the successful combat outcomes. After investigating the military background of the 24 Blue-battle commanders, he is faced with the subdivided array of data shown in Figure 50.

[50] The direct analogy to the ϕ_{xy} coefficient for correlation of interval scale variables is the Pearson product moment correlation coefficient: $r_{xy} = \sum xy / N\sigma_x\sigma_y$. See [McNemar 1962].

	c				~c		
	~x	x	Σ		~x	x	Σ
y	0	8	8	y	4	0	4
~y	4	0	4	~ y	4	4	8
Σ	4	8	12	Σ	8	4	12

Figure 50 Use of contingency planning by combat outcome controlling for training (c)

Equation (2) covariance/partial correlation theorem[51]

$$\phi_{xy} = \phi_{xyc} P_c \sqrt{\frac{P_{xc} Q_{xc} P_{yc} Q_{yc}}{P_x Q_x P_y Q_y}} + \phi_{xy\sim c} P_{\sim c} \sqrt{\frac{P_{x\sim c} Q_{x\sim c} P_{y\sim c} Q_{y\sim c}}{P_x Q_x P_y Q_y}} + \phi_{yc}\phi_{xc}$$

Although we cannot completely reject our contingency planning/warfighting effectiveness hypothesis with the new correlations, we find an equally plausible rival explanation for the findings, *viz*. training at U. S. Military Academy at West Point (c) leads to improved combat effectiveness (ϕ_{cy} = .32), and West Point training is disproportionately associated with use of contingency planning (ϕ_{cx} = .32). So, based on the simple correlations alone one could assert with equal confidence that West Point training caused the success in combat. Indeed, for all we know, it may have been the case that the winning battles were all correlated with yet another potential causal

[51] The analogy to the conditional, within group, ϕ correlation coefficient for interval scale variables is the partial correlation coefficient, $r_{xyc} = \frac{r_{xy} - r_{xc} r_{yc}}{\sqrt{1-r_{xc}^2}\sqrt{1-r_{yc}^2}}$.

This partial correlation coefficient represents the correlation between two variables, x and y, when the influence of a third variable, c, has been controlled. The Covariance Theorem for dichotomous attributes states that for any two attributes, x and y, and a third "control" attribute, c, it is possible to equate the universal covariance, C_{xy}, with a weighted average of covariances within control subgroups, and, in addition, a term involving a product of the covariances between y and c, and c and x:

$C_{xy} = P_c C_{xyc} + P_{\sim c} C_{xy\sim c} + C_{yc} C_{cx} / P_c P_{\sim c}$. Substituting $\phi_{xy}\sqrt{P_x Q_x}\sqrt{P_y Q_y}$ for C_{xy} and similarly for other C_s yields Equation (2). The Covariance Theorem for dichotomous attributes was first established by G. Udny Yule. Paul Lazarsfeld brought it to the attention of American scientists [Lazarsfeld 1958]. We have benefited from its illuminating treatment in [Alker 1971].

factor, *e.g.*, more Blue firepower, and that the users of contingency planning had more firepower. We still don't have all the relevant data.

It is easy to show that while causation implies correlation, the converse is false: simple correlation does not prove causation. Digging deeper into the data by examining the partial correlations within the two training subgroups between use of contingency planning and success in combat, we find a perfect positive correlation within the subgroup that had West Point training (ϕ_{xyc} = 1.00) and a moderate negative correlation within the untrained subgroup ($\phi_{xy\sim c}$ = -.50). So with these data, the relationship between use of contingency planning and success in combat is clearly confounded by training. Indeed, there is an interaction here between use of contingency planning and level of training in impacting combat outcome: with these data, it appears that use of contingency planning caused improved combat effectiveness only under the condition where the commander had West Point training; otherwise, it did not. Hence the findings from partial correlations with these data do not unequivocally support the general hypothesis that use of contingency planning causes successful combat outcomes. So we need a comparison group, more like the treatment group, which is unconfounded by extraneous variables.

We are obliged to consider the role of control variables in general in testing our causal hypotheses. Equation (2) above states the general Covariance/Partial Correlation theorem for correlation of three dichotomous variables, independent (x), dependent (y) and control (c). According to Yule's Theorem, any universal xy correlation is composed of a weighted average of the correlations within control subgroups plus the product of the independent and dependent variable correlations with the control variable. We assume, of course, that independent and control variables precede the dependent variable in time. Control variables, c, which are uncorrelated with the dependent variable, y, are not plausible explanatory factors in the first place; but those that are so correlated may provide rival explanations if they are also correlated with the independent variable, x. Bearing this in mind, is it possible to find a way to conduct an unequivocal test of our causal hypothesis on contingency planning and warfighting effectiveness? In the words of [Lazarsfeld 1958], "If we have a relationship between x and y and if for any antecedent test factor, c, the partial relationship between x and y does not disappear, then the original relationship should be called a causal one." But do we have a way to examine all plausible test factors?

1.2 Controlled Experimentation

Modern science provides us an answer to the question posed above: In the words of Albert Einstein, *"Development of Western science is based on two great achievements: the invention of the formal logical system (in Euclidean geometry) by the Greek philosophers, and the discovery of the possibility to find out causal relationships by systematic experiment (during the Renaissance)*" [Einstein 1953]. It is here that we must advance from the passive Aristotelian mode of empirical investigation to the active, experimental Galilean mode ushered in during the Renaissance. In the assertions

of the Nobel laureate Herb Simon [Simon 1957], and John Stuart Mill[52] a century earlier, the causal interpretation of a simple (or partial) correlation depends upon the presence of a compatible causal hypothesis and the absence of a plausible rival hypothesis to explain the correlation on other grounds. But Yule's Covariance Theorem, (2) above, states that any correlation can be decomposed into the weighted average of the partial correlations within control subgroups plus the product of the independent and dependent variable correlations with the control variable. Hence any new control variable, or combination of control variables, may provide a potential new rival explanation while washing out the original xy correlation in the subtables of partial correlations. Thus in testing our hypothesis that contingency planning causes improved combat effectiveness, we should control not only for training but also for Blue-firepower advantage, Blue-troop advantage, quality of C4I and other factors. Through the judicious use of control variables, which usually are not completely specified, we could then investigate the persistence of the original xy correlation in the control subtables as significant partial correlations and perhaps prune rival explanatory hypotheses down to a surviving root cause; but beyond successive prunings, the conduct of a controlled experiment enables us to ascertain precisely whether an alleged cause is a real root cause.

In a controlled experiment, the observed subjects (or units) are randomly assigned to the treatment group, here use of contingency planning, x, or to the control group, ~x, and the mean effectiveness of their combat performance, y or ~y, is measured and compared. Since the two groups are now statistically equivalent, any discovered difference in performance between the two groups is due solely to the treatment condition. Just such a procedure is followed in running clinical trials in the modern health sciences to determine the true efficacy of potential new healing drugs.[53] In the biologist's terms, this practice ensures that x is truly an exogenous variable. Ultimately,

---

[52] In [Cook and Campbell 1979].

[53] Popper, K.R. *The Logic of Scientific Discovery*, [Popper 1959]. For a generalization of Lazarsfeld's work on causality see H. Simon, "Spurious Correlation: A Causal Interpretation," *op. cit.* p.42-43. Simon shows that for multivariate causal modeling, using interval scale data, if and only if one can ensure the proper temporal sequencing of the variables and ensure that the error terms of the variables are uncorrelated with each other, is it safe to assume that the other variables are in fact "controlled for" or "held constant" or correctly "assumed to be random" as required for true causal relationships to be inferred. Our controlled experiment satisfies these conditions since in this context, X is a random variable with a random error term, u_x, and its correlation with u_y is necessarily zero. Otherwise there could exist some prior variable, C, spuriously affecting both X and Y and contributing to both u_x and u_y. So with controlled experimentation, all plausible rival hypotheses to explain r_{xy} as spurious and stemming from possible r_{xc} r_{cy} correlations are effectively ruled out. This task is much more difficult for non-experimental investigations. For non-experimental investigations involving two, three or more variables, it is necessary to carefully examine the validity of the assumption that the residual error terms of the variables are pairwise uncorrelated with each other. Otherwise there could easily exist some extraneous variable or variables, C_i, correlating both with Y and with any of the specified independent variables, X_i, even after the posited causal relations between Y and X_i have been taken into account. Thus, regardless of whether one's research is experimental or non-experimental, the investigator must somehow isolate sub-systems of variables from the complex environment and verify the non-correlation of residual error terms while making careful use of controls in order to draw legitimate causal inferences.

a controlled experiment affords the best causal test prospect, and it differs from a passive correlational study precisely because the process of active randomization disrupts any lawful relationship between, c, the characteristics of the antecedents of the subjects, *e.g.*, training, and their exposure to the treatment condition, x, *i.e.*, randomization in a controlled experiment effectively sets the value of the correlation between the independent variable (treatment condition) and any control variable to zero, $\phi_{xc} = 0$. Since $\phi_{xc} = 0$ in Equation (2) above for controlled experiments, the universal correlation, ϕ_{xy}, equals simply the weighted average of the partial correlations, ϕ_{xyc} and $\phi_{xy\sim c}$, for all c's: the spurious portion of the xy correlation, $\phi_{xc}\,\phi_{yc}$, has been nullified. Hence, it follows as in Lazarsfeld's assertion above that, in the context of a controlled experiment, if an observed correlation between x and y is significantly greater than zero, then the hypothesized relationship should be called a causal one. In Simon's terms, there is no tenable rival hypothesis to explain the correlation on other grounds. Thus controlled experiments provide the scientist a probative way of posing causal questions to nature such that her reply will always be revealing and sometimes profound.

1.3 Some Controlled Experimental Tests of Causal Hypotheses on Combat Effectiveness

Following this approach, a controlled experimental test of our hypothesis that use of contingency planning causes improved combat effectiveness (H1) was conducted in the TRADOC Analysis Center (TRAC) Lab at the Naval Postgraduate School (NPS) in January 1988 as reported in [Needalman *et al.* 1988].

H1: Use of contingency planning causes improved combat effectiveness.

This experiment utilized the fine grained, two-sided Janus wargame simulator (hence the name Janus for the two-faced Roman god) to provide a realistic combat setting as well as the capability to adjudicate combat moves and measure combat outcome in terms of the summated losses of Red- and Blue-warfighting platforms over the time course of the combat. The validity of the Janus simulator had been previously tested by comparing the time course of the Red- and Blue-attrition data at the battalion level from Janus-T to comparable data from low-intensity laser battles conducted by troops engaged in live exercises at the National Training Center. The fit was found to be, "strikingly similar during the force-on-force part of the battle" [Ingber 1989]. In setting up the experiment, 12 military officers, who were students at NPS, were randomly assigned to one of three, four-man teams. Each team played all four possible conditions resulting from crossing contingency planning/ single thread planning with high and low battle intensity. This procedure yielded a total of 12 three-hour trials, half of which involved the use of contingency planning. There were no significant differences between the trials in training, numbers of Red and Blue troops, available firepower, or available C4I. The question was, would use/non-use of contingency planning make a significant difference in combat outcome. In the combat scenario, US forces opposed Soviet forces who were threatening to close down Bandar Abbas and, with it, all Persian Gulf shipping. The US mission was to prevent Soviet forces from going through the Bam

Darzin Pass. The results of the experiment confirmed the contingency planning/successful combat outcome hypothesis: Across the 16 trials, use of contingency planning resulted in a 16% advantage to Blue in terms of attrition of Red forces per kilometer of advance ($Y = 26$ *cf.*[54] 22, $p < .001$).

The foregoing experiment is a replication of an earlier contingency planning experiment which was conducted utilizing the Joint Theater Level Simulator (JTLS) in the War Lab at NPS in August 1987 [MacMillan, Entin and Lenz 1988]. The subjects consisted of two random assignments of 14 officers to one of two teams, each organized into five command cells. Each team participated in four counterbalanced trials of three hours each for a total of four contingency planning trials and four single-thread trials. The warfighting scenario here also involved a Persian Gulf mission defending against a Soviet invasion. Here, again, use of contingency planning produced significantly greater Red losses than single-thread planning, yielding a 36% advantage for Blue ($Y = .84$ *cf.* .62, $p = .02$). So the general hypothesis that the use of contingency planning causes improved combat effectiveness is once again supported, and this causal relationship is shown to be invariant with respect to the particular wargame simulator or particular officers involved in the experiment. Furthermore, these observations cannot be accounted for with rival explanations of better training, more firepower, more Blue troops or better C4I since these factors were the same in the experimental and control conditions, and teams were randomly assigned to the different conditions.

Such controlled experiments have been conducted not only to test causal hypotheses regarding the combat effectiveness of particular military doctrines, but also to test causal hypotheses about the combat effectiveness of potential new C4I technologies.

H2: Use of a shared COP causes improved combat effectiveness.

A controlled experimental test of the hypothesis that use of a shared COP causes improved combat performance was conducted in the MIT Research and Engineering Corporation (MITRE) Command Center Engineering Lab in the summer of 1991 [Hiniker and Entin 1992]. This experiment utilized the Navy's Research and Analysis for Systems Engineering (RESA) wargame simulator for an air/sea battle set in the Persian Gulf. Eight experienced Naval officers were recruited from the faculty of the Naval War College and were joined with four retired Admirals to compose four, three-man teams. Each team played two COP trials and two control trials for a total of 16, three-hour trials, half of which utilized the COP prototype. There were no significant differences between the trials in numbers of Red and Blue troops, available firepower, training or doctrine. The question was, would teams using a cross echelon shared COP fed by both organic and national sensors perform better in combat than a control team with the high commander using only a national sensor fed big picture view and a pair of subordinate ship captains using only local tactical pictures fed by their organic ship sensors. In the combat scenario, US assets are under attack by Red craft, and the Blue team is required to sort through ambiguous information to determine who the attackers are and then take appropriate combat action. The results of the experiment confirmed

[54] *cf.*: Latin: "confer" which means compare.

the shared COP/combat effectiveness hypothesis: Across the 16 trials the ratio of Red losses to Red plus Blue plus Neutral losses was significantly greater when the Blue teams employed the COP (Y = .68 *cf.* .54, p = .04).

The foregoing experiment is a replication of the original COP prototype experiment conducted at the Naval Ocean Systems Center (NOSC) RESA Lab in spring 1990 utilizing the RESA wargame simulator, which was the first time a prototype of a shared COP was subjected to controlled experimental testing [Hiniker and Entin 1990]. Six experienced Naval officers were recruited in the San Diego area to join three, crisis-tempered, retired Admirals to compose three, three-man teams. Each team played four, three-hour trials, as above, for a total of 12 trials, half of which used the COP prototype. The Persian Gulf scenario was essentially the same as above. Employing the HEAT/OODA Loop Model, we derived two hypotheses on COP effectiveness: H2, as above, use of COP causes improved combat effectiveness; and, as a mechanism for this, H2B, use of COP causes improved situation assessment accuracy, later dubbed "Situation Awareness."[55] The results were inconclusive for H2, but showed substantial support for H2B: When using the prototype COP, Blue teams displayed significantly higher situation awareness, in terms of the proportion of the mission relevant set of warfighting platforms they were able to identify correctly (Y = .56 *cf.* .50, p = .02).

[55] The HEAT model measures the speed and accuracy of the command decision cycle composed of a sequence of six phases: monitoring, situation assessment, course of action development, outcome prediction, decision, direction of action… remonitoring. HEAT was initially applied to higher headquarters planning processes. At the tactical level of command decisionmaking, the abbreviated four phases of the similar OODA loop are applied to the decision cycle: Observe, Orient, Decide, Act…Reobserve. It has become an accepted tenet of military doctrine that warfighters should act fast, and inside the decision cycle of the adversary.

Site of experiment	Wargame simulator used	Combat effectiveness measure (y): for		Number of trials run	Significance of difference [a]
		Experiment group (x)	Control group (~x)		
H1: Use of contingency planning causes improved combat effectiveness.					
NPS TRAC LAB	Janus	.26	.22	12	p < .001
NPS WAR LAB	JTLS	.84	.62	8	p = .02
H2: Use of shared COP causes improved combat effectiveness.					
MITRE CCEL	RESA	.68	.54	16	p = .04
NOSC RESA LAB	RESA	-	-	12	n.s.
Idem but for H2 B	RESA	.56[b]	.50[b]	12	p = .02
DISA JDEF	RESA	.61	.42	5	p = .09
H3: Use of N-KRS replanning aid causes improved combat effectiveness.					
NOSC RESA LAB	RESA	.56	.59	24	n.s.

[a] All Significance of Difference probabilities (p) are from the *F* test tables for the ANOVA used in analyzing the experimental results (n.s.: non-significant).

[b] Situation Awareness Measure defined as the proportion of the mission critical set of Red, Blue, and Neutral warfighting platforms correctly identified.

Table 6 Controlled experimental tests of causal hypotheses on combat effectiveness

Another test of the shared COP/combat effectiveness hypothesis was carried out through another replication of the COP experiment in DISA's new Joint Demonstration and Evaluation Facility (JDEF) Lab in the summer of 1991 utilizing the RESA wargame simulator [Hiniker 1991]. Employing the same Persian Gulf scenario and design as above, albeit with a smaller number of trial runs, more support was found for H2. Across the five trials, the ratio of Red losses to Red plus Blue plus Neutral losses was significantly greater when Blue teams employed the COP (Y = .61 *cf.* .42, p = .09). In summary, three different controlled experiments, conducted in three different laboratory venues, with three different sets of subjects all provided significant support for the hypothesis that use of a shared COP causes improved situation awareness or

improved combat effectiveness. In all three experiments across dozens of trials, akin to dozens of small scale military exercises, the observed superior performance of the Blue teams using the shared COP cannot be explained by their use of more troops, more firepower, better doctrine or better training since all these factors were controlled by randomization of subjects in the design of the experiments. The discovered superior combat performance of the Blue teams that was observed and reported here was due solely to their use of a shared COP. All the observational evidence reported here is consistent with the proposition that use of a shared COP causes improved combat effectiveness; there is no tenable rival hypothesis that accounts for these findings.

H3: Use of N-KRS decision aid causes improved combat effectiveness.

Another C4I technology hypothesized to cause improved combat effectiveness (H3) was experimentally tested in the NOSC RESA Lab using the RESA wargame simulator in spring 1990 [MacMillan and Shaw 1990]. This technology, the Navy Knowledge-based Replanning System (N-KRS), was a computerized replanning aid designed to produce rapid air tasking orders for carrier-based air strike commanders. Six experienced Naval air strike commanders were recruited to play all four conditions of a two-wave Kamchatka Peninsula targeting scenario. Half of these 24 trials involved use of N-KRS. The results showed no significant difference in the proportion of Red targets successfully destroyed (Y = .56 *cf.* .59, p = n.s.). Despite the fact that replanning was accomplished significantly faster by the strike commanders when using the automated N-KRS aid, this advantage was offset in the overall command decision cycle by the fact that the experienced strike commanders made significantly less accurate estimates in their projections of target destruction when using the new aid. Thus H3 was not supported by the controlled experimental results, and N-KRS was sent back to the drawing boards for informed modification.

1.4 Conclusions

We have clearly demonstrated that controlled experiments provide unequivocal tests of causal hypotheses. *Post facto* controlled statistical analyses can approach the validity of such controlled experimental tests, but they seldom, if ever, produce unequivocal tests of causal hypotheses. We have also demonstrated that such controlled experiments are feasible and can be conducted in the warfighting area, in particular, with tests of the efficacy of certain military doctrines and certain C4I technologies alleged to improve command decisionmaking. In the process we have produced significant experimental evidence supporting the twin hypotheses that use of contingency planning and use of a shared COP by Blue commanders cause improved combat effectiveness. These replicated, controlled experimental findings permit no other explanation for the observations. Whenever multiple treatments were employed on a group, possible learning effects were ruled out as explanation by temporal counterbalancing. As summarized in Table 6, these controlled experimental observations supporting the two hypotheses on the causes of warfighting effectiveness are robust: they were found and replicated in five different experimental venues, employing three different wargame

simulators, Army, Navy, and Joint; and they involved more than 50 runs of man-in-the-loop combat trials with five different sets of Army, Navy, and Air Force officers. The scientifically sound practice of random assignment of subjects to treatment conditions employed here serves to define a controlled experiment and thereby rule out any rival explanations for significant findings. We have also demonstrated the utility of the method of controlled experimentation to delay the acquisition of certain immature prototyped C4I technologies as not significantly effective, while providing important diagnostics for improvement as part of an evolutionary development program. In each such experiment, the treatment condition served to isolate the cause. As shown in previous material presented in GUIDEx, one may, of course, also make informative and useful observations of factors thought to cause improved combat effectiveness by making careful use of quasi-experimental designs where the randomization requirement is relaxed; but then one is obliged to rule out, as much as possible, all plausible rival explanations for the findings by other means [Campbell and Stanley 1963].

Both sets of confirmed experimental findings are consistent with the HEAT or OODA Loop Model of command decisionmaking: use of the shared COP makes for more accurate situation awareness, or Observation (first O of the OODA), among Blue warfighters; use of contingency planning permits more rapid responses, or Orient-Decide-Act times, for a changed situation. Currently, both propositions are also in accord with new DoD emphases on Defense Transformation: use of the shared COP contributes to “information superiority”; use of contingency planning contributes to “flexible response.” Historically, DISA adopted the COP in 1995, converting it from an idea and a prototype into an integral part of the Global Command and Control System (GCCS), now DoD’s official C2 system. DISA has evolved and spread the COP continuously since 1995, now to more than 600 sites including the National Military Command Center and all Combatant Commander command centers. Recently the COP has been folded into the Global Information Grid (GIG) as part of DISA’s new Net Centric Enterprises Services. Indeed without a shared COP, current DoD emphases on Network Centric Warfare, as opposed to weapons platform based warfare, would not be feasible for our Joint Forces [Cebrowski and Garstka 1998].

The HEAT Model and OODA Loop variant that inspired the six combat experiments above, also suggest the combat utility of a Shared Map Planning/DCTS, that is a natural companion to the shared COP, to complete the command decision cycle for a warfighting team. Using this technology to speed the iterations of the decision cycle for a warfighting team, even more, should result in increased observations of success on the battlefield.

1.5 Discussion Relative to GUIDEx

This Case Study illustrates a campaign that successfully influenced the evolution of GCCS and the GIG. This discussion of the Case Study relative to GUIDEx follows a format designed to ease the interpretation of the results and hopefully allow comparison between Case Studies. In the discussion of this Case Study, the methods

and approaches used, and the results or lack of results are related to the 14 Principles of this guide as well as the 4 Requirements and 21 Threats to experiment validity in a table which answers the following questions:

1. Which of these 14 Principles are appropriate for this Case Study?

   The answers fall in one of these three categories: yes (Y), this Principle was or somewhat (S) was appropriate, or not at all (N).

2. Which ones were addressed during the Case Study?

   The answers fall in one of these three categories: yes (Y), this Principle was or somewhat (S) was addressed, or not at all (N).

3. How was it addressed during the Case Study?

#	GUIDEx Principle	Relevant	Addressed	How was it addressed?
1	Defense experiments are uniquely suited to investigate the cause-and-effect relationships underlying capability development.	Y	Y	Original experiments on COP prototype
2	Designing effective experiments requires an understanding of the logic of experimentation.	Y	Y	Controlled experiments
3	Defense experiments should be designed to meet the four validity requirements.	Y	Y	Controlled experiments
a	Ability to employ the new capability	Y	Y	In the three replicated COP experiments, players were given an initial throw away trial with a special scenario to familiarize them with the workings of the COP and train them in its use.
b	Ability to detect change	Y	Y	Since all three experiments yielded significant differences between control and experimental groups on the dependent variables of improved situation awareness or combat loss/exchange ratio, significant change was in fact detected.
c	Ability to isolate the reason for change	Y	Y	Ability to isolate cause-and-effect was ensured since subjects were randomly assigned to control group or treatment group or, in some cases, a completely counterbalanced within-subjects design was employed.
d	Ability to relate results to actual operations	Y	Y	Use of real-world scenario and experienced retired Admirals and ship captains enhanced ability to relate results to real operations.
4	Defense experiments should be integrated into a coherent campaign of activities to maximize their utility.	Y	Y	HEAT Program

#	GUIDEx Principle	Relevant	Addressed	How was it addressed?
5	An iterative process of problem formulation, analysis and experimentation is critical to accumulate knowledge and validity within a campaign.	Y	Y	A series of experiments
6	Campaigns should be designed to integrate all three scientific methods of knowledge generation (studies, observations and experiments).	Y	Y	Studies, experiments, exercises and observations
7	Multiple methods are necessary within a campaign in order to accumulate validity across the four requirements.	Y	Y	Replicated experiments with different set-ups and simulators. HEAT/OODA loop model[56]
8	Human variability in defense experimentation requires additional experiment design considerations.	Y	Y	Randomized assignment of subjects to treatments
9	Defense experiments conducted during collective training and OT&E require additional experiment design considerations.	N	N	N/A
10	Appropriate exploitation of M&S is critical to successful experimentation.	Y	Y	Experiments with wargames and HITL simulators
11	An effective experiment control regime is essential to successful experimentation.	Y	Y	JDL's[57] Basic Research Group
12	A successful experiment depends upon a comprehensive data analysis and collection plan.	Y	Y	Designed for MANOVA
13	Defense experiment design must consider relevant ethical, environmental, political, multinational, and security issues.	Y	Y	Employed American and British observers
14	Frequent communication with stakeholders is critical to successful experimentation.	S[58]	S	Early and final results presented to N6 (the primary US stakeholder for this result) and three years later to DISA lead engineers.

Table 7 Relation of CS1 to GUIDEx Principles

[56] Headquarters Effectiveness Assessment Tool, HEAT, Observe-Orient-Decide-Act, OODA loop.
[57] Joint Directors of C3 Laboratories.
[58] Players intended to use insights in their decision process.

If a Principle should have been addressed but not, how could it have been applied?

Principle 14 was only informally implemented, and then only after the initial COP prototype experiment. Engagement with a prototype decisionmaker could have appreciably shortened the three- to four- year time period between initial series of experiments and final government adoption.

If some Principles were applied, were they of any value to the Case Study from addressing them?

Principles 1, 7, and 13 were of particular value in this set of experiments. P1: Since controlled experiments provide a general method of testing causal hypotheses, including those of the form defense capability "x" causes improved warfighting effectiveness, positive results from well designed experiments provided very persuasive evidence on the potential effectiveness of the COP prototype. P7: The HEAT/OODA Loop model we followed also suggested the potential effectiveness of a natural companion technology for the COP, *viz*. collaborative planning tools. P13: Involvement of coalition partners in the experimental development phase of this technology has led to innovations in and tailoring of the technology to suit more rapid incorporation of the technology into effective coalition operations.

Case Study 2. UK Battlegroup Level UAV Effectiveness

2.1 Background

In the late 1990s, the UK MoD had begun to consider a range of roles for tactical UAVs, having had several years experience with the British Army Phoenix UAV, which was procured as a divisional level depth fire targeting system. In particular there was considerable interest in the concept of a unit-level UAV for use by maneuver battlegroups (BGs), to enhance their organic ISTAR capabilities. It was decided to commission a defense experiment to help understand the effectiveness of such a concept and how it compared with existing BG ISTAR capabilities.

2.2 Aim and Hypotheses

The aim of the experiment was to establish the increased effectiveness of UK maneuver BGs when equipped with a unit-level UAV. Thus the hypothesis was that if a UAV was used, the BG would receive better and faster intelligence and consequently their operational effectiveness would improve.

2.3 Type of Experiment

There were two pre-requisites for such an experiment:

1. A number of BGs executing the same types of mission.
2. An exemplar unit level UAV.

These factors played the major part in determining the type of experiment to be carried out.

It was clear from an early stage that sufficient BG-level activity could only be accessed by exploiting collective training exercises, something which would also provide a degree of control over the mission types and scenarios. The only UK-operated facility where consistently similar exercise missions were regularly played out was the British Army Training Unit Suffield[59] (BATUS), which comprises a maneuver area of approximately 50 x 60 km of unwooded rolling terrain (prairie).

The only economically viable means of producing a suitable UAV capability (for a low budget experiment) at the time was through simulation. In 1998, BATUS had acquired a GPS-based remote vehicle tracking system. Near real-time knowledge of the location of all tactical vehicles enabled a simple UAV simulation to be created, whereby a 2D footprint was “flown” over the tactical map and icons appeared when it overlaid a vehicle position. Outer “detection” and inner “identification” zones within the footprint were defined so that different levels of information could be relayed to the tasking HQ (*e.g.*, “unidentified vehicle,” or “tank” respectively). This approach had some obvious

[59] In Alberta, Canada.

limitations: it produced no imagery so only voice reporting could be done; it suffered from no geographic errors; reporting was always accurate (given the part of the footprint that targets were in); and the OPFOR could not react to it—it was ultimately stealthy from that perspective, as it only existed virtually. All of these factors contributed to the threats to the experiment's external validity.

Nevertheless, this simple approach, combined with the general realism of BATUS tactical engagement simulation[60] training was felt to be adequate for the task. Thus in terms of the types of experiment outlined in GUIDEx, this was a field experiment augmented by simple human-in-the-loop simulation stimulated by a live feed.

2.4 Experimental Treatments

There were two treatment groups, comprising six missions where a UAV was used and six where one was not. The experiment was able to exploit the consistency of BATUS missions and use four BG training exercises (notionally numbered A-D) so that each treatment group comprised two missions of each of the three different mission types, as indicated in the table below.

Mission	BG A	BG B	BG C	BG D
Delay mission[61]				
Meeting Engagement[62]				
Advance-to-Contact[63]				

Table 8 Assignment of the four BGs to the three missions.
Grey cells are where the UAV was used and White cells depict no UAV use.

For exercise intrusion reasons the UAV was not able to report on the general BG command net (as normally practiced by BG close recce) but had a dedicated radio link with the BGHQ. In addition there were some artificial ground rules, such as the UAV not being allowed to report the position of the enemy reserve (which was unrealistically close due to training area size constraints). When the UAV was used, a scheme of use was defined and applied rigorously. This fixed the number of flights per exercise mission and the duration of those flights.

[60] *i.e.*, using laser-based weapon effect simulators.
[61] A defensive action in which ground is traded for time.
[62] In which two opposing forces meet unexpectedly.
[63] In which own-force area of responsibility is exploited into enemy territory up to the point at which contact is made.

2.5 Results

The performance of the UAV as an information provider was compared against all of the Battlegroups' (BGs') other assets, *e.g.*, close reconnaissance (recce), Artillery Tactical Groups (arty tac gps), Anti-Tank Guided Weapon (ATGW), mortar platoons, armored squadrons and infantry companies. It was significantly better than all of them in all of the measures taken, including the following:

1. general information contribution in terms of number of reports,
2. the cumulative rate of information provision throughout each mission (*i.e.,* more reports were received earlier),
3. number of cluster sightings (those of most use), and
4. report content accuracy (artificial to a degree, due to the simple nature of the simulation, but nevertheless, restricted sightlines from ground locations often resulted in the content accuracy of reports from other assets being quite poor).

The acid test was the comparison of battle outcomes derived from the weapon effects simulators across the two treatment groups. It is unusual in experiments of this type, especially when piggybacked on training, to achieve statistically significant "bottom line" results when the independent variable is an information-related system, rather than (say) a tank gun upgrade. In this case, significant results were indeed achieved. However, to the great surprise of the analysis team, they showed that use of the UAV actually *reduced* the loss exchange ratio from almost 1.5 to just over 0.5. In other words the concept appeared at first glance to be a highly significant battle-losing weapon. But why was this?

The team had placed a human factors analyst in each of the BGHQs (ostensibly for other purposes) and he was able to shed some light on the problem. He reported that:

1. The lack of prior training in tasking and exploiting the product of UAVs resulted in some enduring problems in those areas.
2. BGHQs manning levels were insufficient to collate and process UAV information into useful and timely intelligence.
3. The sheer quantity of UAV product overwhelmed an already overstretched HQ. This was exacerbated by the direct UAV link into just one cell of the HQ (G2). What appeared to be happening was that the swamping effect of the UAV sighting reports on the G2 cells caused them to become effectively dysfunctional. This appeared to result in other elements of the BGHQs taking over parts of the G2 function and spending less of their effort coordinating the activities of the BG sub units.

In other words, the cause of the reduction in loss-exchange ratio (LER) could be isolated in a superficial sense to the presence of the UAV, but it was not possible to identify the underlying cause with any great degree of confidence. However, it was possible to deduce *associative relationships* between reduced LER and those phenomena described above. Thus an experiment which began as a largely equipment-focused event, managed to conclude that at BG level:

1. the means of communicating UAV product is a major issue;
2. commanders and staff officers need to be trained in UAV use;

3. the manning of the BGHQs needs to be changed if organic UAVs are to be used effectively; and
4. the quantity of product exacerbates any information flow problems in HQs.

This was only the second time that intrusive experimentation had been allowed during BATUS exercises and the first time than any synthetic play had been sanctioned (albeit of a very simple nature). After several more years of close cooperation with the training authorities and BATUS, the same experimentation team was operating:

1. full live-into-virtual sensors (UAVs and airborne battlefield radar platforms);
2. a synthetic CGF wraparound to augment the live activity and provide better stimulation of the above; and
3. deploying networks of surrogate battle management systems.

This enabled the UK MoD to glean useful knowledge about UAVs, digitization and ISTAR integration at formation level.

2.6 Lessons Learned

The key lessons learned from this activity relate to the intrusive exploitation of collective training exercises for experimental purposes. This was only the second time that the British Army training authorities had allowed intrusive exploitation of training at BATUS, their premier maneuver training centre. Detailed “rules of engagement” had to be agreed with Commander BATUS on the ground and sometimes (not surprisingly) compromises on the experimentation side had to be made. This all sounds rather negative, but in fact this was one of several such activities that contributed to a wholesale change of mindset about experimenting during army training in the UK. Only two years after this activity, as the in-service dates of various new C3 ISTAR systems approached, there was a general acceptance that, if done responsibly and professionally, experimentation such as this can actually improve the training and development experience. This is particularly the case for smaller nations when they are training for potential coalition operations with the US, because their next generation of equipment may (in a generic sense) already be in US service.

Also, this was unusual for a field experiment in that it was able to exploit a whole season’s worth of similar field training exercises in the same location and consequently score rather better in experiment validity Requirements 2 and 3 (ability to detect and isolate the cause of change) than is usual for field experiments. Moreover, it showed clearly that in the right circumstances, it is possible to undertake genuine “true” experimentation, not just observational studies, on the back of training exercises.

A key lesson learned from this activity was that before new capabilities are deployed to experiments or exercises, all lines of development must be brought together. In other words, the doctrine, procedures, organizational structures and training must all be adjusted to enable the new capability effectively, as called for by experiment validity Requirement 1. However, it was a constraint of the training environment that there was no possibility of changing the BGHQ ways of working to any meaningful extent, or providing them with any significant pre-training.

Another unusual aspect of this work (in the UK at the time, at least) was that it employed simple human-in-the-loop simulation technology in concert with instrumented live action, so that the UAV "response cell" was able to provide a realistic feed to the supported HQ based on the modeled UAV's technical characteristics and the positions of the exercising vehicles on the ground. This was just the beginning of live/human-in-the-loop simulation at BATUS and in subsequent years, proper DIS-based UAV; Airborne Stand-Off Radar and attack helicopter simulators were used in BATUS exercises, together with simulated rear and flanks action created by computer generated forces. This all enabled:

1. various experiments and observational studies to be carried out during training;
2. much greater contextual richness for the exercises themselves; and
3. the Field Army to become acquainted with, and to begin to develop TTPs for, various near-future capital equipments (which could already be encountered in coalition ops with the US, *e.g.*, JSTARS).

Thus by creating a hybrid between experiments using human-in-the-loop simulations and Field Experiments, it was possible to achieve some of the benefits of both.

Detailed comments, in terms of the GUIDEx Principles, are laid out below in a table, which answers the following questions:

1. Which of these 14 Principles are appropriate for this Case Study?

   The answers fall in one of these three categories: yes (Y), this Principle was or somewhat (S) was appropriate, or not at all (N).

2. Which ones were addressed during the Case Study?

   The answers fall in one of these three categories: yes (Y), this Principle was or somewhat (S) was addressed, or not at all (N).

3. How was it addressed during the Case Study?

#	GUIDEx Principles	Relevant	Addressed	How was it addressed?
1	Defense experiments are uniquely suited to investigate the cause-and-effect relationships underlying capability development.	Y	Y	This was the forerunner of a program of experimentation in support of the UK Watchkeeper UAV program.
2	Designing effective experiments requires an understanding of the logic of experimentation.	Y	Y	Despite the potential confounding factors inherent in training exercises, a reasonable two-treatment design was achieved.

#	GUIDEx Principles	Relevant	Addressed	How was it addressed?
3	Defense experiments should be designed to meet the four validity requirements.	Y	Y	See below.
a	Ability to employ the new capability	Y	S	With hindsight, Requirement 1 was not well satisfied, but failures in this area still yielded new knowledge.
b	Ability to detect change	Y	Y	By exploiting the instruments of the resources listed at Table 8.
c	Ability to isolate the reason for change	Y	S	Partially, the changes in effectiveness could be attributed to UAV use but the underlying reasons for the LER reduction were not entirely clear.
d	Ability to relate results to actual operations	Y	Y	Yes, in that this was largely a field experiment, but some of the simplifications of the UAV simulation placed constraints on the conclusions.
4	Defense experiments should be integrated into a coherent campaign of activities to maximize their utility.	Y	S	This work was done in concert with wider OA modeling programs but not explicitly as part of a deliberately-designed campaign.
5	An iterative process of problem formulation, analysis and experimentation is critical to accumulate knowledge and validity within a campaign.	N	N/A	Not part of a coherent campaign
6	Campaigns should be designed to integrate all three scientific methods of knowledge generation (studies, observations and experiments).	N	N/A	Not part of a coherent campaign
7	Multiple methods are necessary within a campaign in order to accumulate validity across the four requirements.	N	S	Not part of a coherent campaign. This had the potential to be part of a deliberately designed M-E-M campaign, but at the time that wasn't really part of the UK analysis psyche. There was, however, a looser relationship with some related operational analysis.
8	Human variability in defense experimentation requires additional experiment design considerations.	Y	S	Because this was based on four real battlegroups, there were no artificial constraints on human variability.
9	Defense experiments conducted during collective training and OT&E require additional experiment design considerations.	Y	Y	Many training exploitation issues had to be overcome to perform this experiment.

#	GUIDEx Principles	Relevant	Addressed	How was it addressed?
10	Appropriate exploitation of M&S is critical to successful experimentation.	Y	Y	A simple virtual UAV simulation used in concert with instrumented live action was fundamental to the experiment. The key point here is that "appropriate" in this case meant sufficient only to provide a credible response cell and therefore a simple simulation was all that was needed.
11	An effective experiment control regime is essential to successful experimentation.	Y	Y	Made more difficult due to training constraints, close liaison with the training authorities enabled a sufficient degree of control to be achieved. Nevertheless, this experiment was always only going to be able to measure quite large differences between the treatments.
12	A successful experiment depends upon a comprehensive data analysis and collection plan.	Y	Y	A detailed data collection plan was agreed before the experiment. In addition, the fortuitous location of a human observer in the BGHQs emphasized the importance of capturing data at the key nodes of each causal chain, especially if something unexpected occurs.
13	Defense experiment design must consider relevant ethical, environmental, political, multinational, and security issues.	Y	S	The use of human-in-the-loop simulation rather than live UAVs greatly aided the safety aspects of the experiment. However, there were no major ethical, environmental, political, multinational or security issues to be overcome.
14	Frequent communication with stakeholders is critical to successful experimentation.	N	N/A	Not part of a coherent campaign.

Table 9 Relation of CS2 to GUIDEx Principles

Case Study 3. UK NITEworks ISTAR Experiment

3.1 Background

In 2002, the UK MoD, in partnership with industry, set up the "Network Integration Test and Experimentation works" (NITEworks). NITEworks is closely aligned with the UK Network Enabled Capability initiative (broadly akin to NCW) and is intended to be an experimental environment which allows the customer community to assess the benefits of NEC and the options for its effective and timely delivery.

The NITEworks program comprises a number of "themes" and one of the early ones was ISTAR. The first question set for the ISTAR Theme was concerned with how the UK's information requirements management (IRM) process might be improved. The question had a strategic to operational level setting (*i.e.*, UK-based HQs; the in-theatre Joint Force HQ; and the in-theatre component HQs (Maritime, Land and Air).

3.2 Aim and Hypothesis

The aim of the experiment was to establish the relative merits of changing two attributes of the IRM system: infrastructure and toolset. Thus the main hypothesis for the experiment was:

"The efficiency of processing Commanders' requests for information will be improved by the provision of a single domain infrastructure and/or collaborative working tools."

The infrastructure change involved replacing the current partitioned architecture (security; command hierarchy) with a non-partitioned infrastructure. This meant that all players could see and search all information (including answers to requests for information (RFI)) present in the system, irrespective of where the information was held. It also freed players to send RFIs directly to the cell or agency that they thought was in the best position to answer the request.

The toolset change involved introducing a web-based collaborative working environment with a common IRM database. The collaborative toolset provided a unified management system that extended across all four levels of command represented in the experiment. This provided a central numbering system for RFIs, with full visibility to all players of the current location and status of all RFIs in the system.

3.3 Type of Experiment

The experiment was a "true experiment" as described in Part II of this document. It was a bespoke activity (*i.e.*, not piggybacking on a training exercise), but it did use a training facility, the UK Joint Warfare Training Centre, and a related exercise scenario. The various player cells were physically separated but provided with communications and an information infrastructure appropriate to the treatments being run. An exercise control (EXCON) cell fed RFIs into the system at various levels and the cells themselves

created their own, as would be expected. Thus a key EXCON function was to ensure that overall RFI numbers circulating in the system were broadly comparable between treatments.

3.4 Experimental Treatments

There were two discrete independent variables: infrastructure (partitioned and single-domain); and toolset (current—various disparate databases; and the collaborative working tool). This led to a set of four treatments, comprising all combinations of these two factors. Hence:

Treatment 1 (current toolset in a partitioned infrastructure): Treatment 1 was designed to represent the current process. Players receiving an RFI from a demander would first conduct a search in a limited/partitioned database for extant products and, if nothing was found, forward via email to the next cell upward (JFHQ in the case of lower component commands) for processing.

Treatment 2 (current toolset in a non-partitioned infrastructure): Treatment 2 was unchanged from Treatment 1 in terms of toolset used. However, players were now permitted to search for information in a much wider database, simulating the removal of system and security boundaries. Furthermore, players were allowed to pass RFIs to any cell in the simulated system; the command hierarchy adopted in Treatment 1 was abandoned.

Treatment 3 (collaborative toolset in a partitioned infrastructure): the infrastructure reverted to that adopted in Treatment 1, but the collaborative web-based tool was introduced. RFIs were stored on a central repository such that an RFI had only to be entered onto the system for all players to have visibility of it. This contrasted with the current toolset in that each time an RFI entered a new cell, that cell had to copy the information from the RFI and add it to the intra-cell management database system.

Treatment 4 (collaborative toolset in a non-partitioned infrastructure): Treatment 4 permitted players to operate both in a non-partitioned infrastructure (as with Treatment 2) and with the collaborative tool (as with Treatment 3).

Every treatment was run for nine and half hours over two days. Players were told to behave as if the scenario was running continuously (*i.e.*, it was effectively frozen at the end of the first day of each treatment). Because of the relatively short duration of the scenario, RFIs were issued with time-for-completion in the range of hours rather than days.

3.5 Results

The following measures (dependent variables) were taken, covering a range of objective and subjective response data:

1. RFI completion

2. Count of RFIs completed per treatment
3. RFI completion time
4. Functionality assessment
5. Player workload
6. Information awareness
7. Shared knowledge and trust
8. Interviews and observations

For brevity, this short summary will focus on the RFI completion measures.

3.5.1 Count of RFIs Completed per Treatment

One of the major determinants of system efficiency in the experiment was the number of RFIs processed in each treatment. For formal analysis, this was expressed as the number of RFIs completed per half hour of treatment, and was analyzed in a two (toolset: current; collaborative) by two (infrastructure: partitioned; non-partitioned) repeated-measures General Linear Model (GLM) analysis. The results are shown at Figure 51. The pattern of results suggests that toolset was the major determinant of system efficiency: it can be seen that the collaborative toolset approximately doubled the rate at which RFIs were processed under the current toolset.

The analysis results confirmed a significant main effect of toolset, $F(1, 67) = 35.5$, $p < .001$, but no significant main effect of infrastructure. There was also no evidence of a significant interaction between the two factors.

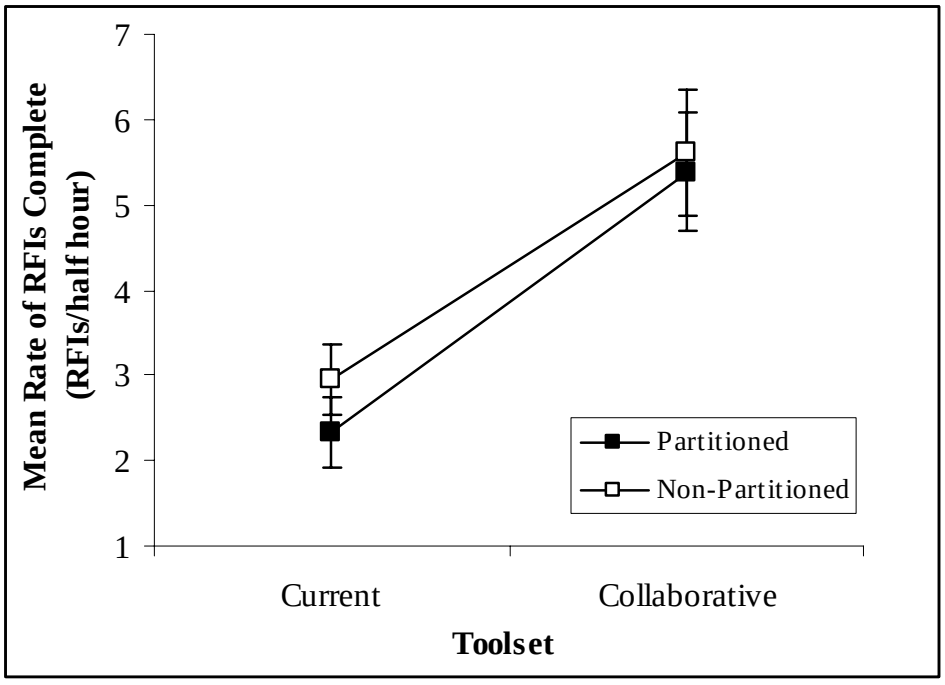


Figure 51 Number of RFIs completed per half hour as a function of toolset and infrastructure. Standard error bars are shown.

3.5.2 RFI Completion Time

To complement the rate of RFI completion data, the average length of time each RFI took from injection (by EXCON/demander) to completion (return to EXCON/demander) was also investigated. Time spent outside of the system under investigation (*i.e.*, any facility represented by EXCON) was not included in this evaluation. Results indicate that RFIs were processed more rapidly under a collaborative toolset than under a current toolset. Results also suggest that under a current toolset, RFI completion time benefited from a non-partitioned infrastructure rather than a partitioned infrastructure.

A two (toolset: current; collaborative) by two (infrastructure: partitioned; non-partitioned) repeated-measures General Linear Model (GLM) analysis on the natural logarithmic function of completion time for an RFI demonstrated a significant main effect of toolset, $F(1, 228) = 207.7$, $p < .0001$, but no main effect of infrastructure. The interaction between the two factors, however, was significant, $F(1, 228) = 6.4$, $p < .05$. The interaction, as Figure 52 suggests, arises because RFI time-to-complete is influenced by infrastructure only under a current toolset condition, not under a collaborative toolset.

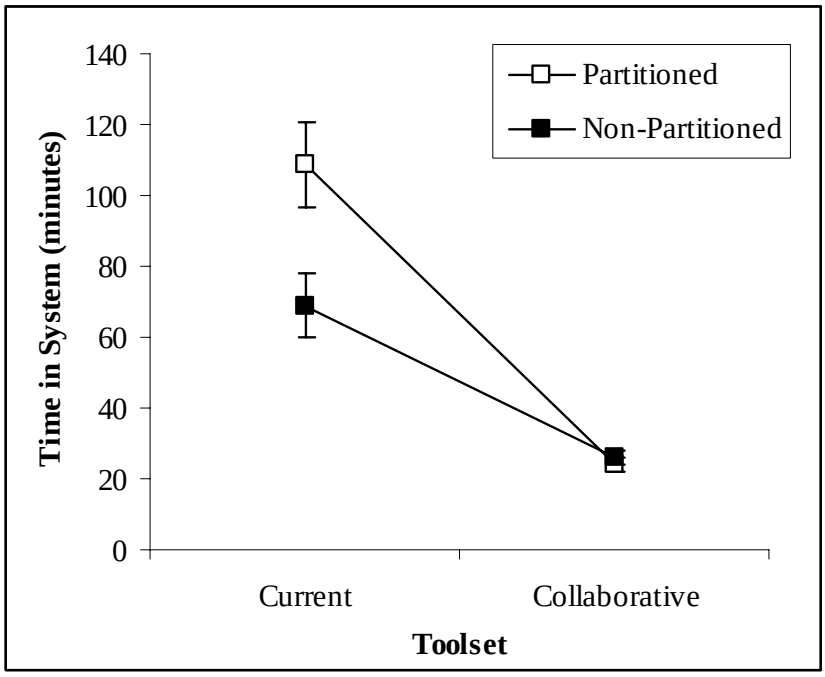


Figure 52 RFI time-to-complete as a function of toolset and infrastructure. Standard error bars are shown.

3.5.3 Results Summary

The results of the experiment were unequivocal in demonstrating the benefits of a collaborative toolset over current instantiations within an IRM system. Indeed, all the MoEs (of which only two were described above) reported a significant main effect of toolset. Compared to the current toolset, players performed more efficiently (in fact players doubled their productivity), experienced less workload, perceived themselves to

have greater information awareness, exhibited more trust and considered shared awareness to be higher under the collaborative tool. Generally, it was felt by players that the benefits of a collaborative toolset included: greater ability to search for existing or duplicate RFIs; greater speed and ease of RFI entry; better consistency of numbering; and an improved ability to check and track RFIs.

The results also suggest strongly that the players benefited from operating under a non-partitioned rather than partitioned infrastructure in some circumstances. Players demonstrated improvements to RFI tracking, favorable functionality ratings and experienced less workload under the non-partitioned infrastructure compared to the partitioned infrastructure.

3.6 Lessons Learned

This activity was essentially an experiment using an analytic wargame focused on information requirements management. However, there was no directly supporting simulation, as in this case it was possible to run the whole thing using EXCON-inserted RFIs. Thus a general lesson learned here was that although modeling and simulation is usually at the heart of defense experimentation, it does not necessarily have to be so.

This was a classic "true" experiment with two independent variables (or factors), resulting in four treatments. It yielded statistically significant results that showed the relative effect of the two factors, a satisfying result, given the difficulty of achieving this in the C3I domain. Given the push to satisfy experiment validity requirements 2 and 3, the whole spectrum of control activities, including rigorously managing the experiment's scope, were the biggest considerations. Thus the most important guiding principle for the experiment was Principle 11.

The lessons learned in terms of the 14 GUIDEx Principles are laid out in a table, which answers the following questions:

1. Which of these 14 Principles are appropriate for this Case Study?

   The answers fall in one of these three categories: yes (Y), this Principle was or somewhat (S) was appropriate, or not at all (N).

2. Which ones were addressed during the Case Study?

   The answers fall in one of these three categories: yes (Y), this Principle was or somewhat (S) was addressed, or not at all (N).

3. How was it addressed during the Case Study?

CS3 NITEworks ISTAR Experiment

#	GUIDEx Principle	Relevant	Addressed	How was it addressed?
1	Defense experiments are uniquely suited to investigate the cause-and-effect relationships underlying capability development.	Y	Y	The results of this experiment directly supported changes in the areas of equipment procurement, doctrine, structures and training.
2	Designing effective experiments requires an understanding of the logic of experimentation.	Y	Y	This was a classic experiment, with a causal hypothesis, two independent variables and sound statistical analysis.
3	Defense experiments should be designed to meet the four validity requirements.	Y	Y	See below.
a	Ability to employ the new capability	Y	Y	All requirements met.
b	Ability to detect change	Y	Y	Many effects detected.
c	Ability to isolate the reason for change	Y	Y	Careful design and good control resulted in the reasons for the detected effects being clearly identifiable. See explanation under P8 of how potentially confounding learning effects were alleviated.
d	Ability to relate results to actual operations	Y	S	The limited manning meant that this was the weakest of the four requirements, although the number of RFIs per day (to which the results could have been sensitive) was considered by the players to be representative of their recent operational experience.
4	Defense experiments should be integrated into a coherent campaign of activities to maximize their utility.	Y	Y	This experiment was part of the UK NITEworks program of visualization and experimentation in support of the delivery of Network Enabled Capability.
5	An iterative process of problem formulation, analysis and experimentation is critical to accumulate knowledge and validity within a campaign.	N	Y	The subject matter of this experiment came about as the result of considerable problem formulation activity. In fact the originally stated problem concerned Collection Coordination (the “CC” of CCIRM) but it became apparent during problem analysis that IRM needed to be tackled first.
6	Campaigns should be designed to integrate all three scientific methods of knowledge generation (studies, observations and experiments).	Y	S	The campaign, or “theme,” of which this experiment was a part, comprised 2-3 experiments and an observation event. However, integration with wider studies was not designed in from the start.

#	GUIDEx Principle	Relevant	Addressed	How was it addressed?
7	Multiple methods are necessary within a campaign in order to accumulate validity across the four requirements.	Y	N	In reference to the M-E-M paradigm, the use of these results in concert with a workflow model and other methods (*e.g.*, a field experiment on the back of a major training exercise) could have enabled the results to be generalized to much greater degree.
8	Human variability in defense experimentation requires additional experiment design considerations.	Y	S	This was a single-group design and therefore could not address the effects of between-group variability on the results. The main potential confounding factor encountered was a learning effect that might have impacted on the treatments. Good toolset training alleviated the problem, and it was also possible to test for a learning effect within each treatment (which lasted two days): none was detected.
9	Defense experiments conducted during collective training and OT&E require additional experiment design considerations.	Y	S	This was a bespoke experiment but used a training establishment to supply the infrastructure and also a training exercise scenario. Some consideration had to be given to the validity of the scenario for experimentation purposes, but otherwise this was not a big issue.
10	Appropriate exploitation of M&S is critical to successful experimentation.	Y	Y	The M&S approach was considered for this experiment, but was deemed an unnecessary complication. A straightforward EXCON-operated MSEL was sufficient to provide all player simulation. So “appropriate” in this case was, in fact, none.
11	An effective experiment control regime is essential to successful experimentation.	Y	Y	This was a well-controlled experiment from design, through planning and into execution, resulting in Requirements 2 and 3 being completely satisfied.
12	A successful experiment depends upon a comprehensive data analysis and collection plan.	Y	S	A detailed data collection plan was agreed before the experiment. Well-briefed observers captured both hard and soft data during execution, as planned.
13	Defense experiment design must consider relevant ethical, environmental, political, multinational, and security issues.	Y	S	No ethical, multinational, political or environmental issues. However, security was a key factor in the experiment (associated with the partitioned infrastructure). In reality, partitions exist between various high security compartments. The difficulties of experimenting with these in practice led the team to use “exercise” security classifications, which still enabled the impact of real security partitions to be reflected.

#	GUIDEx Principle	Relevant	Addressed	How was it addressed?
14	Frequent communication with stakeholders is critical to successful experimentation.	Y	Y	The NITEworks approach is a close partnership with the sponsoring body. The key stakeholders were kept in frequent communication and were regular attendees at the experiment itself.

Table 10 Relation of CS3 to GUIDEx Principles

Case Study 4. Pacific Littoral ISR UAV Experiment

4.1 Background

As part of a campaign to evaluate new ways of delivering intelligence, surveillance, and reconnaissance (ISR) capabilities, the Canadian Forces Experimentation Centre (CFEC) conducted the Pacific Littoral ISR Experiment One[64] (PLIX-1) off the west coast of Vancouver Island in Canada from July 7 to 11, 2003. This experiment was designed to examine concepts of UAV employment and Integrated ISR Architecture. The integrated ISR architecture prototype developed for the experiment connected UAV operations located at Tofino on Vancouver Island to the Maritime Operations Centre at Canadian Forces Base Esquimalt and onward to the National Defence Command Centre in Ottawa. Information acquired from the UAV was to be fused with other sources and used to enhance the Recognized Maritime Picture (RMP).

The critical operational issue within PLIX-1 was the integration of UAV data into an information and intelligence (I2) system. An Integrated Intelligence Surveillance and Reconnaissance Architecture (IISRA) was implemented for the experiment that permitted UAV-sensor contact information to be provided to three levels of command; tactical, operational, and strategic in near real time. UAV radar and optical imagery products were made available to network users at all three levels of command through an Internet-based imagery server.

The Chief Maritime Staff (CMS) was the operational sponsor for the experiment with Commander Maritime Forces Pacific (MARPAC) as the office of primary interest (OPI) together with the Director Joint Force Capabilities (DJFC) as a key stakeholder mandated to establish a joint UAV program in Canada.

4.2 Aim and Hypotheses

The experiment was designed to assess the utility of a multi-sensor payload, medium altitude, long endurance UAV and the UAV integration requirements for an IISRA to support the construction and maintenance of a Recognized Maritime Picture (RMP) within a specific littoral operations area.

A hypothesis was developed for the experiment in consultation with the sponsor:

If a UAV patrols a designated operations area of littoral waters, then all surface contacts are detected, continuously tracked, and positively identified in the experimental RMP of the operations area before the end of the patrol.

This hypothesis was deemed unambiguous and falsifiable. Given the size of the operations area, the utility of the UAV in maintaining a recognized picture was postulated at the maximum possible level.

[64] A second part of the experiment, called PLIX-2, employed an uninhabited surface vessel (USV) to support operations of a ship at sea and to build a local recognized maritime picture.

In retrospect, the hypothesis should have included a statement of conditions under which the hypothesis is expected to hold. Real-time integration of UAV sensor data and contact report management through the ISR architecture is a pre-condition for the subsequent of the above proposition. This precondition was partially achieved after technical problems were resolved in the initial stages of the experiment. The hypothesis was tested on four separate patrols and was falsified at each instance. All surface contacts were not positively identified nor classified to type in the experimental RMP at completion of the patrol. Subsequent analysis revealed many issues and lessons learned as to why the hypothesis was not supported. The knowledge generated could not have been found easily without experimenting with the real equipment.

4.3 Type(s) of Experiment, Series or Campaign

PLIX was conducted as a live field experiment off the West Coast of Vancouver Island using an augmented operational ISR architecture (linked to operational systems), actual ISR sensors including a line-of-sight UAV, trained operators, and planners at the Maritime Operations Centre.

The Israeli Aircraft Industries (IAI) Eagle[65] 1 aircraft was leased for the experiment. This UAV was equipped with a TAMAM Multi-mission Optronic Stabilized Payload (MOSP) electro-optical/infrared camera and an ELTA 2022-A(V3) maritime patrol radar. The Eagle flew from the airport in Tofino, British Columbia. The experiment crew established a deployed UAV squadron at a prepared strip in approximately 10 days with connections to the Maritime Operations Centre classified systems at MARPAC headquarters through a one-way data link. Following a test flight program, four experiment flights (one per day) were completed by July 11. All flights were limited to daylight hours, a line-of-sight data link, VFR flight conditions, and periodic local air traffic de-confliction. Positional data from a USN Yard Craft as a vessel of interest and several at sea Canadian Coast Guard ships was recorded for ground truth purposes.

The experiment was subject to environmental conditions and weather effects, to concurrent maritime surveillance operations, to local procedures, and to operator inputs. Airspace was reserved in advance for the purpose of the experiment although coordination with local air traffic control was required during transits in and out of the operations area. The experiment provided a very realistic setting to test the hypothesis, but was complicated by intervening factors that were not anticipated at the outset.

4.4 Treatments

The design called for a comparison of the recognized maritime pictures and mission plans of two treatments effected by command teams working independently. The ordinary command team (OCOM) acting as a control group was presented with the

[65] This UAV is also known as the “Heron” within Israel and is marketed in Europe and North America as the Eagle 1 by European Aeronautic Defence and Space (EADS) Company. Eagle 1 is the piston engine model and Eagle 2 is the larger turboprop variant with longer range, higher speed, altitude and payload.

current RMP derived from existing surveillance assets while the experimental command team (XCOM) had access to ISR data from the UAV while on continuous patrol as well as other surveillance assets and the classified recognized picture. It is important to note that the deployed UAV mission Commander did not have access to the classified picture although he could discuss operations over a classified phone link.

Figure 53 illustrates that scenario and mission inputs were provided to both teams. This information was used to plan UAV sorties which in turn generated contacts and imagery used to build a local UAV recognized picture (A-RMP). This picture together with the ordinary RMP were available to the experimental command team, thereby producing an experimental RMP to be evaluated with regard to the hypothesis and compared with the ordinary RMP. Similarly the XCOM mission plans were compared with those produced by the OCOM according to various metrics.

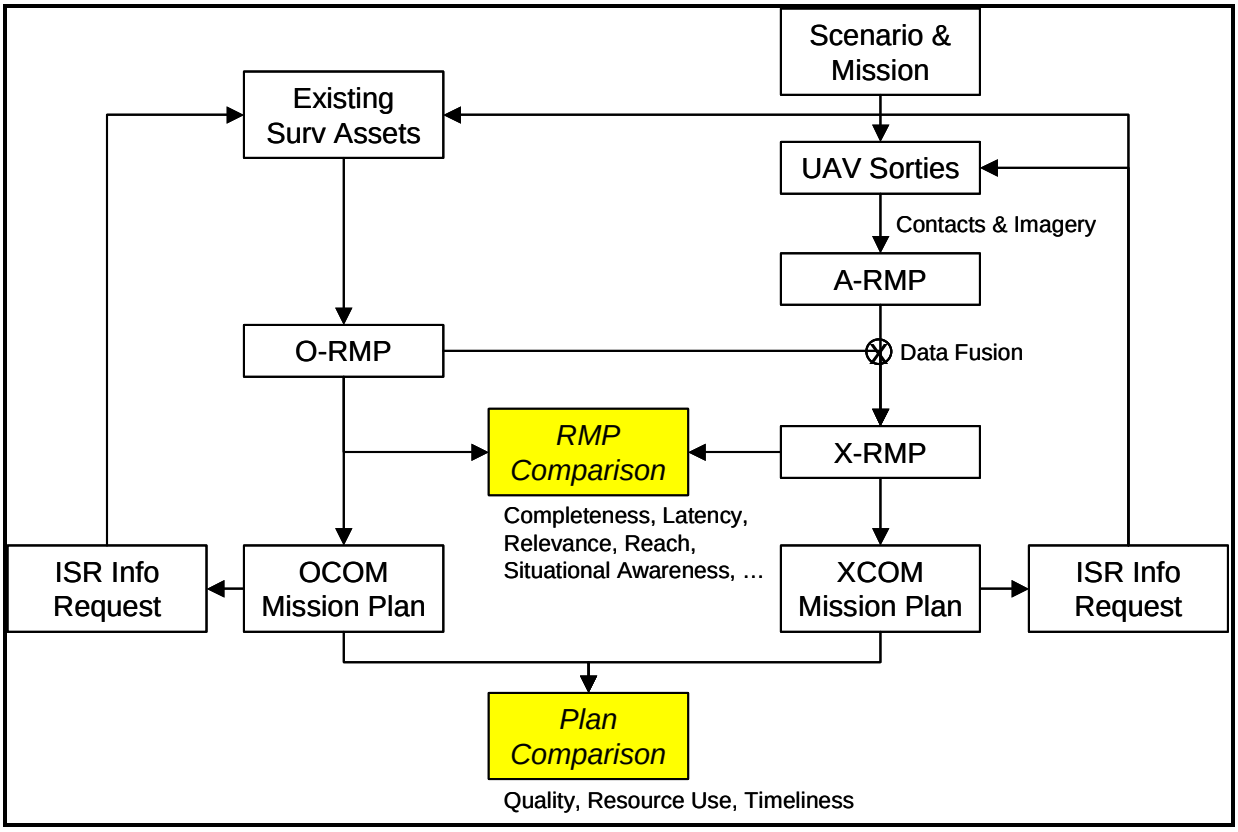


Figure 53 PLIX-1 experiment design schematic

The Canadian Forces Force Planning Scenario 8 (Surveillance/Control of Canadian Territory and Approaches) was used as the generic context for the experiment. A total of four vignettes provided specific events and tasks. In order of complexity and execution, the missions were:

1. Build and maintain a recognized picture within the littoral area of operations as part of normal peacetime surveillance operations, then support a Search and Rescue operation for an overdue simulated vessel.

2. Locate and track a simulated vessel with illegal immigrants. The detection of a vessel suspected of a pollution violation pre-empted the operation. The UAV payloads provided a positive identification of the pollution violation suspect and color imagery of the suspicious emissions were turned over to local authorities.
3. The previous vignette was re-played which required the command teams to track a simulated vessel of interest suspected of smuggling illegal immigrants.
4. Search and locate a simulated terrorist vessel, track and provide real time targeting information to simulated control assets.

The experiment network utilized commercial telephone lines to transfer data between the two primary sites. Network bandwidth of 2 x 56 kb/s maximum capacity was established as a representation of the capacity that is often available in deployed operations.

A USN Yard Patrol craft acted as the vessel of interest (VOI) in all 4 iterations. The experiment analysts became aware that "intelligence" had been acquired by operators prior to commencement of experiment execution. Had the hypothesis not been falsified, it would have been difficult to make definite conclusions regarding the utility of the UAV when integrated within the ISR architecture. This illustrates the importance of a robust experiment control regime.

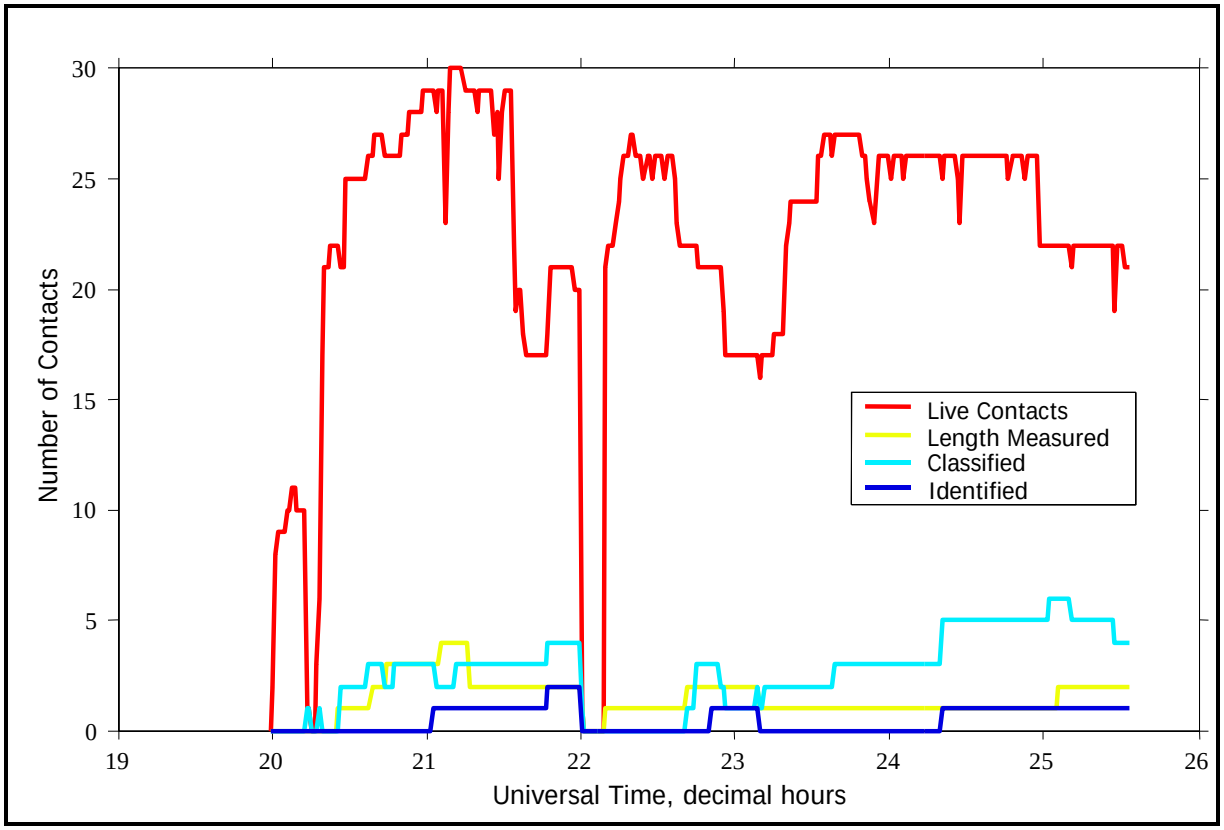


Figure 54 Number of contacts detected, classified to type and by length, and identified over time (universal time) for a typical patrol

As part of the analysis, a theoretical model was developed for the time evolution of detection, classification and identification based on differential equations. The model

was estimated using empirical data as shown above for each iteration and at each of the command team nodes to obtain parameters related to ramp-up and steady state probabilities of detection, classification and identification.

4.5 Broad Results

Results for the ordinary command team (OCOM) are classified and so it is only possible to discuss incremental contributions the UAV made to the recognized picture. Despite limitations and constraints imposed by weather, system architecture and the relative inexperience of the Canadian Forces' operators, it was assessed that the UAV made a measurable contribution to development of the RMP in the Patrol Area.

In terms of ISR capability, the utility of the UAV was assessed in six broad areas:

1. **Detection:** Immediately on activating the radar in search mode in the patrol area, the UAV provided initial detection of approximately 20-30 contact tracks via the Global Command and Control System–Maritime (GCCS-M). No false contacts were apparent although some targets were later classified as flotsam in tide rips after further investigation using Spot SAR and EO capabilities. Probability of detection was estimated at approx. 80-90%.
2. **Tracking:** The tracking of all detected contacts was assessed as good when compared to data provided by ground truth vessels. Latency in contact report dissemination affected tracking performance. The accuracy of the data was limited to one nautical mile because of the current OTH-GOLD format. The accuracy that could have been achieved, given the accuracy of the UAV navigation system as claimed by the manufacturer, was less than 20 meters. Issues related to unique track labeling due to reuse of track identifiers were found upon detailed analysis. This negatively impacted situational awareness at the Maritime Operations Centre.
3. **Classification:** Initially, no accepted methodology for classifying targets was established but the payload operators quickly developed a process to prioritize contact classification. The Inverted Synthetic Aperture Radar (ISAR) mode and the EO capabilities of the UAV were used to eliminate targets that were larger (or smaller) than the Vessel of Interest (VOI).
4. **Identification:** The main tool used for identification was the EO sensor. This capability was limited by the flying altitude restrictions in the OP area and frequent cloud cover. When conditions permitted, the identification was relatively quick and accurate and was often achieved by imaging the target's nameplate. Identification was also achieved by correlating other data available in the GCCS-M database through the regular RMP feeds such as the Tofino Vessel Traffic Management System data, shipping databases, and other source position reports.
5. **Tasking:** The UAV ISR mission tasking methodology evolved over the duration of this experiment as XCOM developed a tasking process to ensure clear and concise direction was given to the Mission Commander for any contingency tasking. A tasking template based on the land ISR doctrine was created and made available for ISR tasking but was not used. There was at least one instance of mis-communication between the XCOM team and UAV mission commander that resulted in a lost opportunity to identify the target.
6. **Situational awareness:** The increased situational awareness achieved by XCOM due to the UAV resulted in appreciably different mission plans when compared to OCOM. Given the incompleteness and latency of the OCOM recognized picture and resulting higher levels of uncertainty, OCOM took on a risk averse posture by planning to deploy significant resources to investigate and react to the simulated crisis. Unfortunately some assets would have been unavailable to deploy during parts of the operation (this aspect was played with paper assets with control providing feedback). In the case of the terrorist vessel, XCOM had high confidence of

locating and tracking the VOI having eliminated many contacts, while OCOM assigned significant assets to carry out ISR tasks to prosecute. Identification and continuous tracking of the VOI by the UAV was not achieved in the last two missions due to weather and track management.

There were many lessons learned as a result of this experiment, which have been documented in an experiment Quicklook report [Newton *et al.* 2003]. Primary lessons learned include:

1. **Recognized picture**: The unavailability of the classified operating picture to the UAV commander resulted in unnecessary effort classifying contacts that had been reliably acquired by other sources. The one-way data communication channel between UAV operations and the Maritime Operations Centre was imposed due to security restrictions and resulted, at least partially, in platform centric operations. Additionally, a wide variance in GCCS experience and knowledge was observed among operators. Manual operator data fusion procedures must be formalized in standard procedures and trained to sufficient levels of proficiency.
2. **Collaborative information environment:** The design, implementation, and maintenance of appropriate communications, information technology, and information management infrastructure are vital to the passing and fusion of sensor data for the benefit of personnel connected to the IISRA. Dedicated C4 technicians were essential for systems and network integration and maintenance.
3. **LOS UAV data link**: UAV altitude limitations caused by the use of line-of-sight data links dramatically limited the operational capability of the MALE UAV. Any Medium or High Altitude UAV acquired by the Canadian Forces must be capable of beyond-line-of-sight (BLOS) operations. Because of the limited operating ranges and endurance of Tactical UAVs, it is desirable that Tactical UAVs have an option to include a BLOS mode.
4. **Weather**: Weather was a major factor in all flight operations. Visual meteorological conditions were required to launch and recover the UAV and for transit to the patrol area. Once in the patrol area, reduced weather limits were acceptable as long as the UAV avoided icing conditions. Launch and/or recovery times were often adjusted to meet minimum weather conditions and in some flights a solid undercast layer below the minimum operating altitude prevented the use of the EO/IR sensor to identify radar contacts by name.
5. **Sensors**: Since the effective employment of sensor payloads of the experiment UAV was limited by weather conditions, various sensors need to be considered in future ISR experiments that can classify and identify contacts in all Canadian weather conditions. This could include automated identification system (AIS) as well as electronic support measure (ESM) sensors capable of identifying contacts through unique signatures.

The experiment involved the efforts of a 15 person integrated project team (IPT) from multiple organizations led by CFEC over a period of approximately 8 months. The cost to lease and operate the LOS UAV was on the order of one million Canadian dollars (CADs) while other network, administration, consulting services and travel costs amounted to approximately two million CADs.

4.6 Lessons Learned and Interpretation in Terms of the GUIDEx Principles

The importance of problem definition and scoping was demonstrated during PLIX. Initial planning for the Integrated ISR and UAV experiment proposed a much larger set of objectives to be accomplished. Working with sponsors, experiment designers advised

planning staff of the risks in conducting a multi-node large-scale multi-objective experiment without risk reduction. The understanding of how to integrate UAVs into an existing ISR architecture was insufficient at the initial stage of experimentation. It was decided to conduct a simpler experiment on the West Coast, and to examine related objectives in a subsequent experiment the following year on the East Coast (ALIX). This was a very wise decision since several unforeseen issues were encountered that had to be resolved related to point-to-point integration.

The revised hypothesis below could have been established for the experiment. It would nonetheless have been falsified due to a number of factors highlighting that other factors were at play for effective ISR such as mission coordination, information sharing, and effective data fusion.

If a UAV patrols a designated operations area of littoral waters, then all surface contacts are detected, continuously tracked, and positively identified in the experimental RMP of the operations area before the end of the patrol ***when*** *UAV sensor data is made available in real time and contact reports are managed and uniquely tagged.*

Much of the learning comes from insights about conditions under which relationships hold true and the correctness of a theoretical model.

The lessons learned in terms of the 14 GUIDEx Principles are laid out in a table, which answers the following questions:

1. Which of these 14 Principles are appropriate for this Case Study?

   The answers fall in one of these three categories: yes (Y), this Principle was or somewhat (S) was appropriate, or not at all (N).

2. Which ones were addressed during the Case Study?

   The answers fall in one of these three categories: yes (Y), this Principle was or somewhat (S) was addressed, or not at all (N).

3. How was it addressed during the Case Study?

#	GUIDEx Principles	Relevant	Addressed	How was it addressed?
1	Defense experiments are uniquely suited to investigate the cause-and-effect relationships underlying capability development	Y	Y	Gathered empirical data in the physical, information and cognitive domains on effects and underlying conditions.
2	Designing effective experiments requires an understanding of the logic of experimentation.	Y	Y	Design established to make direct comparisons between treatments and record conditions.

CS4 Pacific Littoral ISR UAV

#	GUIDEx Principles	Relevant	Addressed	How was it addressed?
3	Defense experiments should be designed to meet the four validity requirements.	Y	S	See below.
a	Ability to employ the new capability	Y	S	A training program and more lead-time would have been beneficial to stabilize the setup, and resolve technical issues encountered in the first iteration.
b	Ability to detect change	Y	Y	Extensive data collection on multiple MoEs
c	Ability to isolate the reason for change	Y	S	Control group established. Human variability noted but subjects not randomized across iterations.
d	Ability to relate results to actual operations	Y	Y	Scenario vignettes and capabilities representative of actual surveillance operations
4	Defense experiments should be integrated into a coherent campaign of activities to maximize their utility.	Y	Y	CFEC multi-year UAV and ISR campaign plan
5	An iterative process of problem formulation, analysis and experimentation is critical to accumulate knowledge and validity within a campaign.	Y	Y	A series of progressively more complex experiments was conducted. Results and lessons learned from two previous experiments were included in this experiment.
6	Campaigns should be designed to integrate all three scientific methods of knowledge generation (studies, observations and experiments).	Y	Y	National and TTCP UAV studies were consulted, team members observed other nation's experiments (*e.g.*, UK JUEP, US UAV Time-sensitive operations), and exercises.
7	Multiple methods are necessary within a campaign in order to accumulate validity across the four requirements.	Y	S	Live experiment using commercial technologies and operational systems in typical missions. Lack of constructive and HITL simulations prior to going live. Experimental results were used to fit a theoretical dynamic ISR model. Additional M&S should have been used prior to execution.
8	Human variability in defense experimentation requires additional experiment design considerations.	Y	S	Assignment of subjects to teams was randomized, but not across scenarios during execution. Should have capture more data on operator knowledge and skills to better fulfill P8.
9	Defense experiments conducted during collective training and OT&E require additional experiment design considerations.	N	N/A	N/A
10	Appropriate exploitation of M&S is critical to successful	Y	N	Insufficient time and resources were assigned to experiment preparation. A theoretical model should have been developed and investigated in a controlled

#	GUIDEx Principles	Relevant	Addressed	How was it addressed?
	experimentation.			environment.
11	An effective experiment control regime is essential to successful experimentation.	Y	S	Experiment designers worked with military controllers to ensure unbiased treatments during execution. Initial scenarios were conducted with variation in the architecture due to technical problems. VIP tours impacted on data collection.
12	A successful experiment depends upon a comprehensive data analysis and collection plan.	Y	S	Recorded conditions and quantified outcomes in relation to hypothesis. SA data collection would have benefited from a survey pre-test.
13	Defense experiment design must consider relevant ethical, environmental, political, multinational, and security issues.	Y	Y	Subjects were tasked within their military occupations. Local authorities were consulted for operations near a national park. Security policies were followed given presence of foreign contractors.
14	Frequent communication with stakeholders is critical to successful experimentation.	Y	S	Regular situation reports and briefings were provided during the experiment to sponsors and primary stakeholders. Quicklook results were shared with a wide audience within 30 days. Final documentation of detailed analytical results have been delayed due to competing demands.

Table 11 Relation of CS4 to GUIDEx Principles

In this experiment, more attention should have been given to Principle 8 related to human variability. Variability in GCCS operator proficiency only became evident during execution. The control group also suffered from challenges in motivation given the lack of visibility of new capabilities that were being examined. Randomizing teams across scenarios may have resulted in increased operator engagement and additional lessons learned. As per Principle 7, the exclusive reliance on live experiments may have limited return on investment since basic integration issues, dependencies, and training requirements, could have been identified in a controlled environment.

The ability to develop and maintain the recognized maritime picture was affected by track management. The re-use of track labels by the UAV ground control station only became evident during post-analysis and caused track fusion errors. Applying Principle 12 on data collection requires that few or no assumptions be made as to the correctness of data elements at key data collection points. It is highly desirable to have simple data summary and reporting tools in place during execution to recognize problems early and to incorporate in control mechanisms in subsequent iterations.

Principle 14 on effective communication was problematic. Experiment planning for the next event started immediately upon return from execution due to required lead times. Lessons learned from the analysis of PLIX were incorporated in a subsequent event as results became available.

Principles applied to the design and execution of this experiment which were of particular value are:

1. **Principle 3 The four requirements:** The ability to use the capability was achieved through flight tests and architecture validation. Results can be readily related to real operations given the scenarios that were used and the area of operations in consideration.
2. **Principle 4 Campaign planning:** ISR and UAV campaign plans were instrumental in ensuring progressive and logical goals. Given the multiple concepts being examined, it was important to recognize dependencies and to work on critical elements in a logical fashion. The integration of UAV information in real time and the management of imagery on a common server was a precursor to follow-on experiments involving multiple nodes. Each experiment in the campaign plan involved an increasing degree of complexity and risk.
3. **Principle 12 Data analysis and collection:** The hypothesis was unambiguous in its data requirements. Checklists were produced to ensure all data were captured at source. Additional assets could have been employed to enhance ground truth data and situation assessment (SA) surveys would have benefited from a pre-data test.

Despite the hypothesis being falsified in all case iterations, valuable lessons were learned about necessary conditions required to achieve improved ISR mission effectiveness. Whether these conditions are sufficient remains the task of follow-on experimentation.

Case Study 5. An Integrated Analysis and Experimentation Campaign: Army 21/Restructuring the Army 1995-99

5.1 Background

In 1986, the Defence White Paper *"Defending Australia*" instituted "Layered Defence" of Australia's northern approaches. Air and naval forces were to close the air-sea gap and the Army had restricted warfighting roles in defense of continental Australia against "adversary forces that penetrated the air and naval barriers." Defence's conceptual response was "Army in the 21st Century" (A21) a review aimed at optimizing the army for the defense of Australia.

The basis of this concept was, "detect, protect, and respond." Surveillance and reconnaissance forces operated in the broad and focal areas around critical civilian and military infrastructure (detect/respond) with close protection forces forming barriers around the key points (protection). The review was followed by an evaluation, the "Restructuring the Army" Trials (RTA) that in turn laid the analytical foundations for the Army Experimental Framework [Australian Army 2000].

A21/RTA was a Defence initiated review, which meant it had high-level endorsement; the trials officer (a brigade commander) had direct access to the Chief of Army and reported frequently (indication of good P14 adherence), was well funded and had the interest and participation of all relevant parties. In hindsight it can be reviewed as a complete campaign although A21 was initiated as a stand-alone study. Consequently, when sections within Defence questioned early A21 conclusions, it was determined that trials had to be conducted [Fisher, Brennan and Bowley 2003].

5.2 Aim and Objectives

The aim of the A21 Review was to define an appropriate organization, equipment, doctrine and preparedness for the Army of the future within the defense concept of defending Australia. In particular the review was to define the strategic tasks to which the Land Force could contribute, identify options for achieving the core tasks, and propose options for optimization, considering capability mix, readiness and equipment upgrades.

The aim of the RTA Trials was to optimize the RTA Task Force, the core of the optimized Land Force envisioned in the A21 Review. However, from an analytical perspective, the intent of the Chief of Army was more important; his intent was:

"that RTA should initiate a development process for Army that is…dynamic and evolutionary…, …setting doctrine and force structure on constantly converging paths which anticipate the requirements of future operations…"

5.3 The RTA Task Force Trials

There were two components of the trials: the RTA Task Force, including organizational, personnel, equipment and facilities requirements, and doctrine for the land environment based on the concepts for operations recommended in the A21 Review.

The trial focused on five objectives, themselves the product of a yearlong problem definition process, and based on the following five critical areas[66].

1. **Effectiveness of Embedding.** The embedding of combined arms at unit level so contributes to synergy and therefore tempo and combat effectiveness, that it is worth the trade-off in penalties for training and technical control, maintenance, logistic support, and cost.
2. **Depth of Embedding.** The embedding of combat arms at low levels within the unit structure enhances combat power by its responsiveness at the actual point of need on the battlefield, while achieving all the advantages of tempo and synergy of embedding. This does not preclude the concentration of combat power as necessary by the unit or TF commander, and incurs an acceptable cost.
3. **The Effect of Information on the RTA Task Force (Information).** The ability of the RTA Task Force to collect, process and distribute information so increases the tempo of decisions and the flexibility of the main effort that less combat power is required.
4. **The Adequacy of the RTA TF Combat Power (Combat Power).** The enhancements to the detection capability and the timeliness of response, mean that the TF does not need to hold dedicated reserves of combat power as traditionally was the case, and still has adequate capacity to meet the routine tasks necessary to conduct effective operations within its TAOR.
5. **The Endurance of the RTA TF (Endurance).** TF aims through its organization and doctrine to generate a high tempo of operations and has the administrative and logistics capability and flexibility to sustain 15 months in conflict of which up to 30 days may be on combat operations.

5.4 Methods Employed

A21 used a seven-step method for data analysis and collection summarized in Table 12. Subsequently, the RTA Trial addressed the uncertainties of Step 5 “Determine the level of effectiveness.”

The RTA Trials presented a complex multi-dimensional problem that had to be refined before any analysis could be conducted. It was essential to use techniques that determined the critical tasks, elements and factors in order to focus the field trials on gathering data that was relevant. Consequently, the most important aspect was to ensure the critical issues of the system were stressed during the analysis, if this had not occurred there may have been no measurable impact, reflecting the fact that each system was performing well within its capabilities. The key therefore was to determine stressful scenarios through seminars, TEWTs[67], CPXs[68] and wargames to focus the

[66] During the trials they were called hypotheses and are reproduced in this document verbatim. They can however be written as IF-THEN hypotheses, for example Critical Area 1: IF combined arms are embedded at unit level THEN synergy and therefore tempo and combat effectiveness increases and it is worth the trade-off in penalties for training and technical control, maintenance, logistic support, and cost.
[67] Training Exercise Without Troops.

modeling, simulation and field trials on the critical operational issues. The process had to be iterative because it was possible that the initial studies would not identify issues subsequently discovered in the field trials. Therefore the field trials had to be a mix of data collection at the lower force levels and "full operational situations" where the entire unit or formation was stressed. The campaign was flexible enough to revisit areas if new factors were discovered at any stage (related to P5).

Step	Title	Details
1	Tasks	Determine the tasks that the information collection system must perform.
2	Methods	How may these tasks be achieved?
3	Force elements	What force elements are available, and what are their performance characteristics?
4	Measures of effectiveness	Note that the results were not reported against MoE
5	Desired level of effectiveness	Determine desired level of effectiveness within a given geographical region against a given adversary.
6	Matching	Match methods and force elements to desired levels of effectiveness.
7	Force structure	Combine the preferred force elements into an organization structure, integrate the system into the larger combat system, and determine the C3I system for the regional force.

Table 12 A21 methodology

The analysis method for RTA phase 1 was adapted from OT&E and operations analysis with a detailed problem definition phase and model-exercise-model iterative analysis. The most important aspect of problem definition was to ensure the critical operational issues of the system were analyzed as missions and tasks were conducted under stressful conditions. This was attempted to maximize the chance of identifying the impact of changes to the system.

The analytical method employed was based on the iterative cycles of model-exercise-model (P5). The objectives of the modeling phases were to measure the performance of the system, determine the reasons for this performance, confirm the selection of trial units and critical operating issues, and focus the field trials. The ultimate aim of the modeling was to predict the performance of the system beyond the environment of the field trials. Modeling also provided one of the means of aggregating the performance of the activities tested. The test phases confirmed the predictions of the modeling in a specific environment, provided data for subsequent modeling, and tested aspects that could not be modeled. A range of tools was used for modeling and testing (as recommended by P7), see Figure 55.

[68] Command Post Exercises.

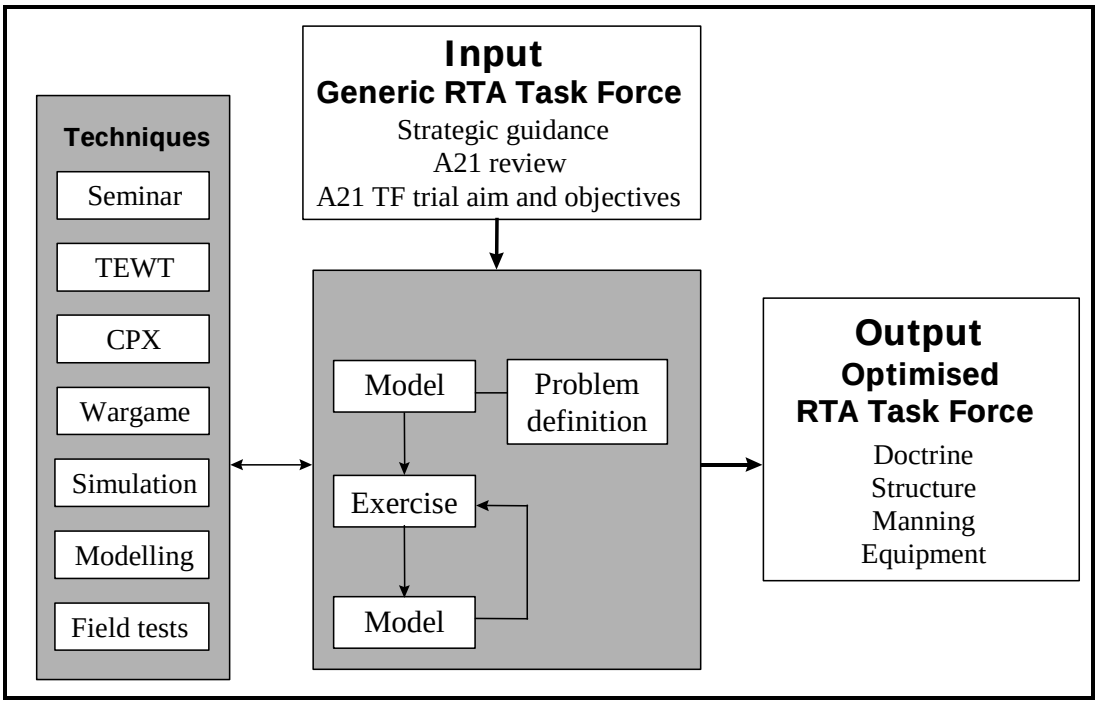


Figure 55 Outline process for the RTA Trials

5.5 Types of Campaign

The relationships between the activities conducted during RTA Phase 1 are shown diagrammatically in Figure 56. The nature of the overall campaign (P4) was a problem definition phase that utilized wargames and field exercises supported by seminars and studies (P6), followed by a detailed analytical phase where the results of each activity were used to define the objectives of subsequent activities (P5). The grey arrows indicate the main effort of the brigade. The first year, 1997, was devoted to problem definition and the major products were the analytical framework and trials campaign. 1998 was the main analysis phase, which produced the major report and supporting analytical papers. Each type of activities, *i.e.*, seminars, field exercises, wargames, simulations and review/reporting, are outlined in Figure 56 and discussed below.

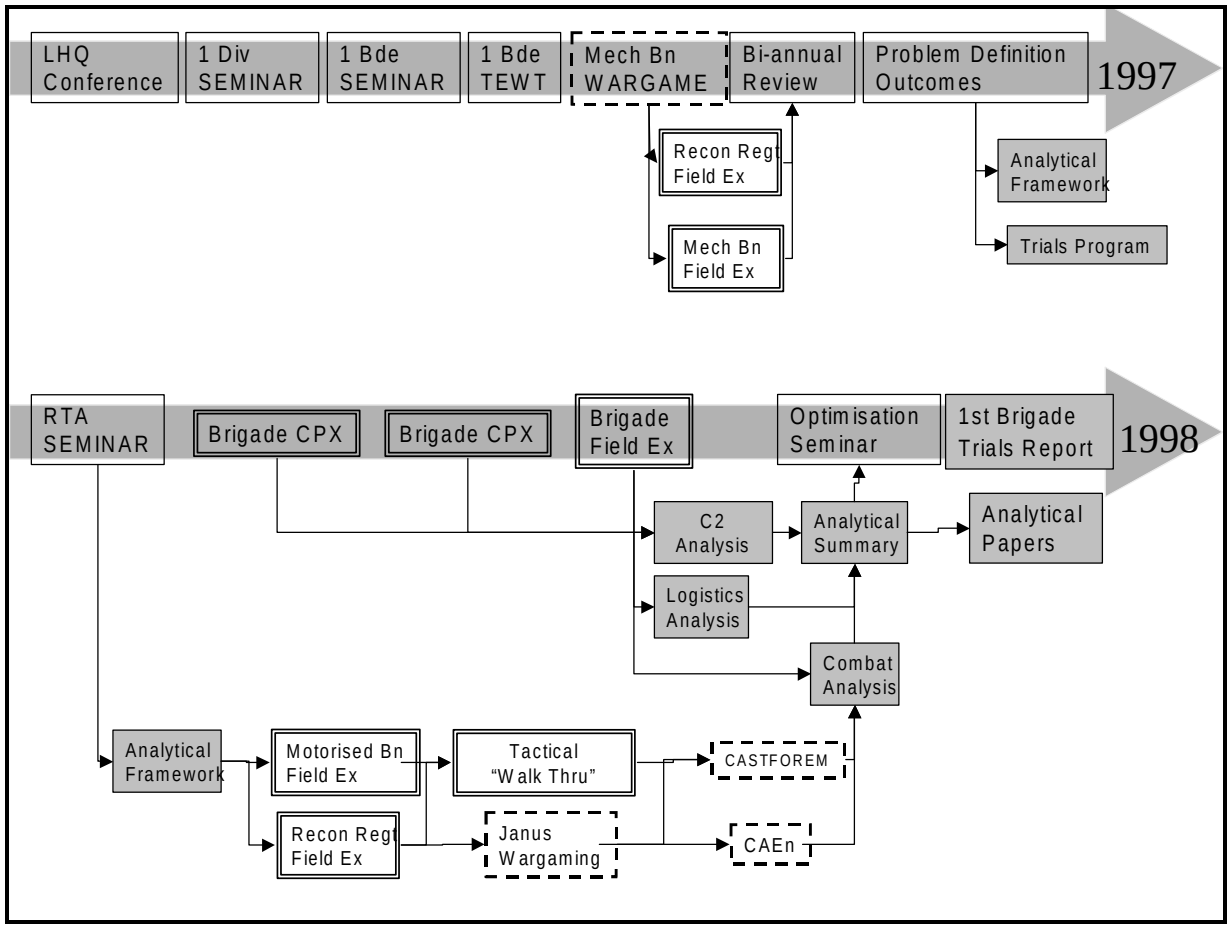


Figure 56 Outline of RTA Analysis

Seminars (Solid Lines). A number of seminars were conducted during this period, initially they were used to identify the critical areas and other aspects of the trials, finally they were used to integrate the results of the other activities into a coherent set of recommendations.

Live Experimentation (Double Lines). Although referred to as live experiments there was no attempt to control the activities to increase internal validity, however precise observations of actions and their environment were recorded to:

1. determine subsequent activities
2. provide specific insights
3. inform CAEN and CASTFOREM modeling
4. develop discussion points for the RTA Phase 1 Optimization Seminar

This ensured that the experiments conducted with constructive simulations and wargames were addressing the correct issues, with the correct forces in the correct environment.

Date	Trial or Exercise	DSTO reports
May 97	Delphis Oracle 97	Command Post Exercise (CPX)
Aug 97	Silicon Safari 97	CPX
Aug 97	Tiger Rage 97	Field Exercise
Sep 97	Flashing Sabre 97	Field Exercise
Oct 97	Northern Trilogy II	CPX
May 98	Predators Crawl 98	Field Exercise
June 98	Sabre Draw 98	Field Exercise
June 98	Rising Sun 98	Field Exercise
July 98	Silicon Safari 98	CPX
July 98	Winter Sun 98	Field Exercise
Aug 98	Phoenix 98	Field Exercise

Table 13 Trials, exercises and experiments studied during RTA

Wargaming and Simulation (Dashed Lines). Three formal sets of wargames were conducted; the first to explore concepts for protecting vital assets such as airfields; the second to evaluate concepts for the employment of a motorized battalion; and finally a major series investigating close combat operations in a range of operational environments with various force structures. These games refined the scenarios used in constructive simulation addressing close combat in open and urban terrain.

Modeling and Simulation. Modeling and simulation had three roles in the methodology of this Case Study. In the problem definition and the initial modeling phases it provided early system problem identification, and focused and highlighted critical operational issues. During the predictive modeling phase it allowed the system to be assessed against unavailable threats, supplemented and extended test data, and permitted extrapolation of the results by modeling the capabilities in additional physical environments and against threats. The most important characteristic of the use of modeling and simulation in this evaluation process is that it provided insights as well as absolute quantitative answers. Simulation replay, used as an activity to drive an After-action review (AAR), is as useful as a statistical analysis of the loss exchange ratios.

Studies were used to validate the simulations to increase confidence in the results. The studies targeted three areas through the validation process; vegetation representation, systems data (specifically lethality and vulnerability) and the representation of motorized company operations. The physical models of detecting, moving and engagement were not validated in Australia because all the simulations were validated by the developmental agencies.

Simulations and wargames offer powerful representations of the mechanics and psychology of military operations. In the past their use has been limited to the quantitative investigation of loss exchange ratios, and system comparison. By integrating the wargames and simulations into a network of complementary analytical tools ranging from seminars to field trials, it is possible to use the rich synthetic environment they offer to gain insights into the workings of military systems. The

important prerequisites are to clearly identify the scenario that offers the best information and to understand the assumptions and limitations of the wargame or simulation.

5.6 Conclusion

In this case the original concept was found to fail during the testing period, and it was through the accumulation of reinforcing evidence across a range of techniques that led to a coherent case to reject the concept. Consequently, the field exercises gained increasing significance because they had to compellingly demonstrate failure of the concept while clearly retaining their objectivity. Subsequently a clear case was made for an alternative concept, which was eventually adopted as the over-arching force development concept for the Australian Army.

The campaign demonstrated the importance of detailed problem definition; a process that took 12 months and involved extensive analysis in itself. The task force reorganized and developed doctrine in parallel with the analytical development so both the analytical framework and the participants were prepared for the extensive trial program in the second year.

The other feature of the campaign was the refinement of options through wargaming prior to simulation, and validation of the results in field trials and finally the extrapolation of the field results through simulation. This allowed the strengths of each technique to be exploited (for example the external validity of field trials and the internal validity of simulation) and exploit other techniques to increase the validity of the campaign as a whole.

The lessons learned in terms of the 14 GUIDEx Principles are laid out in a table, which answers the following questions:

1. Which of these 14 Principles are appropriate for this Case Study?

   The answers fall in one of these three categories: yes (Y), this Principle was or somewhat (S) was appropriate, or not at all (N).

2. Which ones were addressed during the Case Study?

   The answers fall in one of these three categories: yes (Y), this Principle was or somewhat (S) was addressed, or not at all (N).

3. How was it addressed during the Case Study?

CS5 Restructuring the AU Army

#	GUIDEx Principle	Relevant	Addressed	How was it addressed?
1	Defense experiments are uniquely suited to investigate the cause-and-effect relationships underlying capability development.	Y	S	Constructive simulation was used for the experiments, and provided the additional rigor required for compelling force development arguments. However the resource allocation in the campaign was biased toward live experiments to the detriment of other analytical methods.
2	Designing effective experiments requires an understanding of the logic of experimentation.	Y	Y	Experiments with constructive simulations were designed to isolate the changes and the variables manipulated were force elements. Results were extrapolated through studies and observation.
3	Defense experiments should be designed to meet the four validity requirements.	Y	S	No individual activity provided high levels of statistical power and correlation to operations, so an iterative process was employed.
a	Ability to employ the new capability	Y	Y	Wargames were conducted to develop TTPs for force elements, and studies were used to identify the stressful scenario, finally the TTPs were employed in the field.
b	Ability to detect change	Y	Y	The development of stressful scenarios ensured that changes in performance, if present, could be measured, and batch size allowed statistical analysis.
c	Ability to isolate the reason for change	Y	Y	The selection of constructive simulation allowed force elements to be changed, the result being valid if tactics were relevant, which was confirmed through wargaming.
d	Ability to relate results to actual operations	Y	Y	Results were related to actual operations through exercise observations and historical analysis.
4	Defense experiments should be integrated into a coherent campaign of activities to maximize their utility.	Y	Y	An iterative campaign of wargaming, simulation and field trials, integrated with studies and operations research was designed. Resource allocation in the campaign was biased toward studies (in particular the A21 review) and field trials (in particular the final exercise).
5	An iterative process of problem formulation, analysis and experimentation is critical to accumulate knowledge and validity within a campaign.	Y	Y	Biannual reviews with all stakeholders were conducted to endorse insights and redirect the campaign and iterations of field observations. Wargames and simulation built the evidence for force structure changes.
6	Campaigns should be designed to integrate all three scientific methods of knowledge generation (studies, observations and experiments).	Y	Y	An iterative campaign of wargaming and constructive simulation (experimental) and field trials and historical studies (precise observations), integrated with operations research and studies (rational-deductive).

#	GUIDEx Principle	Relevant	Addressed	How was it addressed?
7	Multiple methods are necessary within a campaign in order to accumulate validity across the four requirements.	Y	Y	Constructive simulations were conducted with CAEn and CASTFOREM, and key activities were replicated in field events. Unit exercises, field trials, wargames and constructive simulations were used in iterative cycles to build confidence. Key factors such as vegetation effects were modeled separately and integrated into the campaign. Strict pre-exercise and post-exercise modeling was not conducted.
8	Human variability in defense experimentation requires additional experiment design considerations.	Y	N	Large command post exercises were conducted and observed then combined with seminars to determine the human requirements for C2. There were no attempts to develop statistical power through manipulating the assignment of the players. Therefore while the CPXs and field exercises were realistic, it was difficult to isolate the reasons for change or mitigate the impact of human variance.
9	Defense experiments conducted during collective training and OT&E require additional experiment design considerations.	Y	Y	Units developed their own issue lists and collated lessons and observations from all their field activities. Data logging was conducted during the major field exercise to capture critical special and temporal data.
10	Appropriate exploitation of M&S is critical to successful experimentation.	Y	Y	Wargaming and simulation were the critical adjuncts to field exercises, and were the best force on force representation available during the trials due to a lack of weapon simulators.
11	An effective experiment control regime is essential to successful experimentation.	Y	S	The wargames were designed and controlled to explore TTP's for combinations of force elements to take the "best" Red and Blue options into constructive simulation. The constructive simulations were tightly controlled to ensure cause-and-effect could be clearly identified. The field exercises were not controlled, but carefully observed to identify discrepancies and consistencies with the simulation results.
12	A successful experiment depends upon a comprehensive data analysis and collection plan.	Y	S	Two tiers of planning were used, an over plan linking specific events (the campaign plan) and detailed plans for each activity. More effort was applied to the data analysis and collection plans for the field exercises due to fleeting opportunities.
13	Defense experiment design must consider relevant ethical, environmental, political, multinational, and security issues.	Y	S	Normal exercise planning and safety procedures were followed.

#	GUIDEx Principle	Relevant	Addressed	How was it addressed?
14	Frequent communication with stakeholders is critical to successful experimentation.	Y	Y	Trial results were reported semi-annually, with major progress reports annually. The final report was then presented to government.

Table 14 Relation of CS5 to GUIDEx Principles

Case Study 6. The Peregrine Series: a Campaign Approach to Doctrine and TTP Development

6.1 Background

The Australian Army introduced its new armed reconnaissance helicopter (ARH) capability commencing in 2005. The Army's vision for ARH employment is to operate in Troops of two aircraft, as members of combined arms teams at company and battalion levels. The Australian environment, context and concepts for employment are unique and call for original development of the doctrine and TTPs as well as adaptation from other Army's that already have an attack helicopter capability. Synthetic environment experimentation was identified as a primary means to develop and assess doctrine and TTPs in advance of acquisition, with further testing and refinement achieved through other methods (*e.g.*, constructive simulation)—according to GUIDEx Principle 1.

6.2 Aim and Hypothesis

The Peregrine Series commenced in late 2002, and was conceived as a series of experiments using human-in-the loop simulation (*i.e.*, synthetic environment) to inform doctrine and TTP development in advance of individual and collective training programs which was planned to commence in mid-2005. It focuses on the roles and functions of the ARH, at the system-of-systems level, as part of a combined arms team. The series is an example of using experimentation for the ***design*** of a system component (in this case the TTP's) rather than a causal relationship testing campaign trying to prove or disprove some capability development related decision. The design is done through posing a number of questions (that can be worked into a hypothesis), and by allowing the other system elements to evolve through human immersion into the problem.

The notion of a campaign as a multi-methodology approach, designed to accumulate validity and mitigate the weaknesses of individual approaches, is described in Principles 4 to 6. Figure 57 is adapted from the CCRP COBPE (2002). It illustrates the concept that the top-level question (which need not necessarily be at the whole-of-force level) is the starting point for broad, exploratory, holistic appreciation of the problem space. Such experiments (or studies) identify the first-order critical issues which spin-off further experiments or studies at relatively finer scale and "fidelity." When the more detailed issues are better understood, a re-synthesis should occur to build the concepts back up toward the top-level question (although, in reality, this will most likely occur across multiple, inter-connected campaigns).

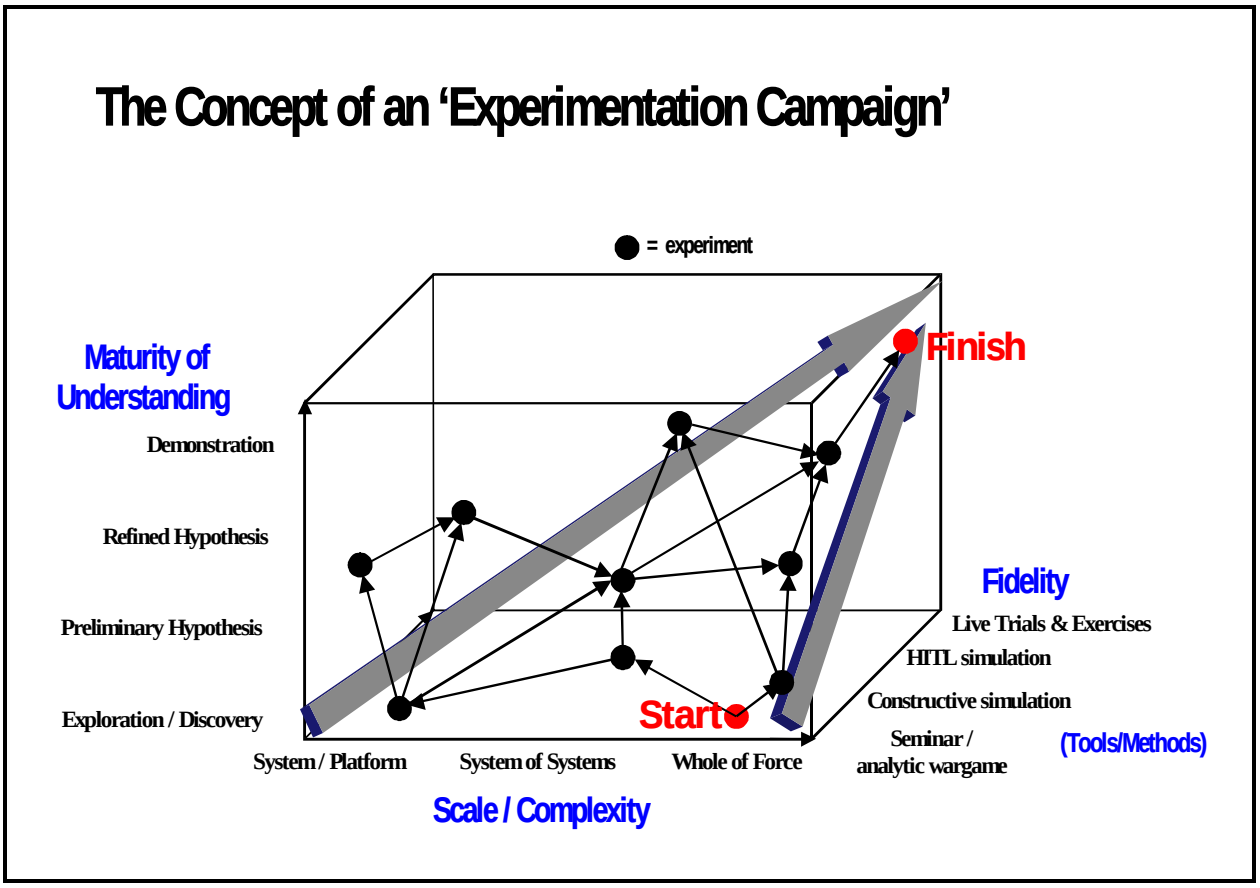


Figure 57 - The progress of a campaign

In this way, the Peregrine series, which addresses system-of-systems level issues using human-in-the-loop simulation methods, forms the backbone of the Peregrine campaign (in accordance with Principle 3 of GUIDEx). It is connected with (is informed by and informs) events within the Army Experimental Framework which address broader, whole-of-force concepts and force structure questions for Army; and similarly incorporates events at more detailed systems and sub-systems levels for understanding weapons, sensor and platform capabilities, and human factors questions. Each of these elements of the campaign is further linked to observational data (reviews of historical data and lessons learned from recent operations, *i.e.*, Afghanistan and Iraq) and studies, including operations research, which addresses particular sub-questions, or informs experiment design (in accordance with GUIDEx Principle 6).

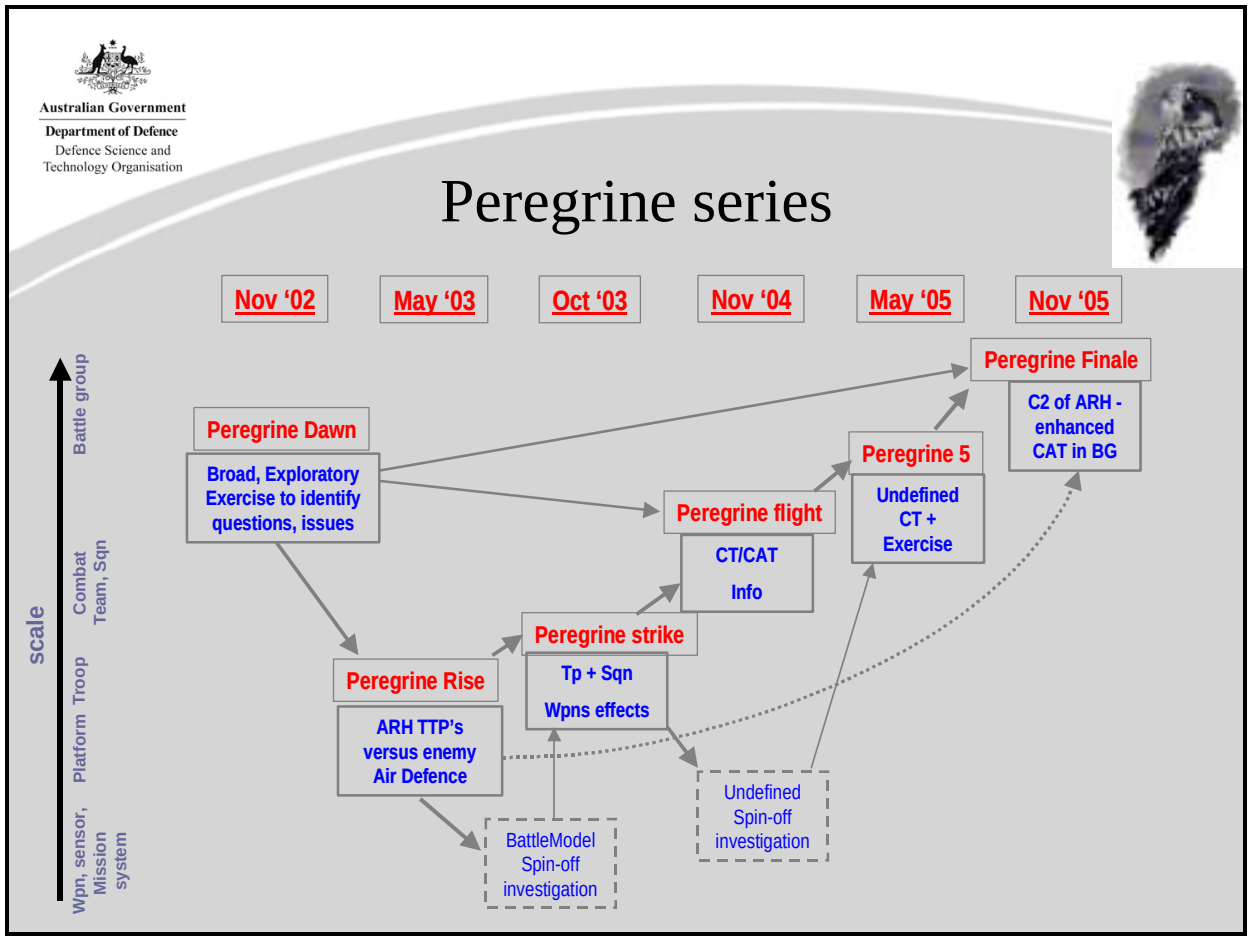


Figure 58 The Peregrine Series

As this Case Study was being written, the Peregrine Series was approaching its fourth major experiment (Nov. 2004), as shown in Figure 58. *Peregrine Dawn* addressed the client's top-level question for the battalion sized (battle group), combined arms team. Even though the experiment employed immature methods and simulation systems, those systems were adequate to identify the critical issues requiring further investigation in: *Peregrine Rise*, which focused on doctrine and TTPs for ARH in a layered air defense environment; and *Peregrine Strike*, which informed weapons employment and effectiveness issues at the single ARH Troop level (*i.e.*, one pair of helicopters). With the benefit of better understanding of Troop capabilities and methods, *Peregrine Flight* (November 2004) was the first attempt to integrate those capabilities into a Company-sized combined arms team (a Combat Team) to investigate the information needs and processes of team members.

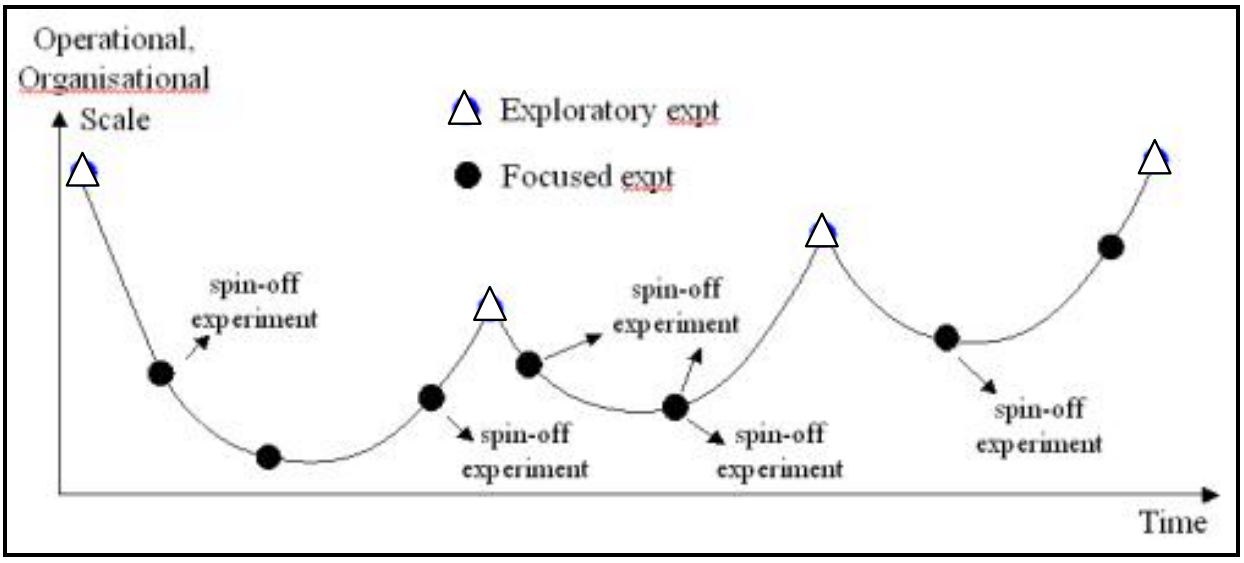


Figure 59 Exploratory and focused experiments in a campaign

Experience in the Peregrine Series suggests that as the primary critical issues are addressed in focused experiments, a re-synthesis or integrated, exploratory experiment is required at a higher operational, or organizational scale to reformulate the problem (re-visit problem definition) and identify critical issues at the next order (addressing GUIDEx Principle 5). The first cycle addresses first-order issues arising from the exploratory exercise. A subsequent exploratory exercise makes use of improved understanding and re-synthesizes elements of the problem back toward the higher-level. It then exposes secondary issues, which are further addressed in focused experiments. This looping process (illustrated in Figure 59) iterates back up toward the top-level problem, accumulating knowledge gained along the journey, such that the quality of insights and conclusions progressively improves through each cycle.

6.3 Results

The outcomes of the campaign to date relate mainly to the change in culture required by the existing Reconnaissance pilots toward a more aggressive stance for operating an armed helicopter. This in turn has resulted in the revision of a significant number of the draft standard operating procedures for ARH employment and the commencement of development of the doctrine publications for the operation of the ARH in the combined arms team. Other findings provide insights as to the degree of teaming required between air and ground elements in order to maximize the overall combat team effectiveness.

6.4 Lessons Learned

At the time of writing, the campaign is not complete and work is on-going to design the next experiment in the sequence. However it is clear that the approach of conducting broader exploratory activities, followed by a sequence of focused activities is paying

dividends in assisting the Army to understand how best to employ the new ARH capability. The broad experiments are by their very nature operating at a lower level of internal validity (lower fidelity representation of many systems) than the focused experiments. However they cover a broader spectrum of military tasks in a realistic environment, and hence have a higher degree of external validity than the focused activities. As a result, the combination of the activities in a carefully designed sequence allows the campaign to build both internal and external validity as the series continues. This assists the military stakeholder in understanding the objective analytical findings within the context of their experiential, subjective lessons learned.

On the negative side, the campaign has shown that, despite the best intentions of the designers, it is very hard to plan much beyond the next activity. This has been the result of a number of factors, including the difficulty in getting long-term direction from stakeholders, the problem of not knowing what the experiment after next will focus on until the current experiment is complete, and also the huge amount of personnel resources required just designing the current experiment, preventing any consideration of future activities. Using broad exploratory events as part of the problem definition process for future focused activities makes it very difficult to plan the focused experiments until the analysis from the broad experiments is complete.

The lessons learned in terms of the 14 GUIDEx Principles are laid out in the following table, which answers the following questions:

1. Which of these 14 Principles are appropriate for this Case Study?

   The answers fall in one of these three categories: yes (Y), this Principle was or somewhat (S) was appropriate, or not at all (N).

2. Which ones were addressed during the Case Study?

   The answers fall in one of these three categories: yes (Y), this Principle was or somewhat (S) was addressed, or not at all (N).

3. How was it addressed during the Case Study?

#	GUIDEx Principle	Relevant	Addressed	How was it addressed?
1	Defense experiments are uniquely suited to investigate the cause-and-effect relationships underlying capability development.	Y	Y	The Peregrine campaign addresses ARH doctrine and TTP development and assessment in advance of acquisition.
2	Designing effective experiments requires an understanding of the logic of experimentation.	N/A	N/A	Not applicable at the campaign level

CS6 Armed Reconnaissance Helicopter

#	GUIDEx Principle	Relevant	Addressed	How was it addressed?
3	Defense experiments should be designed to meet the four validity requirements.	N/A	N/A	Not applicable at the campaign level
a	Ability to employ the new capability	N/A	N/A	
b	Ability to detect change	N/A	N/A	
c	Ability to isolate the reason for change	N/A	N/A	
d	Ability to relate results to actual operations	N/A	N/A	
4	Defense experiments should be integrated into a coherent campaign of activities to maximize their utility.	Y	Y	The Peregrine Series lies within a broader campaign involving a series of experiments, studies and observations.
5	An iterative process of problem formulation, analysis and experimentation is critical to accumulate knowledge and validity within a campaign.	Y	Y	Through an iterative (looped) process of exploratory and focused experiments, within which the problem is re-formulated based on acquired knowledge.
6	Campaigns should be designed to integrate all three scientific methods of knowledge generation (studies, observations and experiments).	Y	Y	The broad Peregrine campaign does include experiments, studies and observations.
7	Multiple methods are necessary within a campaign in order to accumulate validity across the four requirements.	Y	Y	The Peregrine Series is part of a larger campaign that uses all of the methods for individual activities. The doctrine and TTPs that emerge from individual activities are passed forward for validation in constructive simulations. In addition, in some experiments within the series, some pre-modeling, using simplistic spreadsheet models was undertaken in order to inform the design process.

#	GUIDEx Principle	Relevant	Addressed	How was it addressed?
8	Human variability in defense experimentation requires additional experiment design considerations.	Y	Y	Army aviation (in Australia) is a relatively small (busy) community. Availability of aviators to play roles in experimentation is therefore limited. Nevertheless, the Peregrine Series enjoys strong support from that community and the crews are played by committed pilots who will soon own the real ARH. Rather than rotating the players to mitigate the effects of their learning, the campaign tends to use many of the same people for multiple experiments. In order to limit and identify the effects of their learning on our interpretations, the design includes substantial pre-experiment training so that players come into experiments with a strong understanding of the peculiarities of the systems' representations. The experiments also consider trial sequencing so as not to bias observed outcomes with additional consistent learning components.
9	Defense experiments conducted during collective training and OT&E require additional experiment design considerations.	Y	S	Collective training for Australian ARH may commence in mid-2006. Nevertheless, the Peregrine Experimentation Series is anticipated to continue beyond the collective training period and is intended to integrate with, and take advantage of such opportunities as and when appropriate.
10	Appropriate exploitation of M&S is critical to successful experimentation.	Y	Y	As indicated above, the Peregrine Series integrates with a broader campaign that uses other methods and techniques to address components of the high-level "introduction into service" problem for ARH. Each major experiment or event in the campaign is informed by prior steps, and by models and analysis that feed the iteration loop involving problem definition, experiment design and analysis planning for each experiment. This process is also informed by a higher level methodology study underway, alongside and independently from Peregrine, to develop more objective methods for breaking problems down and apportioning elements to those techniques best suited to address those elements, and to deliver the kind of outputs most needed. This work is described in P10 and will also be published elsewhere.
11	An effective experiment control regime is essential to successful experimentation.	N/A	N/A	Not applicable at the campaign level
12	A successful experiment depends upon a comprehensive data analysis and collection plan.	N/A	N/A	Not applicable at the campaign level

#	GUIDEx Principle	Relevant	Addressed	How was it addressed?
13	Defense experiment design must consider relevant ethical, environmental, political, multinational, and security issues.	N/A	N/A	Not applicable at the campaign level
14	Frequent communication with stakeholders is critical to successful experimentation.	Y	S	Involvement of the principal client and all relevant stakeholders must remain paramount and ongoing at the campaign level. It is rarely the case that an experiment will deliver all that a client wishes, so that as the design and development proceed, it is crucial to keep the decisionmakers in the loop and ensure that there are no surprises to any members of the stakeholder groups as the planning, development and conduct phases proceed. These conversations are progressed, in the Peregrine Series by various visits to the client organization, to the experimentation organization, and to other groups, as required, in order to inform critical components (*e.g.*, practitioner units to design relevant scenarios). These conversations continue, supplemented by email, right up to the experiment itself. However, as always, there could be better interaction at the problem definition level in order to get clearer direction for where the future activities should go, rather than simply focusing on the next experiment. This would assist in developing a more detailed, longer-term plan than is currently possible.

Table 15 Relation of CS6 to GUIDEx Principles

Case Study 7. Multinational Experiment Three (MNE 3)

7.1 Background

Multinational Experiment 3 (MNE 3) was the third event in a series of United States Joint Forces Command (USJFCOM) multinational experiments. MNE 3 was a process-refinement experiment whose goal was to build on the lessons learned from Multinational Limited Objective Experiments I and II, and to continue exploring concepts and supporting tools for effects-based planning (EBP). Results will assist the development of future processes, organizations, and technologies at the operational and joint task force level of command. Additionally, MNE 3 provided the participating nations an opportunity to examine issues associated with operational net assessment (ONA), a coalition interagency coordination group (CIACG), coalition intelligence, surveillance, and reconnaissance (CISR), multinational information sharing (MNIS), logistics, coalition based health services support (CBHSS), information operations (IO), and knowledge management (KM).[69] The North Atlantic Treaty Organisation (NATO) also examined concepts associated with their NATO Response Force (NRF).

Results of multinational experimentation will support further development of a standing joint force headquarters (SJFHQ) and will provide data for information sharing, multilevel security, and collaborative operational net assessment development to both the NATO Concept Development and Experimentation (CDE) Working Group and to the Multinational Interoperability Council (MIC). MNE 3 participants included Australia, Canada, France, Germany, the United Kingdom, the United States, and NATO. The United States, with USJFCOM, Joint Experimentation (J9) as the executive agent, led this event.

MNE 3 was a worldwide-distributed experiment with USJFCOM and key coalition players situated at the USJFCOM Distributed Continuous Experiment Environment (DCEE) facility located in Suffolk, VA and other coalition players participating from their national experimentation facilities. NATO utilized its Castlegate, Germany facility.

The scenario for MNE 3 was set in 2004 Afghanistan utilizing real-world data and scripted vignettes reflecting possible future developments in the area.

[69] A more detailed explanation of all concepts experimented on may be found within the J9 Knowledge Management Portal. http://www.jfcom.mil/about/abt_j9.htm

7.2 Aim and Hypotheses

Objective 1: Develop and Assess Processes to Support Coalition/NRF EBP[70, 71]

Proposition 1: The application of EBP will improve an operational commander's ability to:

1. broaden the range of effects considered,
2. broaden the range of actions considered,
3. respond in an agile fashion to changing conditions,
4. coordinate actions with multinational military and non-military participants,
5. enable exploitation of military and non-military knowledge, and
6. create a comprehensive effects tasking order (ETO).

Objective Context:

EBP depends on a complex array of processes. These processes involve the integration of many planning efforts into one coordinated endeavor. The concepts that describe these supporting efforts include EBP, ONA, Collaborative Information Environment (CIE), CIACG, CISR, MNIS, and logistics. If these concepts are employed to their full extent, and each contributes necessary information to EBP, the coalition/NRF will be successful in performing EBP.

The measures for each Critical Operational Issue (COI) identified metrics of performance for each concept relative to the six EBP performance requirements.

COIs:

1. Does the MNE 3 implementation of EBP facilitate the operational level of command's ability to address their objectives?
2. What are the critical human and mechanistic processes, dependencies and information flows in EBP?
3. How well do the current logistic planning processes support the construction of a coalition force deployment plan within the EBP process?
4. Does the MNIS concept support effective sharing and exploitation of information for the EBP process in a multinational, virtual and distributive environment?
5. What JISR/CISR process is required to support EBP?
6. To what extent was the CIACG able to coordinate and harmonize operational planning between the coalition military planners and the relevant civilian agencies or departments of their respective governments?
7. How does ONA support the EBP process?
8. What CIE procedures are required to support EBP?

[70] During the planning for and execution of the experiment, multinational effects based planning and supporting processes and organizations were developed and refined.

[71] Assess is defined as evaluating the importance, significance, value, or merit of the processes, organizations and technologies examined in the experiment.

9. Does the Coalition-Based Health Services Support process support EBP?

Objective 2: Develop and Assess Organizations to Support Coalition/NRF EBP

Proposition 2: The EBP organizational design will:

1. enable a managed flow of information,
2. facilitate the generation of knowledge,
3. enhance planning,
4. improve decisionmaking, and
5. create a comprehensive ETO.

Objective Context:

The Coalition Task Force Headquarters (CTFHQ) and NATO Experimental Deployable Joint Task Force Headquarters (XDJTFHQ) organizational structures identified for MNE 3 are based upon the design of cross-functional teams that are connected in a habitual way to distributed experts including non-military government and civilian agencies and coalition partners. The staff is organized to enable the effective flow and integration of information. The elimination of functional stovepipes should reduce coordination time and allow synergistic planning and execution. This fluid movement of information, people, and machine interfaces is both a challenge and an opportunity for commanders. Managed properly, this organizational structure can produce better decisions faster. Ultimately, the ETO should produce the desired effects when implemented, and better than current planning practices.

The measures for each COI identified if and how the organizational structures implemented support the EBP and supporting processes.[72, 73]

COIs:

1. What organizational structure is required for EBP?
2. What behaviors and competencies are required for EBP?

Objective 3: Identify Technology Requirements to Support Coalition/NRF EBP.

Proposition 3: Technology will augment the human ability to conduct EBP through a suite of tools Objective Context.

Objective Context:

Technologies support the human ability to communicate and collect, process and display information from diverse sources to conduct effects based planning in a CTFHQ/NRFHQ.

[72] There is a relationship between structure and process (structure implies process).
[73] There is a relationship between the information environment and technology.

The measures for each COI identified if and how the technologies implemented support the EBP and supporting processes, as well as identify technology functionality shortfalls and requirements.

COI:

1. What functional requirements are necessary to conduct EBP within a coalition/NRF environment?

7.3 Background

Type of experiment, series or campaign: MNE 3 was an experiment using human-in-the-loop simulation with additional elements of an analytic wargame. Although MNE 3 was a single trial experiment, it is just one piece of a larger experimentation plan conducted by USJFCOM concerning the concepts examined.

7.4 Treatment

The treatment (independent variable) examined in this experiment was an operational level EBP process. The experimental unit was a subset of a functionally organized CTF headquarters staff. The effect (dependent variable) was an assessment of whether the EBP process as played had the potential to provide the capabilities described in the proposition statements, and recommend changes to that process to improve its effectiveness.

As several nations had already developed their own variations of an EBP concept prior to MNE 3, it was necessary to write a new version of the EBP concept specifically for MNE 3. This version was based on features from the national EBP concepts and provided a common baseline for MNE 3.

To examine the viability of and the procedures required for implementing EBP, certain overarching and supporting concepts were required to accurately depict the planning environment. These included: ONA, CIE, CIACG, CISR, MNIS, Logistics, CBHSS, IO, and KM.

The United States chose to implement a CTFHQ based on the SJFHQ organizational construct. The experiment was designed so that NATO and Allied Command Transformation (ACT) would examine and implement the EBP process in parallel with the CTFHQ.

For MNE 3 a CIE was developed. The CIE enabled collaboration at will between selected groups of individuals or organizations. The CIE was defined as the aggregation of infrastructure (hardware, software, and communications links), capabilities (synchronous and asynchronous), people, procedures, and information for the common purpose of creating and sharing data, information, and knowledge necessary to plan, execute, and assess coalition/NRF operations.

Elements of the CIE deployed for MNE 3 included:

1. a secure, reliable network built on the CFBL network using virtual private network (VPN) technology,
2. the InfoWorkSpace (IWS) collaborative tool,
3. a Voice Over IP (VOIP) telephony,
4. a web portal,
5. a situation awareness through the WEBCOP, and
6. the ONA Database

The MNE 3 EBP processes and the use of the CIE were introduced to participating nations through on-site training in the partner nations, collaborative training, and workshops prior to experiment execution. To facilitate experiment execution, a week of training (Week 0) was followed by two weeks of live play. A single scenario and vignette were used to stimulate the CTFHQ and XDJTFHQ EBP process. Both the NRF XDJTFHQ and the CTFHQ worked through the EBP process simultaneously.

The assessment team was organized to support the analysis functions of the experiment, which included: assessment planning, data collection, data analyses, and results reporting. All partner-nation analysts were integrated into the USJFCOM analysis team to contribute to the assessment process, from planning to reporting. Assessment focused on two primary areas:

1. a qualitative comparison of the conceptual and applied models of the EBP and supporting processes, organizations, and technologies; and
2. non-intrusive observations and participant perceptions and insights on specific aspects of the EBP and supporting processes, organizations, and technologies.

Conceptual models representing the functional and temporal aspects of EBP and supporting processes, organizations, and technologies were developed using the G2 and C3TRACE[74] process model tools. These models captured internal and external tasks, processes, organization, and communications played during the experiment. The models were developed during the experiment validation Rock Drill and the experiment. Figure 60 depicts the components of the conceptual models.

[74] Command, control, and communications - techniques for the reliable assessment of concept execution (C3TRACE).

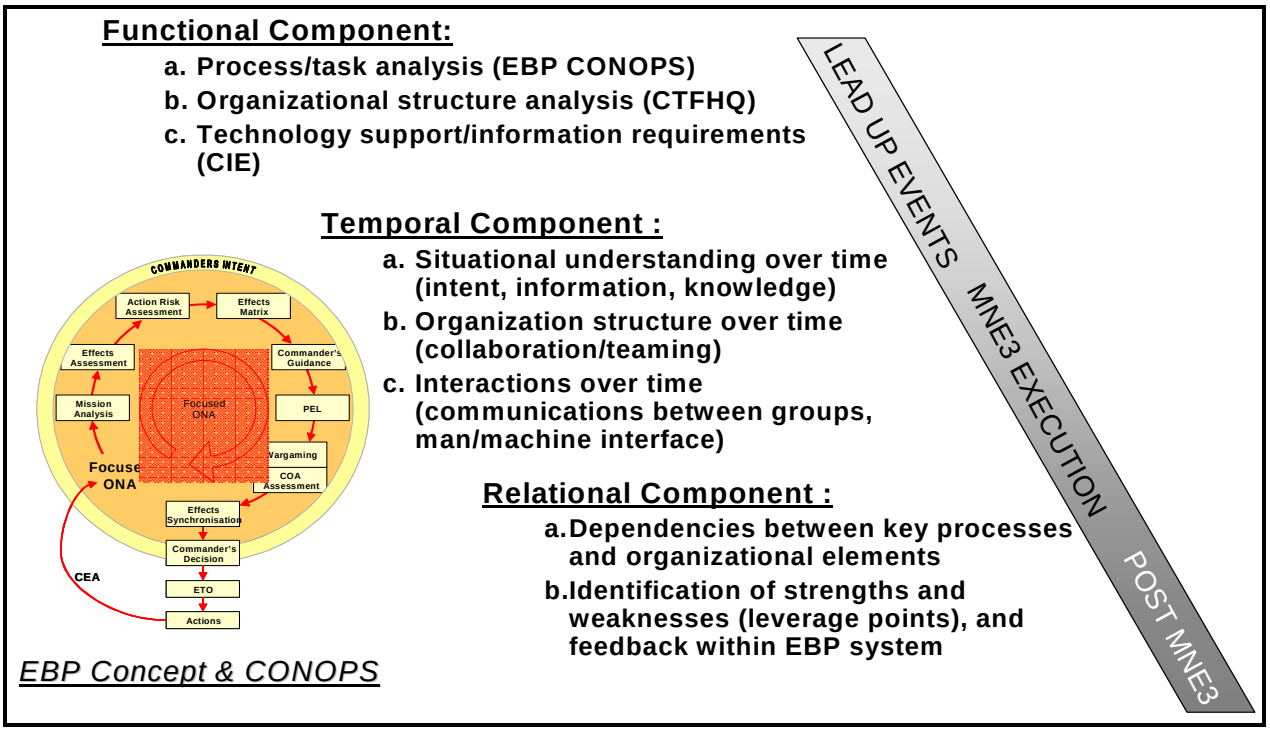


Figure 60 Conceptual model components

The output of these models can then be compared to the physical model of the processes, organizations, and technologies employed by the CTFHQ as well as the XDJTFHQ during experiment execution to build a relational understanding of key processes and organizational elements. These measurements and observations are then used to update conceptual models so as to document the developed EBP process for subsequent experimentation.

The experiment generated qualitative and quantitative data needed to gain insight into the EBP process and supporting processes, organizations, and technologies. Qualitative data include subjective evaluation of events by participants and observers through the use of surveys, SCD insights and observations, participant seminars, daily end-of-the-day reviews ("hotwashes") and end-of-week after-action reviews (AARs). Quantitative data sets are objective measurements of events from nonjudgmental observers or instrumentation such as command, control, communications, computers, and intelligence (C4I) system usage. This data is critical to finding inefficiencies in process, organization, and technology. Qualitative data were analyzed for trends and commonalities and for differences in rating metrics. Quantitative data were also analyzed for time-and-event-frequencies associated with the EBP process, as well as task, communications, and workload analyses.

Objective 1 Methodology. The analytical design to support the process objective was divided into effectiveness and performance of the EBP process. The effectiveness COI was analyzed through synthesis of survey responses and participant insights. The results of the effectiveness COI provides a high-level assessment of the EBP process and indicators of areas for further development. The aim of the analysis under the

performance COI was to identify specific issues that impacted execution of the process. The performance COI was analyzed through synthesis of survey responses, participant insights, numerical C4I data, and direct observations.

Objective 2 Methodology. Qualitative and quantitative data were used to assess the CTFHQ and XDJTFHQ organizational structures, as well as the human behaviors and competencies required to conduct EBP. The ultimate aim was to identify critical organizational constructs and to understand organizational relationships since these are vital for effective C2 of joint, allied, or coalition task forces.

Objective 3 Methodology. Qualitative and quantitative data were used to assess MNE 3 implemented technologies, and to identify functionality requirements for EBP. The aim was to identify technology requirements to support coalition and NRF EBP.

Furthermore, an experiment analysis workshop was convened to enable all analysts to contribute their inputs to the final report. Partners discussed their insights into the objectives and concepts for which they had lead analysis responsibility, as well as proposed experiment findings.

7.5 Broad Results

Key findings from MNE 3 were:

1. The effects-based planning concept has the potential to make the coalition task force and NATO Response Force more effective instruments of power. However, the effects-based planning concept as developed for MNE 3 is not operationally mature and requires further refinement.
2. Players stated that the best features of the effects-based planning process were:
    a. it forced military planners to think in terms of effects, which expanded alternative ways to achieve objectives beyond military actions, and
    b. collaboration brought out the best ideas from a collective thought process.
3. Players stated the most difficult parts of the MNE 3 effects-based planning process were:
    a. the complexity of the process inhibiting thought and analysis,
    b. confusing terminology, and
    c. lack of an integrated tool suite.
4. There is a need to create a coalition logistics structure and plan as a coalition, not as a group of individual nations.
5. The Coalition Interagency Coordination Group brings a valuable civilian perspective to military planners, the coalition task force staff, and the command group that is essential to an effective effects-based planning process.
6. Contributions from subject matter experts such as Coalition Interagency Coordination Group, medical, and information operations need to be integrated in the operational net assessment.
7. The staff organization should be driven by effects-based planning process requirements.
8. Leadership in a coalition collaborative information environment requires different skills than those required in today's command and control environment.

9. Effects-based planning calls for an integrated suite of tools to support distributed collaborative planning as well as tools specifically designed to support the effects-based planning process.

A significant element of defense experimentation is the participation of senior concept developers - a select group of former general and flag officers and civilian equivalents. These individuals participate in a variety of activities as a source of experience and knowledge that contributes to the growing understanding of concepts being examined during the experiment. Senior concept developers identified three overarching, emergent themes from the experiment:

1. The effects-based approach to coalition planning in a collaborative information environment is essential and challenging—but doable. It poses new and significant interoperability challenges with promising opportunities.
2. The observations gleaned from this experiment have greater value and credibility because the effects-based planning that was accomplished used a real-world scenario.
3. The use of an Afghanistan scenario emphasized that stability operations are inherently multinational and interagency and require a common doctrine.

7.6 Discussion Relative to GUIDEx

Answers to questions relating this Case Study to the GUIDEx Principles are summarized in the table below, which answers the following questions:

1. Which of these 14 Principles are appropriate for this Case Study?

   The answers fall in one of these three categories: yes (Y), this Principle was or somewhat (S) was appropriate, or not at all (N).

2. Which ones were addressed during the Case Study?

   The answers fall in one of these three categories: yes (Y), this Principle was or somewhat (S) was addressed, or not at all (N).

3. How was it addressed during the Case Study?

#	GUIDEx Principle	Relevant	Addressed	How was it addressed?
1	Defense experiments are uniquely suited to investigate the cause-and-effect relationships underlying capability development.	Y	S	Ability to isolate cause-and-effect was limited by the number of confounding variables. Use of process/organization/technology matrix assisted capability development decisions.

CS7 Multinational Experiment 3 (MNE 3)

#	GUIDEx Principle	Relevant	Addressed	How was it addressed?
2	Designing effective experiments requires an understanding of the logic of experimentation.	Y	S	The 21 threats to valid experimentation were taken into account in experiment design. Compromises in control of variability were made to accommodate desires to investigate changes in the EBP process during the experiment. The event incorporated aspects of demonstration and discovery. There were stated propositions in place of a formal hypothesis. The analysis was effective however.
3	Defense experiments should be designed to meet the four validity requirements.			See below.
a	Ability to employ the new capability	Y	S	Players did not fully employ the new concept: Participant training and knowledge of the concepts was less than envisaged upon fielding of the concepts. Training and rehearsal intended to occur during the experiment validation event (Rock Drill) did not happen as planned. Instead the time was used to complete the development of tactics, techniques, and procedures (TTP) for the process. Disconnects and gaps between steps in the planning process were not adequately resolved prior to the experiment. It was unclear how the supporting concepts should be integrated into the planning process. Players had no prior exposure to the EBP planning tool before Week 0.
b	Ability to detect change	Y	Y	Player perceptions of the strengths and weaknesses of the process, organization and technology were fairly consistent across experiment sites and nationalities. Players came to the experiment with different perceptions of what a command-led process should be.
c	Ability to isolate the reason for change	Y	S	Ability to isolate cause-and-effect was limited. Because of training difficulty and process immaturity the process and organization evolved over the course of the experiment.
d	Ability to relate results to actual operations	Y	Y	Real-world scenario and robust headquarters staffs enhanced ability to relate results to operations. Component and higher-level command play was limited to White cell responses.

CS7 Multinational Experiment 3 (MNE 3)

#	GUIDEx Principle	Relevant	Addressed	How was it addressed?
4	Defense experiments should be integrated into a coherent campaign of activities to maximize their utility.	Y	S	USJFCOM MNE series supports both Concept Development and Prototype pathways. This Principle was well addressed in the US but not by all of the coalition members. It was a problem for nations like Canada where some concept development had been done, but little preliminary experimentation.
5	An iterative process of problem formulation, analysis and experimentation is critical to accumulate knowledge and validity within a campaign.	Y	S	For US, a series of national and multinational experiment events contributed to the overall examination of the concepts. Only US applied an iterative process, the UK did so partially, but it was baptism by fire for the rest of the nations.
6	Campaigns should be designed to integrate all three scientific methods of knowledge generation (studies, observations and experiments).	Y	Y	Results of all three methods used in EBP concept formulation and design.
7	Multiple methods are necessary within a campaign in order to accumulate validity across the four requirements.	Y	S	Use of human-in-the-loop structure enhanced ability to relate the process to real-world operations but increased variability. MN LOE II and MNE 3 used the same approach, but other experiments outside the MNE series used alternative approaches to examine the concepts. Original intent was to employ M-E-M using the G2 process model. However, process immaturity precluded completion of the model prior to experiment execution. SJFHQ model in G2 had only limited use after the experiment due to process being insufficiently defined.
8	Human variability in defense experimentation requires additional experiment design considerations.	Y	S	Compromises in control of human variability were made to accommodate desires to investigate changes in the EBP process during the experiment. The analysis looked for consensus in the player reactions and observations.
9	Defense experiments conducted during collective training and OT&E require additional experiment design considerations.	N	N	N/A
10	Appropriate exploitation of M&S is critical to successful experimentation.	S	N	M&S support was not required for the conduct of MNE 3. Some M&S efforts were used as mentioned in #8, however more could have been applied for experiment design and execution.

#	GUIDEx Principle	Relevant	Addressed	How was it addressed?
11	An effective experiment control regime is essential to successful experimentation.	Y	S	Differences of opinion on the methods and limitations to be imposed by the experiment controllers emerged across partner nations during experiment execution.
12	A successful experiment depends upon a comprehensive data analysis and collection plan.	Y	Y	All partner analyst teams contributed to analysis methodology, plan, collection, assessment and reporting.
13	Defense experiment design must consider relevant ethical, environmental, political, multinational, and security issues.	Y	S	Utilized multinational directors, controllers, analysts and observers to ensure all factors considered, although some were determined to be not relevant.
14	Frequent communication with stakeholders is critical to successful experimentation.	Y	Y	Frequent planning briefs to JFCOM and partner nation experiment management. Senior Leader Seminar presented initial results to US and multinational leaders. Experiment reports distributed to comprehensive distribution list.

Table 16 Relation of CS7 to GUIDEx Principles

All but one of the 14 Principles were applicable in MNE 3. In the case of Principle 10, it was only listed as “Somewhat” applicable as M&S was not needed for stimulation in the conduct of the experiment. M&S would have been useful, however, in the application of the M-E-M paradigm (see Principle 7). Had it been possible to explore EBP and the organization of the SJFHQ through a complete process model, the problems identified under Principle 3 (problems with process and the integration of supporting concepts) might have been avoided. The same process model would also have contributed to the analysis. This is not to imply that there were any problems with the analysis, it would only have been enhanced by additional M&S.

That thirteen out of fourteen Principles were addressed in this experiment is a good indication that the overall experimentation process, from design and planning through to analysis, was sound and thorough. That many of the Principles were only addressed “Somewhat” was largely due to the difficulty in planning and conducting a very large multinational event. One aspect of this is the differences in nation’s preparedness (note Principles 4 and 5). As noted under Principle 2, for some this was an exploratory activity while for the US it was clearly a process refinement event. Regardless, or perhaps because of this, many valuable lessons were learned and have been put to use in the writing of this guide, Principles 11 and 12 in particular.

Case Study 8. Improved Instruments Increase Campaign Values

8.1 Background

The last four decades showed substantial improvements to the capabilities of sensors and weapon systems that greatly increased the area in which naval commanders are interested in or over which they have responsibility. For example missiles that require a Force Over-the-horizon Track Coordinator (FOTC) concept to plan targeting using sensors from other platforms since weapon extended ranges were in excess of own platform's sensor detection capabilities. Because early systems used to monitor the wide-area naval tactical picture (WAP) were found inadequate to provide the information needed in a timely and effective fashion, improved alternatives were sought. Systems tested evolved into the Global Command and Control System (GCCS) that greatly influenced the evolution of the Global Information Grid (GIG) and Network Centric Warfare (NCW) concepts.

Consequently, the AUSCANNZUKUS Organization[75], which is responsible for identifying and solving interoperability problems among allied maritime units, initiated in 1991 a work program to investigate the Management of Organic and Non-organic Information in a Maritime Environment (MONIME) and created a multidisciplinary *Ad-hoc* Working Group (AHWG) for this program. To address the broad spectrum of issues at stake the AHWG multidisciplinary team was selected from military, requirements, engineering, OA/OR and scientific communities. Note that no human factor engineers were involved since at that time the organization was focusing on technological challenges. Today we recognize the need to address these issues more holistically and so would include the cognitive and social domains for delivering the desired force multiplier as well as using the GUIDEx integrated analysis and experimentation campaign approach. Nevertheless, the AHWG, after a review of available data from operations and systems came to an agreement with the concerned organization by defining the scope, methodology and deliverables for the MONIME campaign.

Improved instruments: During the course of this campaign a critical activity was initiated by Canada to resolve an outstanding issue regarding the ability to observe or detect the effects on operations of some of the changes (control interventions) in specific MONIME campaign experiments by developing new instruments for experimentation, the model-based-measures (MBMs) [Labbé and Proulx 1998a, 1998b].

[75] Counterpart of the TTCP information exchange agreement for operational and in-development systems, related standards and problems. Member countries: Australia, Canada, New Zealand, United Kingdom and United States.

8.2 Objectives and Hypotheses

The investigation campaign agreed six objectives were:

1. Identify Information: Identification of existing information, and amount of detail, including such aspects as accuracy (positional precision and correct classification) and timeliness, available at various command levels of a force using the Composite Warfare Commander (CWC) structure.
2. Identify Deficiencies: Identification of current differences in the production of a coherent and accurate (accurate means both correct classification and positional granularity) Allied wide-area picture, as developed by Link 11, Joint Operational And Tactical System (JOTS), Flag Data Display System (FDDS), Tomahawk Weapons Control System (TWCS), the RN's Fleet Ocean Surveillance Product (FOSP), the Royal Australian Navy (RAN)'s Maritime Intelligence Support Terminal (MIST) system and other national equipment using the same sources of information.
3. Evaluate C3I Capacity: Determine if current national information and communications systems can support the information exchange requirements (IERs) essential to a task within each command and unit level and, where deficiencies exist, whether projected systems will provide solutions.
4. Improve C3I Capacity: Given existing information systems' capabilities, identify and evaluate systems' changes that will improve systems performance.
5. Interoperability: Identification of the interoperability standards and rules, both technical and procedural, required to ensure effective information management (acquisition, display, analysis, and dissemination of tactical information).
6. Local Procedure: Determine intra-platform requirements for information management; *i.e.*, the most effective way to communicate, filter, tailor and present information.

Some of the hypotheses included the followings:

1. If the channel capacity from the FOTC to the participating units is increased, the observed quality of information received by the participants will increase.
2. If two surface actions groups (SAGs) are provided with different channel capacities (data rate, bandwidth) this difference will be reflected in their relative information quality.
3. If opposing forces used different strategies (*e.g.*, collaborative, *i.e.*, they radiate using radars and radios, versus non-collaborative, *i.e.*, radio silence) this will have an impact on the timeliness and completeness of the picture of hostile contacts.
4. If perfect correlation, association and fusion could happen, the resulting picture will display improvements for most practical cases.
5. If dead reckoning is used, the quality of the tactical picture will not degrade as fast as without this first-order position predictor based on last speed, course and position report.

8.3 Type(s) of Experiment, Series or Campaign

Given that the MONIME campaign was not exclusively designed around experimentation but exploited studies and observations as well, several of the final recommendations for

future systems were based on common sense from experts and correlational evidence[76] ranging from system design invasive testing to system specifications, including changes from one collective training activity to the next one, *e.g.*, the RIMPAC[77] series. The soundness of the recommendations came from the cumulative knowledge acquired from years of instrumented exercises (for adjudication purposes, substantial instrumentation and data collection were performed to construct the ground-truth required by most of the MoPs and MoEs of interest). From NAVNEWS,[78] "...RIMPAC '96, a training exercise involving 300 aircraft, 40 allied warships, and 30,000 combat personnel. One of the largest naval exercises in the world." Also, exercises like RIMPAC involve thousands of non-combat people, various air bases, submarines, headquarters, data analysis and collection centers. The MONIME campaign included studies from RIMPAC series prior to 1992 and observations with proper instrumentation and ground-truth reconstruction from the 1994 and 1996 events. The campaign included several experiments and other observational studies from other venues, *e.g.*, MARCOT 95.[79]

The MONIME analysis methodology evolved from the study of the accumulated information from operations, previous exercises, related standards and analysis reports, and thus, exploited the three scientific methods of knowledge generation with an emphasis on observations during training exercises as well as experimentation using human-in-the-loop simulations, wargaming and constructive simulations. An analysis requirements document [AUSCANNZUKUS MONIME 1992] was prepared to define MoPs and MoEs, and data collection required to support the recommendations for future systems. This document included a matrix of measures-resources to identify gaps or areas where little or no data were available, or expected to be available from planned trials, to support recommendations.

Between 1991 to 1997, the AHWG conducted a series of studies, trials and experiments to collect data to better analyze and characterize WAP (wide-area picture) systems, and develop adequate requirements for future command and control (C2) information management (IM) systems. This campaign concluded with Handbook Five (HB5) guidelines [AUSCANNZUKUS MONIME 1997] to be used in the procurement of national C3I WAP-based systems for the compilation and sharing of accurate WAPs, which in turn was used to develop the "Major NATO Commanders CONOPS for Information Management" in 1998.

Figure 61 describes the interactions among the resources and methodologies identified by the AHWG [Labbé 1997]. For the complex problems under consideration, it was recognized that no tractable analytical solutions or single national resource could provide all the results required to formulate credible recommendations. The global

[76] Melvin M. Mark, Inferring Cause from Passive Observation, in [Cook and Campbell 1979] excerpt, "...to infer causal processes based on observations of concomitancies and sequences as they occur in natural settings, without the advantage of deliberate manipulation and controls to rule out extraneous causal influences." page 295.
[77] Rim of the Pacific countries LIVEXs (RIMPACs).
[78] http://www.chinfo.navy.mil/navpalib/news/navnews/nns97/nns97050.txt
[79] Canadian Second 1995 Maritime Command Operational Training Exercise (MARCOT 95-2).

approach to be proposed combined the advantages of top-down and bottom-up methodologies, reducing the weight of the results on any one of the experiments conducted with national resources:

1. (collective training) National and Coalition Forces LIVEXs and post exercise analyses, *e.g.*, Rim of the Pacific countries LIVEXs (RIMPACs);
2. (HITL) TIMSIMs using the US RESA[80], which combined simulation and existing C3I systems with commanders at different command levels in various types of warfare; and
3. (analytic wargame + constructive simulation) the resources of UK's NISAS[81], which provided simulation and modeling tools.

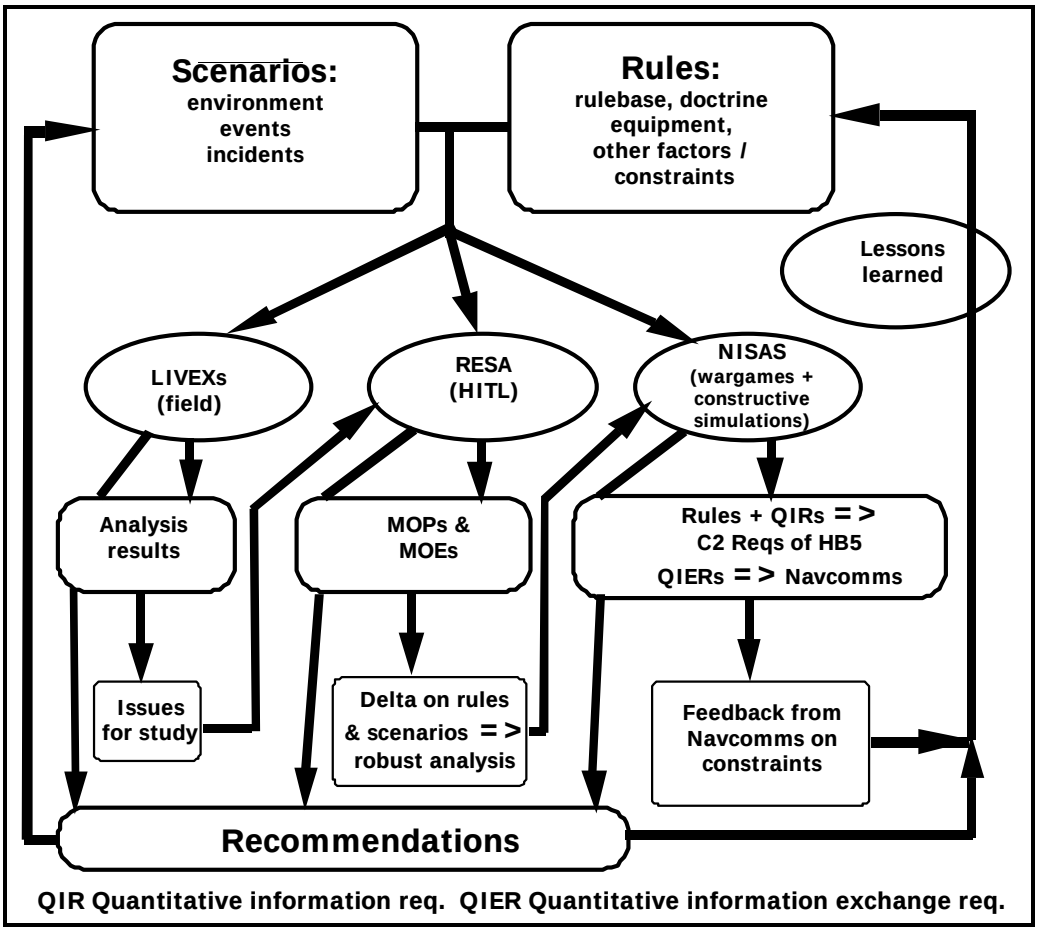


Figure 61 MONIME's methodology: resources and data relationship

Each trial or experiment was designed to increase the level of insight into issues to address, in order to attain the campaign's objectives. Any given experiment focused on issues that could be best tackled using a specific national resource. Each trial raised

[80] TIMSIM, Tactical Information Management Simulation, the name given to AUSCANNZUKUS experiments using the methodology of the US TIMEX, Tactical Information Management Exercise conducted at the Research, Evaluation and Systems Analysis Wargaming Facility (US).
[81] Naval Information Systems Architecture Study (UK).

additional issues and provided new avenues for generating the required knowledge. Consequently, each experiment was designed to facilitate the exploitation of its results in subsequent experiments, using either the same national resource or, more likely, one or more of the other resources, perhaps employing new methodologies.

The NISAS resources and methodology, which were part of a substantial R&D activity conducted in the UK, offered worthy simulation and modeling capabilities, and good internal validity (control and observation opportunity) but less external validity than TIMSIM. NISAS run against a fully (analyst) scripted scenario, whereas TIMSIM ran with a battle-group staff that fought the battle "live" against a pre-scripted Red (opponent) scenario or played Red forces' activities.

NISAS resources had two main components: Command and Control Information Requirement (C2IR) and Information Systems Architecture Tool (ISAT). The C2IR component generates information-exchange requirements without communication or computing constraints (wargame). Its traffic is script-specified, according to information generation and information exchange rules defined by the scenario, by operational procedures and by doctrine. The ISAT component (constructive) imposed restrictions on the information flow provided by the C2IR in order to determine:

1. the COP degradation due to environment-dependent finite channel capacity, or
2. the communications requirements to satisfy the information flow prerequisites of a given scenario and the C2IR load that achieves a given picture quality.

NISAS emphasized analyses within an overall system context with an approach based on the specification of the role of a platform or node and of the scenario within which it was to carry out that role.

The RESA capabilities for HITL TIMEX and TIMSIM experiments are described in CS1.

A subsequent study conducted by Canada revealed that the level of progress via MONIME's methodology was limited by the capacity of available instruments to detect effects on changes of control variables. Consequently, a method for observing a relationship between WAP information quality and mission effectiveness was designed, implemented and exercised on the data generated during the MONIME's trial series and expanded for research purposes beyond the campaign life time. The Canadian study also advocated formalizing a synthesis activity as part of a global methodology to better address the client's interest and be consistent with the scientific method (GUIDEx Principle 4 argues for sufficient synthesis in order for a campaign to be successful).

The MBM method combines a decision model and a set of measures to determine how systems performance and information quality affect mission effectiveness. The decision model selects the tactical information necessary for a postulated action, and the set of measures assesses the value of that information to the postulated decision based on the actual tactical situation (ground-truth) and the outcome of the decision. The new method examines targeting ability as a function of the correctness and timeliness of the information held compared to ground truth at decision time. This method shows a relationship between the quality of the tactical information used and the corresponding

command and control effectiveness in general, and improves the statistical significance of its results by using more systematic and finer grain decision processes than used during military exercises, thereby increasing the sample size from a few tens to several hundreds of decisions.

8.4 Treatments

Apart from the studies of previous exercises and observations of exercises instrumented to collect data relevant to the mandate, the treatments focused onto two systems issues: 1—How much more information would a common database offer to a decisionmaker? 2—How much does the data rate between units affect the completeness of such databases? The initial questions included clauses that could not be addressed with the resources available to MONIME, *e.g.*, what is the minimum information exchange required for a successful task or mission?

In one of the experimental setups, TIMSIM 93, the game umpire and analysts had access to the following data:

1. ground truth for all contacts: positions, speeds, and what they do, *e.g.*, radiate or not,
2. what the sensor suites of the entire asset provided before loss due to alone communications, and
3. what was received by each unit:
    a. the FOTC unit was responsible for combining what was received by its unit from organic and non-organic sources/sensors and was responsible to send (broadcast) the resulting COP onto two different channels (with the same channel capacity for some days and different capacities at specific other times or days) to the two surface action groups (SAG 1 and 2),
    b. the HIT (high interest tracks) broadcast was a subset designed to send the COP over low data rate or low capacity channels,
    c. each SAG was receiving either a full or HIT broadcast on its particular channel and had access to own sensor data,
    d. a fourth unit, the Tomahawk Weapons Control System (TWCS), did a similar data fusion as the FOTC using the same input data but its operators were provided with different rules and training: its database was to be used for Tomahawk targeting and its data were not sent to the SAGs.

Furthermore, data from the MONIME's series were replayed using MBMs for a large number of hypotheses associated with the impact of systems architecture and procedure changes on mission effectiveness [Labbé and Proulx 2000]. An encompassing definition of a MBM follows:

1. A MBM is a measure in which a particular decisionmaker (DM) has been removed from the command and control loop in order to assess the value of a set of MoPs for certain MoEs, systematically by simulation. Since several DMs may influence a function, they are removed individually, one at a time.
2. MBMs replace the complex, man-in-the-loop decision process with simplified models.

3. All staff other than the decisionmaker for the function under study is included in the system assessment.
4. The simulation models link MoPs to MoEs by evaluating the results of actions, based on ground truth.

Specifically, the reported MBMs are defined for over-the-horizon targeting (OTH-T). Such MBMs assess the value of the information made available to a commander by examining each tactical report of track data that meets a particular set of engagement conditions. Location, systems and temporal data are used to establish the engagement parameters and scenarios. Outcomes subsequent to decisions are assessed using both decision-process model definitions and algorithms that include hit-probability calculations. The measures assign reward values that take into account the allegiances of contacts in the interception area and a utility cost for firing a missile.

Using MBMs as a yardstick based on OTH-T effectiveness, various potential changes to the architecture and procedures used in Coalition exercises that might improve the timeliness and accuracy of the information made available to decisionmakers at time of decision (an MoP) are assessed in terms of their impact on OTH-T potential success rates (an MoE). For the reported MBMs, information processing includes sensor data processing, data fusion, situation assessment, weapon pairing, action planning and other deliberative processes that take place before sending the engagement data to the shooter. The information exchange concerns the geographical distribution of the required engagement data from an information-processing node to a shooter. Updated information is used during weapon deployment until final interception or success is confirmed. Resource optimization would benefit from decision support based on OTH-T MBM characteristic curves and the critical age of information required for a given mission success rate.

8.5 Broad Results

Results from this campaign were commended by the AUSCANNZUKUS Organization and as indicated in the Background, it resulted in HB5 recommendations for national procurement of information management systems and related CONOPS for maritime coalition operations.

To reach their full potential, the results of this campaign would have greatly benefited from a more persistent campaign. Because of this lack of persistence, there are issues noted during this campaign that are still not properly addressed today ten years later, as reported in CS4: “a wide variance in GCCS experience and knowledge was observed among operators. Manual operator data fusion procedures must be formalized in standard procedures and trained to sufficient levels of proficiency.”

Beside the campaign results of HB5, it is worth noting the MBMs post-campaign published results that include the following examples of generalization and emergent properties for such operational environments and scenarios (Figure 62 and Figure 63) [Labbé and Maamar 2002].

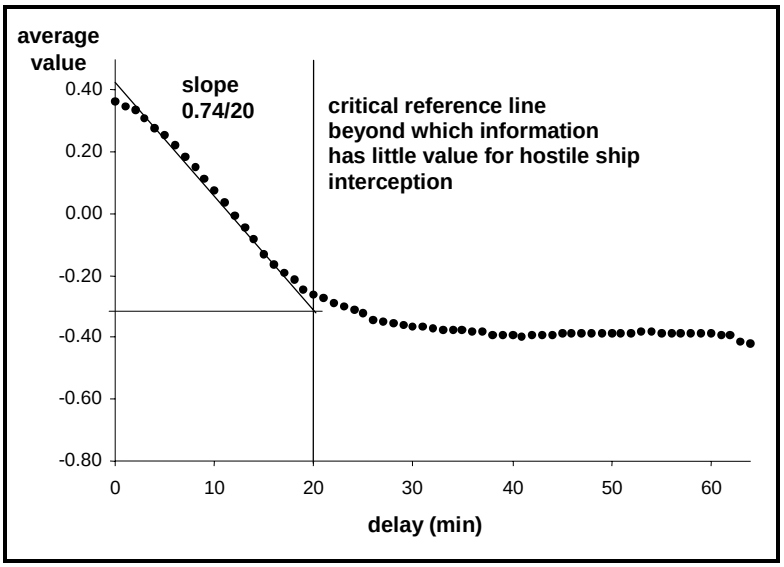


Figure 62 Ship-engagement effectiveness as function of information age

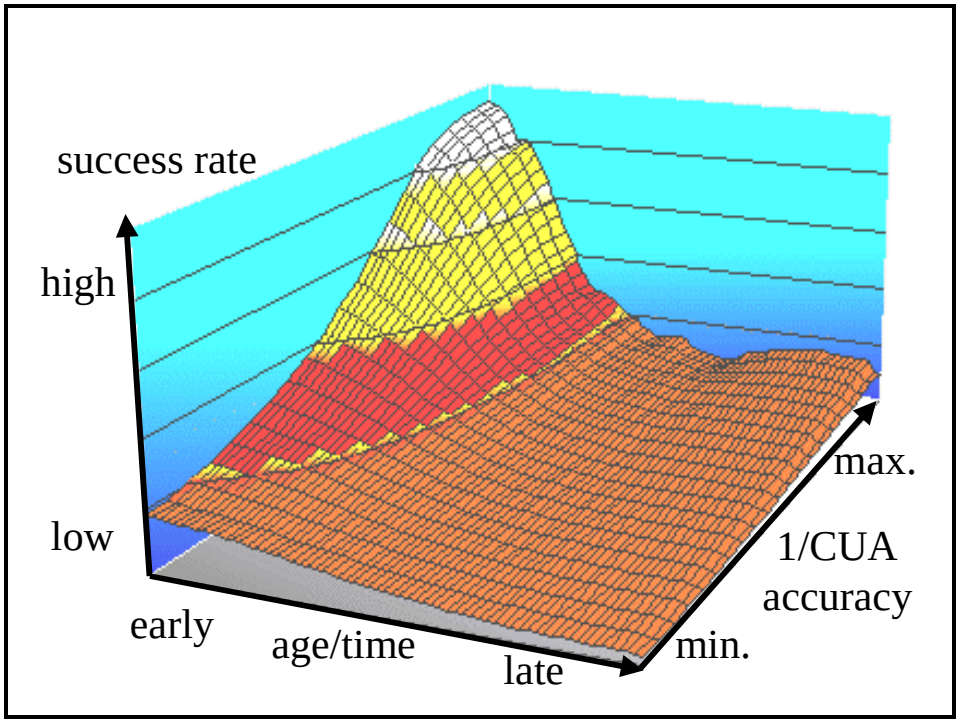


Figure 63 Potential mission success rate as function of input information age and accuracy expressed by MBM's circular uncertainty area (CUA)

Another important result, based on the findings of this Case Study shown in Figure 64, relates to information management in a strongly distributed data fusion environment typical in a GIG concept and for net-centric operations.

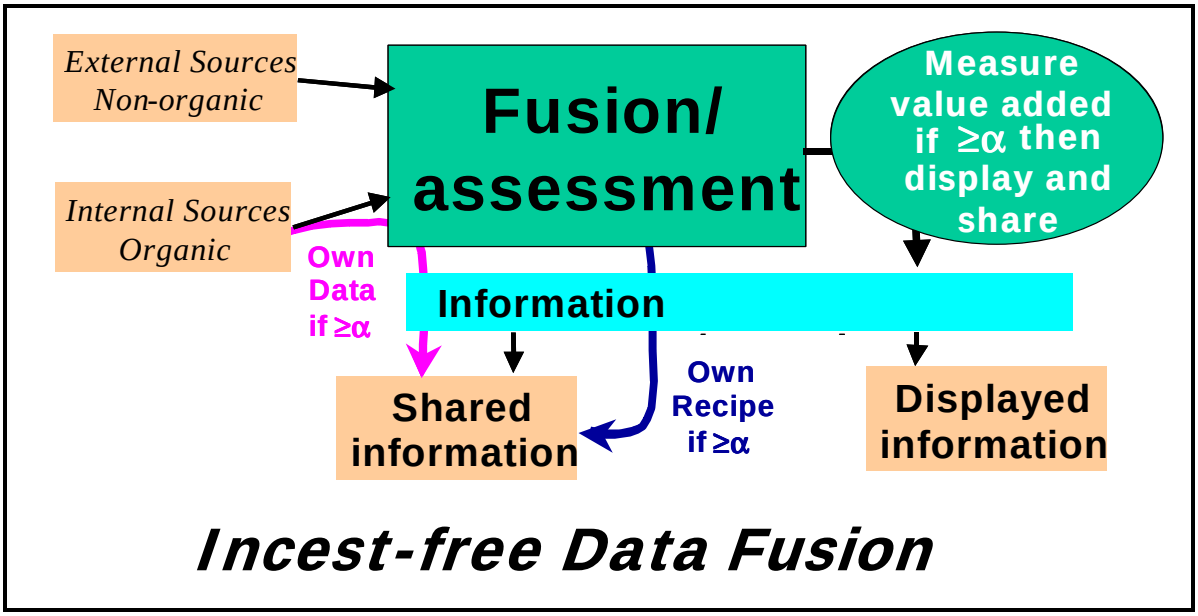


Figure 64 Generalization from MONIME and MBM observations: modified recommendation for improved network enabling activities such as GIG/NCW/NCO (network centric operations)

8.6 Lessons Learned and Interpretation in Terms of the GUIDEx Principles

Success of the MONIME campaign was due to proper problem definition (P4); an iterative process to reach an agreement between analysts and management (P5); integration of the three scientific methods of knowledge discovery and synthesis (P6); exploitation of all the methods available from national resources supported by adequate experiment design to increase analysis robustness (P1-3, 7); techniques to counter human variability (P8); special considerations in exploiting collective training (P9), adequate exploitation of M&S (P10), impressive (exhaustive) data analysis and collection plans (P12); and most importantly a continuous (P14) review of progress with the customer.

The following table summarizes which of the 14 experimentation Principles were appropriate for this Case Study , which ones were specifically addressed, and how they were addressed. Answers fall in one of these three categories: yes (Y), this Principle was or somewhat (S) was addressed, or not at all (N).

CS8 Improved Instruments Increase Campaign Values

#	GUIDEx Principle	Relevant	Addressed	How was it addressed?
1	Defense experiments are uniquely suited to investigate the cause-and-effect relationships underlying capability development.	Y	Y	As indicated in MONIME's methodology, applying deltas (small changes) on rules and scenarios were to deliver a robust analysis, *i.e.*, providing evidences for cause-and-effect relationships. The MBMs have an intrinsic causality clause included in their design simplifying the task of identifying causal relations to changing a single variable at a time (everything else being the same) and observing the hypothetical or potential effects (through an MoE approximation).
2	Designing effective experiments requires an understanding of the logic of experimentation.	Y	Y	All the measures and "If" and "then" clauses were evaluated against resources made available to MONIME, best methods and resources for a given problem were identified in the requirement document (included definitions of measures needed to attain the objectives and how to compute them, and a matrix of measures versus experimental capabilities).
3	Defense experiments should be designed to meet the four validity requirements.	Y	Y	For the field exercises and the HITL experiments, training was not always done as thoroughly as MONIME would have liked due to staff availability. Instrumentation of some of the platforms was difficult and data collectors did not always follow the established procedures.
a	Ability to employ the new capability	Y	S	Yes in most of the trials but not much for the field exercises since MONIME was not the primary driver.
b	Ability to detect change	Y	Y	Yes for most of the cases, but some required further investigations, *e.g.*, one extension was the development of the MBMs to detect the effect of specific changes.
c	Ability to isolate the reason for change	Y	S	Yes for several cases but some required further analyses. MBMs provided for several of these challenges but did not resolve all of them.
d	Ability to relate results to actual operations	Y	Y	The methodology used included data from real operations, which were re-injected in the experiments, and the active participation of the customer into the IPT provided evidence to support such generalization of experimental results to operational capabilities.

CS8 Improved Instruments Increase Campaign Values

#	GUIDEx Principle	Relevant	Addressed	How was it addressed?
4	Defense experiments should be integrated into a coherent campaign of activities to maximize their utility.	Y	Y	MONIME's globally endorsed methodology consisted of exploiting a variety of methods in order to benefit from each of them, *e.g.*, lessons learned from real-operations, field trials for identifying issues that needed further investigation, HITL for testing hypotheses where decisionmakers were a key component of the system, wargames to relate systems to mission effectiveness, and constructive simulations to quantify systems issues in terms of warfighting activities. Moreover, the addition of new tools, the MBMs, allowed us to observe effects that were lost due to experiment noises.
5	An iterative process of problem formulation, analysis and experimentation is critical to accumulate knowledge and validity within a campaign.	Y	Y	As shown by Figure 61, which illustrates the flow of data and the relation between methods of the MONIME's global methodology, an iterative process was used to acquire and accumulate the knowledge needed for specifying future systems that better support decisionmakers for more diverse scenario conditions than observed before some S&T advancements and socio-cultural shifts. Unfortunately, due to time and budget limitations, more iteration was prevented. This raises the issue that organizations should plan for continuous experimentation in order to increase campaign benefits to the combatants, providing more timely, well-tailored concepts and integrated systems.
6	Campaigns should be designed to integrate all three scientific methods of knowledge generation (studies, observations and experiments).	Y	Y	Fortunately, this campaign integrated well the three scientific methods of knowledge generation, providing an economy of contingent resources for the experimentation segments.
7	Multiple methods are necessary within a campaign in order to accumulate validity across the four requirements.	Y	Y	As indicated above, besides the difficulty to train the users for a new system, tactic, technique or procedure, and given that MONIME had to provide recommendations for future systems, several of the tests required some creativity from the military staff. MONIME's global methodology is an example and an extension of this paradigm by providing explicit information flows and processes for its implementation toward the global objectives of delivering recommendations supported by customers' interventions.

CS8 Improved Instruments Increase Campaign Values

#	GUIDEx Principle	Relevant	Addressed	How was it addressed?
8	Human variability in defense experimentation requires additional experiment design considerations.	Y	S	For HITL experiments, MONIME selected a strategy that alleviates some of this variability by changing few parameters at a time and using differential measurement techniques (one FOTC with two SAGs allowed to compare between all the following combinations: ground-truth, perfect fusion of all sensor data or local plus received COP, FOTC, HIT, SAG1, SAG2). Post-experiment analyses of data and re-enactment using MBMs reduced substantially the problems encountered.
9	Defense experiments conducted during collective training and OT&E require additional experiment design considerations.	Y	Y	MONIME considered field exercises as good knowledge generators as long as there were no questions about causality. Such an approach increases the risk of accepting wrongly counter-intuitive relations. For example, operators thought that the COP was better than what MONIME reported. A more significant one was the belief in the technical and military communities that the available dead-reckoning function improved the COP value for targeting, which was confirmed to be false for the hostile tracks as shown by using MBMs.
10	Appropriate exploitation of M&S is critical to successful experimentation.	Y	Y	The MONIME campaign and the MBMs are almost impossible to do without appropriate M&S, the former because of the lack of availability of battle groups dedicated to experimentation and the latter due to the order of magnitude of the computation required (tera-floating-point operations).
11	An effective experiment control regime is essential to successful experimentation.	Y	Y	As for P8 one can say that a good level of control was attained. However, during the conduct of the second HITL experiment MONIME decided to change some of the parameters assuming that sufficient data had been collected with the original parameters. With appropriate post-experiment analysis this was confirmed. It was a calculated risk according to the experiment director. This was not possible for RIMPACs and MARCOTs.
12	A successful experiment depends upon a comprehensive data analysis and collection plan.	Y	Y	Effectively, those documents were more elaborate for the field exercises than for the wargames. These plans allowed MONIME to channel its efforts into the tasks necessary to deliver the products agreed with the customer.
13	Defense experiment design must consider relevant ethical, environmental, political, multinational, and security issues.	Y	S	For the MONIME series, no particular problems were encountered. It was understood that security issues were to be the major handicap in conducting this international experimentation campaign. Note that more attention to human interfaces would have been a plus.

#	GUIDEx Principle	Relevant	Addressed	How was it addressed?
14	Frequent communication with stakeholders is critical to successful experimentation.	Y	Y	Continuous communication with the executive committee, AUSCANNZUKUS, allowed MONIME to adjust the campaign to what was possible to deliver, given battle group, new equipment, laboratory, SME and funding availability. The resulting handbook (April 1997) was accepted and almost immediately used to modify the AUSCANNZUKUS CONOPS (early 1998).

Table 17 Relation of CS8 to GUIDEx Principles

Among the lessons learned we include the fact that it was not expected that nations would have difficulties committing experienced GCCS operators, a situation that we recognized even today (See CS4).

Overall, despite the fact that most of the data were from observational studies of collective training, adjunct experimental methods allowed to identify sufficient causality relationships for a valid generation of the desired future systems requirements, an important generalization process for such campaigns.

Annexes

Annex A: Acronyms, Initialisms and Abbreviations

A21	Army in 21st Century (Australia)
AAR	after-action review or report
ABCA	American, British, Canadian, Australian Armies
ACT	Allied Command Transformation (NATO)
AEF	Army Experimental Framework (AU)
AFV	armored fighting vehicle
AG	Action Group
AHWG	*Ad-hoc* Working Group
AIS	automated identification system
ALIX	Atlantic Littoral ISR Experiment
ANCOVA	analysis of covariance
ANOVA	analysis of variance
ARH	Armed Reconnaissance Helicopter
ARTD	Applied Research Technology Demonstrator (UK)
ATGW	anti-tank guided weapon
AU	Australia
AUSCANNZUKUS	Australia, Canada, New Zealand, United Kingdom, and United States Naval C4 Organization
AWE	Advanced Warfighting Experiment
BARS	behavioral anchored rating scale
BATUS	British Army Training Unit Suffield (in Alberta, Canada)
BG	battlegroup
BLOS	beyond line-of-sight
BOS	behavior observation scale
C2	command and control
C2IR	command and control information requirement
C3	command, control, and communications
C3I	command, control, communications and intelligence or TTCP Command, Control, Communications, and Information Systems Group
C3TRACE	command, control, and communications—techniques for the reliable assessment of concept execution
C4	command, control, communications and computers
C4I	command, control, communications, computers and intelligence
C4ISR	command, control, communications, computers, intelligence surveillance and reconnaissance
CA	Canada
CAD	Canadian Dollars
CAEn	Close Action Environment

Annex A: Acronyms, Initialisms and Abbreviations

CAMEX	Computer Map Exercise
CAST	Command and Staff Training
CBHSS	Coalition Based Health Services Support
CBS	Corps Battle Simulation
CCIR	commander's critical information requirement
CCIRM	collection, coordination and information requirements management process (UK)
CCRP	Command and Control Research Program
CD&E or CDE	concept development and experimentation
CDC	concept development conference
CFBL	Combined Federated Battle Laboratories
CFBLNet	Combined Federated Battle Laboratories Network
CFEC	Canadian Forces Experimentation Centre
CG	control group
CGF	computer generated forces
CIACG	Coalition Interagency Coordination Group
CIE	Collaborative Information Environment
CIS	command information system
CISR	coalition intelligence, surveillance, and reconnaissance
CJCS	Chairman of the Joint Chiefs of Staff
CMS	Chief Maritime Staff
COBP	code of best practice
COBPE	code of best practice for experimentation
COI	critical operational issue
CONOP	concept of operations
COP	common operational picture
COTS	commercial off-the-shelf
CPX	command post exercise
CRT	cathode-ray tube
CS	Case Study (With capitals for GUIDEx CSs, case study otherwise)
CUA	circular uncertainty area
CWC	composite warfare commander
CWID	Coalition Warrior Interoperability Demonstrations
DCEE	Distributed Continuous Experiment Environment
DCTS	Defense Collaboration Tool Suite
DIS	Distributed Interactive Simulation
DISA	Defense Information Systems Agency
DMSO	Defense Modeling and Simulation Office
DoD	Department of Defense
DoDAF	Department of Defense Architecture Framework
DOTMLPF	doctrine, organization, training, material, leadership, personnel and facilities
DRDC	Defence Research and Development Canada

Annex A: Acronyms, Initialisms and Abbreviations

Dstl	Defence science and technology laboratory
DSTO	Defence Science and Technology Organisation
EADS	European Aeronautic Defence and Space
EBO	effects-based operation
EBP	effects-based planning
EO	electro-optic
ESM	electronic support measure
ETO	effects tasking order
EUCLID	European Cooperation Long-term In Defence
EXCON	exercise control
EXFOR	experimental force
FDDS	Flag Data Display System
FEDEP	Federation Development and Execution Process
FOSP	Fleet Ocean Surveillance Product
FOTC	Force Over-the-horizon Track Coordinator
FPC	Final Planning Conference
FWC	Future Warfighting Concept
GCCS	Global Command and Control System
GCCS-M	Global Command and Control System - Maritime
GICP	good idea cutoff point
GIG	Global Information Grid
GLM	general linear model
GMTI	ground moving target indicator
GUIDEx	TTCP Guide for Understanding and Interpreting Defense Experimentation
HB5	AUSCANNZUKUS Handbook Five: Guidelines for Maritime Information Management
HEAT	Headquarters Effectiveness Assessment Tool
HIT	high interest track
HITL	human-in-the-loop
HLA	High Level Architecture
HQ	headquarters
HSP	health and safety plan
HUM	TTCP Human Resources and Performance Group
HW	hardware
HW/SW	hardware/software
I2	information and intelligence
IAEC	integrated analysis and experimentation campaign
IAI	Israeli Aircraft Industries
IERs	information exchange requirements
IISRA	Integrated ISR Architecture
IM	information management
IO	information operations

IPB	intelligence preparation of the battlespace
IPC	Initial Planning Conference
IPT	Integrated Project Team
IR	infra-red
IRM	information requirements management
ISAR	Inverted Synthetic Aperture Radar
ISAT	Information Systems Architecture Tool
ISR	intelligence, surveillance and reconnaissance
ISTAR	intelligence, surveillance, target acquisition and reconnaissance
IWS	InfoWorkSpace
Janus	Roman god that is identified with doors, gates, and all beginning… represented artistically with two opposite faces (Merriam-Webster online) In GUIDEx, a computer generated (HITL) wargame
JBC	Joint C4ISR Battle Center
JCATS	Joint Conflict and Tactical Simulation
JCD&E	joint concept development and experimentation
JDCAT	JBC Data Collection and Analysis Tool, or Joint Data Collection and Analysis Tool
JDEF	Joint Demonstration and Evaluation Facility
JDL	Joint Director of C3 Laboratories
JFC	joint force capability
JFCOM	Joint Forces Command
JOTS	Joint Operational Tactical System (US)
JP	joint program
JROC	Joint Requirements Oversight Council
JSA	TTCP Joint Systems Analysis Group
JSAF	Joint Semi-Automated Forces
JSF	Joint Strike Fighter
JSTARS (ASTOR)	Joint Surveillance Target Attack Radar System (Airborne Stand-off Radar)
JTF	Joint Task Force
JTLS	Joint Theater Level Simulation
JUEP	Joint UAV Experimentation Program
JWARS	Joint Warfare System
JWID	Joint Warrior Interoperability Demonstration
KM	knowledge management
LAN	local area network
LER	loss exchange ratio
LIVEX	live exercise
LOCON	lower control
LOS	line-of-sight
M&S	modeling and simulation

Annex A: Acronyms, Initialisms and Abbreviations

MALE	Medium Altitude Long Endurance
MANOVA	multivariate analysis of variance
MAPEX	Map Exercise
MAR	TTCP Maritime Systems Group
MARCOT	Maritime Command Operational Training Exercise (CA)
MBM	model-based-measures
MEL	master events list
M-E-M	model-exercise-model
MIC	Multinational Interoperability Council
MIST	Maritime Intelligence Support Terminal
MITRE	MIT (Massachusetts Institute of Technology) Research and Engineering
MN LOE II	Multinational Limited Objective Experiment II
MNE	Multinational Experiment
MNIS	multinational information sharing
MoD	Ministry of Defence (UK)
ModSAF	Modular Semi-Automated Forces
MoE	measure of effectiveness
MoM	measure of merit
MONIME	Management of Organic and Non-organic Information in a Maritime Environment
MoP	measure of performance
MOSP	Multi-Mission Optronic Stabilized Payload
MPC	Main Planning Conference
MRC	multiple regression correlation
MSEL	master scenario event list
M-T-M	model-test-model
M-W-M	model-wargame-model
NAMRAD	Non-Atomic Military Research and Development
NATO	North Atlantic Treaty Organisation
NCO	network centric operations
NCW	network centric warfare
NEC	network enabled capability
NG	New-capability Group
NISAS	Naval Information Systems Architecture Study (UK)
NITEworks	Network Integration Test and Experimentation works
N-KRS	Navy Knowledge-based Replanning System
NL	National Leader
NMCC	National Military Commander Center
NOSC	Naval Ocean Systems Center
NPS	Naval Postgraduate School
NRF	NATO Response Force
OA / OR	operational analysis / operational research

Annex A: Acronyms, Initialisms and Abbreviations

OCOM	ordinary command
ONA	operational net assessment
OODA	observe-orient-decide-act (model, loop, cycle…)
OOTW	operation other than war
OPFOR	opposing force
OPI	Office of Primary Interest
OT&E	operational test and evaluation
OTH-T	over-the-horizon targeting
OV	operational view
PLIX	Pacific Littoral ISR Experiment
PMESII	political, military, economic, social, information and infrastructure
QDR	Quadrennial Defense Review
QIER	quantitative information exchange requirement
QIR	quantitative information requirement
RAN	Royal Australian Navy
Recce	reconnaissance
REPEAT	Repeatable Performance Evaluation and Analysis Tool
RESA	Research and Analysis for Systems Engineering
RFI	request for information
RIMPAC	Rim of the Pacific In this document it refers to a naval LIVEX of Pacific Rim countries.
RMP	recognized maritime picture
RTA	Restructuring the Army (Australia)
RTO	Research and Technology Organisation
S&T	science and technology
SA	situation assessment
SAG	surface action group
SAR	synthetic aperture radar
SCD	Senior Concept Developer
SE	synthetic environment
SEDEP	Synthetic Environment Development and Exploitation Process
SIMEX	Simulation Exercise
SIMNET	Simulation Network
SJFHQ	Standing Joint Force Headquarters
SME	subject matter expert
SOPs	standard operating procedures
SV	system view
TAOR	Tactical area of responsibility
TEWT	Training Exercise Without Troops
TFXXI	Task Force XXI
TIMSIM	Tactical Information Management Simulation
TP	Technical Panel

TRAC	Training and Doctrine Command Analysis Center
TRADOC	US Army Training and Doctrine
TSI	Total Systems Intervention
TTCP	The Technical Cooperation Program
TTPs	tactics, techniques and procedures
TUAV	tactical unmanned air vehicle
TV	technical view
TWCS	Tomahawk Weapons Control System
UAV	unmanned air vehicle
UJTL	Universal Joint Task List
UK	United Kingdom
US/USA	United States of America
USJFCOM	US Joint Forces Command
USN	United States Navy
USV	uninhabited surface vehicle
UTC	Coordinated Universal Time
VFR	visual flight rules
VOI	vessel of interest
VPN	virtual private network
VVA	verification, validation and accreditation
WAN	wide area network
WAP	wide-area naval tactical picture
XCOM	Experimental Command Team
XDJTFHQ	Experimental Deployable Joint Task Force Headquarters

Annex B: Lexicon for Defense Experimentation

Purpose. The purpose of this lexicon is to assist with the understanding of concepts elaborated in GUIDEx. Many of these terms have formal (and differing) definitions across the nations, but this lexicon explains specifically how they are used (or some cases why they are not) in GUIDEx. Many are taken from "Experimental and Quasi-Experimental Designs for Generalized Causal Inference" [Shadish *et al.* 2002]. An asterisk (*) identifies these definitions. Where no source is identified they have been agreed during the deliberations of AG-12. Cross-references to other terms in the lexicon are emboldened.

Warning. Although the term "warfighting experimentation" is used by all of the TTCP nations, AG-12 has found that its meaning is not consistent across the nations and it is not helpful in communicating GUIDEx's message. For example: in some countries it is taken and used to imply experimentation only in warfighting scenarios, rather than in all military operations; in some it is taken to mean only experimentation involving the presence of warfighters in their operational role; and in some it is taken to cover all empirical military analyses, not just experimentation as described in this guide. Consequently, this expression and others creating interpretation problems have been avoided as much as possible in GUIDEx and this is reflected in its lexicon.

Term	Definition	Source
acceptance test	A **test** or series of tests which are undertaken on a capability to show that it meets the criteria laid down by the government for acceptance into service.	
advanced warfighting experiment	A US term, usually meaning **defense experimentation** tackling complex transformational issues on a large scale.	
analytic wargame	Analytic Wargames typically employ command and staff officers to plan and execute a military operation, often with some form of **constructive simulation** adjudicating outcomes between turns (sometimes overnight).	
between-participants design	[A design where] different units are studied in different conditions. See also **multiple group design**.	*

Annex B: Lexicon

Term	Definition	Source
brainstorming	i A process used to generate new ideas in a team environment using creativity techniques. ii A process that attempts to solve a problem by a method in which the members of a group spontaneously propose ideas and solutions, disallowing critique until brainstorming is completed.	
campaign	See **integrated analysis and experimentation campaign.** Not used in GUIDEx to mean "military campaign" unless explicitly stated.	
capability	Capability is the power to achieve a desired operational effect in a nominated environment within a specified time and to sustain that effect for a designated period. Capability is delivered through the **lines of development** or "**DOTMLPF**."	
capability development	Development of military **capability**, short-to-long term. Note that this term has more specific meanings in some countries.	
cause	A variable that produces an effect or result.	*
closed loop modeling or simulation	See **constructive simulation**.	
confound	An extraneous variable that covaries with the variable of interest.	*
construct	A concept, model or schematic idea.	*
constructive simulation	The closed-loop force-on-force simulations employed by the modeling and simulation and military operational research communities. Once designers choose the initial parameters, start the simulation, and run it to completion, there is no human intervention in the play of the simulation. **Analytic wargames** sometimes use such simulations but the human intervention is essentially between runs. In some quarters, the term **constructive simulation** is used to describe large scale command post exercise (CPX) drivers such as JTLS. In GUIDEx the term is NOT used in this way and such tools would be considered to be **HITL simulations**.	

Annex B: Lexicon

Term	Definition	Source
control group	In an **experiment**, this term typically refers to a comparison group that does not receive a treatment but may be assigned to a no-treatment condition (to a wait list for treatment, or sometimes to a placebo intervention group).	*
control variable	One can prevent the effects of a specific identifiable extraneous variable from clouding the results of an experiment by holding the value of this extraneous variable constant, *e.g.,* all selected subjects have the same level of training, **C**. A variable that is thus held constant is called a **control variable**. Similarly, in a multiple regression equation, specific extraneous independent variables, *e.g.,* **C**, can be held constant or statistically controlled in examining the impact of **A** on **B**, the dependent variable. The resulting correlation is then called a partial correlation between **A** and **B** controlling for **C**.	
correlational study	See **observational study**.	
data collection plan	A plan that explains how the requisite data will be collected and validated prior to analysis. The plan will identify what data are being collected, the collection techniques and the method of validation.	
defense experiment / experimentation	The application of the experimental method to the solution of complex defense capability development problems, potentially across the full spectrum of conflict types, such as warfighting, peace-enforcement, humanitarian relief and peace-keeping.	
demonstration	An **event** to exhibit a prototype or explain an already known fact or observation. May be a source of information for a decision, or may provide evidence or justification for further experimentation.	Oxford English Dictionary (OED)
dependent variable	Often synonymous with effect or outcome, a variable with a value that varies in response to an **independent variable.**	*
DOTMLPF	A US term meaning the components of military capability: doctrine, organization, training, materiel, leadership, personnel, and facilities. See also **lines of development**.	
effect size	A measure of the magnitude of a relationship.	*

Annex B: Lexicon

Term	Definition	Source
effectiveness	How well an intervention works when it is implemented under conditions of actual application.	*
efficacy	How well an intervention works when it is implemented under ideal conditions.	*
empirical study	See **observational study**.	
evaluation	The process of determining, by whatever means, the quality of a concept or system of interest by comparing it against appropriate criteria or requirements. When done practically or empirically, this is enacted by **testing**.	
event	A generic term which may be used to describe a **demonstration, test**, **experiment** or **observational study** prior to the designation of those terms.	
exercise	A simulated maneuver or operation involving [some or all of] planning, preparation, and execution (usually for the purposes of training). When used in **model-exercise-model** the usage is the general sense above, not specifically training.	UK MoD Official
exercise exploitation (intrusive)	Exploiting a training **exercise** for experimental or other non-training-related purposes where there is a need for some deliberate and pre-agreed intervention into the running of the exercise.	
exercise exploitation (passive)	Exploiting a training **exercise** for experimental or other non-training-related purposes where there is no interference in the running of the exercise and only unobtrusive, passive data collection will be performed.	

Annex B: Lexicon

Term	Definition	Source
experiment	i. (Generally) To try something new and see what happens. ii. To explore the effects of manipulating a variable. iii. An empirical means of establishing cause-and-effect relationships through the manipulation of independent variables and measurement of dependent variables in a controlled environment. **Experimentation** is enacted by the testing of hypotheses. iv. Experiments are empirical deductive activities.	* UK NITEworks
experiment design or experimental design	A detailed description of the methods, techniques, analytical methods and tools that will be used in undertaking an **experiment**. The plan of the experiment which specifies the treatment conditions (independent variables), what is to be measured (dependent variables) and methods of assigning subjects to groups.	http://psy.st-andrews.ac.uk/resources/glossary.shtml
experiment methods	The tools, techniques, manipulations and perturbations that are used as part of the **experiment**, and are used in data reduction and analysis.	US COBP for Experimentation
external validity	The ability to generalize the cause-and-effect relationship found in the **experiment** environment to the operational military environment.	
fatigue effects	The effects of participants tiring over time, causing performance deterioration in later conditions or later assessments.	* (paraphrase)
field exercise	Any **exercise** using **live simulation**.	
field experiment	A **defense experiment** based on **live simulation**.	
human-in-the-loop (HITL) simulation	Any simulation with which humans interact in real time. Includes computer generated forces (CGF) (*e.g.,* JSAF); **virtual simulators**, simulations designed for multi-sided **wargaming** (*e.g.,* Janus); and CPX drivers (*e.g.,* JTLS).	

Term	Definition	Source
hypothesis	An assertion, proposition or statement about relations or constraints whose truth-value is as yet unknown, but in principle is determinable by **tests** (definition ii of **test**) involving generally empirical but also logical evidence.	Web Dictionary of Cybernetics and Systems
independent variable	Often synonymous with **cause** or **treatment**, a variable that purports to be independent of other influences.	*
insight	A set of observations that suggest, but do not prove, a hitherto hidden truth.	
integrated analysis and experimentation campaign	A planned sequence of related **defense experiments**, studies and/or analytical activities designed to advance the understanding of a military force development problem. Within the campaign, the key role of an experiment is to generate some linkage between cause-and-effect. **Integrated analysis and experimentation campaigns** can mitigate the risks associated with particular analytical techniques using the strengths inherent in other methods and thus build validity in the campaign outcomes.	
internal validity	The ability to determine if a causal relationship exists between two variables.	
learning effects	See **practice effects.**	
lines of development	UK term meaning the components of military **capability:** training, equipment, personnel, information, doctrine and concepts, organization, infrastructure, logistics. See also **DOTMLPF**.	
live simulation	Simulation of military operations in a live environment with actual military units and with real military equipment and operational prototypes, with only weapon effects being simulated. For example, Air Combat Maneuvering Instrumentation (ACMI) ranges and field environments using laser-based weapon effects simulators.	
measure	A measure is a standard by which some attribute of interest is recorded.	[Alberts and Hayes 2002]

Annex B: Lexicon

Term	Definition	Source
measure of effectiveness	A **measure** that describes the influence or benefit of a concept within its operational context.	
measure of performance	A **measure** that describes the influence or benefit of a concept in terms of its internal structure, characteristics and behavior.	
metric	A set of measurements, not just one, that quantify results.	
model	A mathematical representation of something. Often implemented on a computer.	
model-exercise-model	A process that maximizes the relative benefits of modeling (usually **constructive simulation**) and empirical techniques of analysis (**experimentation** or **observational studies** using **analytic wargames, virtual simulation** or **field exercises**). It normally comprises three phases: i. Initial use of a **constructive simulation** to help understand key drivers and sensitivities and to assist in designing the second phase; ii. An empirical event ("exercise") whose conditions replicate one or more of the modeled conditions; iii. A subsequent modeling phase, which has been validated, calibrated and/or modified by repopulating the original simulation with empirical data or results from the previous phase. This produces the final results and enables extrapolation from the empirical test condition.	
model-experiment-model **model-test-model** **model-wargame-model**	These terms are avoided in GUIDEx but are synonymous with **model-exercise-model**.	
multiple group design	See **between-participants design.**	
non-experimental study	See **observational study**.	

Term	Definition	Source
observational study	An objectively-observed, practical **event** which does not involve the deliberate or purposeful manipulation of independent variables to establish cause-and-effect relationships. Observational studies may be used to establish associative or correlative relationships. See also **correlational study, empirical study** or **non-experimental study.** These are empirical-inductive activities.	
operational analysis (OA)	See **operational research**.	
operational assessment	An evaluation of Operational Effectiveness and Operational Suitability made by an independent operational test activity, with user support as required, on other than production systems. The focus of an operational assessment is on significant trends noted in development efforts, programmatic voids, risk areas, adequacy of requirements, and the ability of the program to support adequate Operational Testing. An operational assessment may be conducted at any time using technology demonstrators, prototypes, mock-ups, Engineering Development Models, or simulations, but will not substitute for the Initial Operational Test and Evaluation necessary to support Full Rate Production decisions.	http://www.dau.mil/pubs/glosary/preface.asp
operational research or **operations research (OR)**	OR looks at an organization's operations and uses mathematical or computer models, or other analytical approaches, to find better ways of doing them. Often applied to problems of military **capability development.**	UK OR Society
operational test and evaluation	Formal testing conducted prior to deployment to evaluate the operational effectiveness and suitability of the system with respect to its mission.	http://foldoc.doc.ic.ac.uk/foldoc
order effects	The effects on the outcome of a study produced by the order in which the treatments were presented. See also **practice effects** and **fatigue effects**, both of which can contribute to these.	* (paraphrase)

Term	Definition	Source
practice (practise) effects	The effects of participants' learning and performance improvement due to repetition are important problems in **within-participants designs** in which repeated tests are given to the same participants. See also **learning effects.**	* (paraphrase)
scenario	A description of the area, the environment, means, objectives and events related to a conflict or a crisis during a specified time frame suited for satisfactory study objectives and the problem analysis directives.	[NATO 2002]
seminar	An occasion when a group of experts meet to study and discuss.	Oxford English Dictionary (OED)
seminar wargame	A structured discussion between experts in several fields to elicit opinions and judgments from them, and to increase understanding. It is more structured than **brainstorming** (or **seminars**), but is not normally supported by any kind of simulation (like **analytic wargames).**	UK Dstl
simulation	A time-variant model.	
simulation method	Broad category of simulation techniques with identifiably different benefits and disadvantages for supporting **defense experimentation**. For example; **constructive simulation, analytic wargame, HITL simulation, live simulation.**	
single-group design	See **within-participants design**.	
synthetic environment	A computer based representation of the real world, usually a current or future battlespace, within which any combination of "players" may interact. The players could be computer **simulations**, people or instrumented real equipment. SEs are usually taken to include a set of networked and interoperating **simulations.** In a broader sense, SEs are credibly synthesized military environments other than real operations.	UK MoD Official

Annex B: Lexicon

Term	Definition	Source
test	i. A practical or empirical **event** to **evaluate** a concept or system of interest by measuring it against appropriate criteria or requirements; ii. In the experimentation sense, the means of determining the veracity of a **hypothesis**.	
threats to validity	Reasons why an inference might be incorrect.	*
training	The process of teaching, familiarizing and bringing to a known and common skill level operators or users of a concept or system. Often categorized as individual, team or collective training.	
treatment	A set of controlled experimental conditions in which the various **independent variables** are fixed. Comparison of different treatments is the normal means of testing the experimental **hypothesis**.	
trial	i. (Experimental sense) A single opportunity to observe the effect of a particular **treatment** in an experiment. ii. Often used to mean the same as **test**, as in Field Trials, Flight Trials, Sea Trials, *etc*.	
validity	The truth of, or correctness of, or degree of support for an inference.	*
venue (for experimentation)	The location of an **event**. Sometimes taken to mean "simulation method" but not used in that way within GUIDEx.	
VV&A (validation, verification and accreditation)	A process, often formalized, of ensuring that something (often a model or simulation) is fit for its intended purpose, or adequate, and is accredited as such by relevant stakeholders.	
virtual simulation	That subset of **HITL simulations** in which manned equipments (usually platforms such as aircraft or AFVs) are explicitly simulated.	

Annex B: Lexicon

Term	Definition	Source
warfighting experiment	*A term not used in GUIDEx.* Although this term is used by most of the TTCP nations, AG-12 has found that its meaning is not consistent across the nations and it is not helpful in communicating GUIDEx's message. For example: in some countries it is taken and used to imply experimentation only in warfighting scenarios, rather than in all military operations; in some it is taken to mean only experimentation involving the presence of warfighters in their operational role; and in some it is taken to cover all empirical military analyses, not just experimentation as described in this guide.	
wargaming	A synthesis of warfare with a defined ruleset, involving the multi-sided and adversarial engagement of human players. Wargames may or may not use an experimental approach as described in GUIDEx. The possible range of underlying computer simulation support is: i. none (*i.e.,* seminar or tabletop wargames); ii. an Analytic Wargame (*i.e.,* turn-based adjudication); or iii. a HITL simulation (*e.g.,* Janus or JSAF) (*i.e.,* continuous human interaction). iv. Human interaction with wargames is usually, but not necessarily, abstract, in that the real organizational structures and manning levels are not accurately represented. For example, two or three officers may represent an entire headquarters.	
within-participants design	[A design where] the same units are studied in different conditions. Also known as **single-group design.**	*

Annex C: Bibliography and Selected References

ABCA. 2004. "American, British, Canadian, and Australian Armies' Standardization Program Analysis Handbook (draft for unlimited distribution)." 66 p. http://abca.hqda.pentagon.mil/

Alberts, David S. and Richard E. Hayes. 2002. *Code of Best Practice for Experimentation*. Washington, DC: CCRP. 436 p. http://dodccrp.org/html/pubs.html

—. 2005. *Campaigns of Experimentation; Pathways to Innovation and Transformation*. Washington, DC: CCRP. 227 p. http://dodccrp.org/html/pubs.html

Alker, Hayward. 1971. *Mathematics and Politics*. New York: Macmillan. 152 p.

Alley, Michael. 1987. *The Craft of Scientific Writing*. Englewood Cliffs, NJ: Prentice-Hall. 225 p.

Alred, Gerald J., Charles T. Brusaw, and Walter E. Oliu. 2003. *Handbook of Technical Writing*. New York: St. Martin's Press. 645 p.

AUSCANNZUKUS MONIME. 1992. "The Management of Organic and Non-organic Information in a Maritime Environment; Analysis Requirements." AUSCANNZUKUS: The Management of Organic and Non-organic Information in the Maritime Environment (MONIME) *Ad-hoc* Working Group (Command, Control and Communications Committee). 124 p.

—. 1993. "Tactical Information Management Simulation (TIMSIM) 1993." The Management of Organic and Non-organic Information in the Maritime Environment (MONIME) *Ad-hoc* Working Group (Command, Control and Communications Committee). 79 p.

—. 1997. "Handbook 5: Guidelines for Maritime Information Management." The Management of Organic and Non-organic Information in the Maritime Environment (MONIME) *Ad-hoc* Working Group (Command, Control and Communications Committee). 238 p.

Australian Army. 2000. "The Army Experimental Framework." Australian Army Publication.

Bowen, C. and K.R. McNaught. 1996. "Mathematics in Warfare: Lanchester Theory." p. 141-156 in *The Lanchester Legacy, Volume III - A Celebration of Genius*, edited by N. Fletcher. Coventry, UK: Coventry University Press.

Campbell, Donald T. and J. C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally. 84 p.

Annex C: Bibliography and Selected References

CBE (Council of Biology Editors) Inc. 1994. *Scientific style and format - The CBE Manual for Authors, Editors, and Publishers*. Cambridge, UK: Cambridge University Press. 825 p.

Cebrowski, VADM A. and J. Garstka. 1998. "Network Centric Warfare: Its Origins and Future." *Proceedings of the Naval Institute* 124(1):28-35.

Cook, Thomas D. and Donald T. Campbell. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston: Houghton Mifflin. 405 p.

Dagnelie, Pierre. 2003. *Principes d'expérimentation: planification des expériences et analyse de leurs résultats*: Electronic edition. 397 p. http://www.dagnelie.be

DoD Modeling and Simulation Office (DMSO). 1996. "Verification, Validation and Accreditation Recommended Practices Guide." Department of Defense.

DoDAF Working Group. 2004. "DoD Architecture Framework, Version 1.0." 87 p.

DRDC HREC. 2002. "Guidelines for Human Subject Participation in Research Projects." Defence R&D Canada (DRDC) - Toronto, Human Research Ethics Committee (HREC). 69 p.

Einstein, Albert. 1953. Letter to J. S. Switzer. Einstein Archive-61-381. 23 April 1953. Jerusalem: The Hebrew University.

Einstein, Albert. 1950. Einstein, Albert (1879-1955) Pencil Sketch Signed ("A. Einstein"), at lower left, by New York artist Barney Fagan, who signed ("B. Fagan, 1950") at lower right, n.p., 11 x 8½ in. Fagan drew celebrities from life in various New York restaurants from 1933-53. The head-and-shoulders portrait is a wonderful likeness, featuring Einstein's large, expressive eyes. Fagan titled the sketch, "Dr. Albert Einstein," and writes a quote below: "The grand aim of all science, is to cover the greatest number of empirical facts by logical deduction from the smallest number of hypotheses or axioms."

Feynman, Richard P. 1999. *The Meaning of It All: Thoughts of a Citizen Scientist*. USA: Perseus Books Group. 133 p.

Fisher, J.L., M.J. Brennan, and D.K. Bowley. 2003. "A Study of Land Force Modernisation Studies in DSTO 1996 to 2000 (RESTRICTED DISTRIBUTION)." DSTO-GD-0358.

Flood, R. L. and M. C. Jackson. 1991. *Creative Problem Solving: Total System Intervention*. Chichester, UK: Wiley. 250 p.

Hiniker, P.J. 1991. "HEAT Experiments: Use of the Experimental Method to Evaluate the Effectiveness of HQ C2 Insertions." Presented at 1991 Symposium on Command and Control, July 1991, Washington, D.C.

Hiniker, Paul and E. Entin. 1990. "The Effects of Shared Battle Graphics on Team Performance in Crisis Situations: HEAT Experimental Results." Presented at Proceedings of the JDL BRG C2 Research Symposium, July 1990.

—. 1992. "Examining Cognitive Processing in Command Crises: New HEAT Experiments on Shared Battle Graphics and Time Tagging." Presented at Proceedings of the JDL BRG C2 Research Symposium.

Ingber, L. 1989. "Mathematical Comparison of Combat Models to Exercise Data." Presented at Proceedings of the JDL BRG C2 Research Symposium, June 1989.

Johnson, B. 2000. It's (Beyond) Time to Drop the Terms Causal-Comparative and Correlational Research in Education.
http://it.coe.uga.edu/itforum/paper43/paper43.html

Kass, Richard A. 1984. "Calibrating Questionnaires and Evaluators." *The ITEA Journal of Test and Evaluation* 3:26-36.

—. 1997. "Design of Valid Operational Tests." *International Journal of Test and Evaluation* (June/July):51-59.

Kerlinger, F.N. 1986. *Foundations of Behavioral Research*. New York: Holt, Rinehart and Winston. 667 p.

Krepinevich, Andrew. 2001. "The Bush Administration's Call for Defense Transformation: A Congressional Guide." Center for Strategic and Budgetary Assessments. 3 p.

Labbé, Paul. 1997. "Analysis Methods for the Management of Tactical Information in a Maritime Environment." DREV R-9614. Defence Research Establishment Valcartier. 125 p.

Labbé, Paul and Zakaria Maamar. 2002. "Emergent Properties of E-activities for Location- and Time-dependent Information." Presented at Scuola Superiore G. Reiss Romoli (SSGRR) 2002w, January 2002, L'Aquila, Italy.

Labbé, Paul and René Proulx. 1998a. "Model-based Measures for the Assessment of Engagement Opportunities." DREV R-9712. Defence Research Establishment Valcartier, APG Solutions & Technologies Inc. 102 p.

—. 1998b. "Model-based Measures for the Assessment of Engagement Opportunities: Implementation and Test Results." DREV R-9807. Defence Research Establishment Valcartier, APG Solutions & Technologies Inc. 94 p.

—. 2000. "Impact of Systems and Information Quality on Mission Effectiveness." Presented at 5th International Command and Control Research and Technology

Symposium, 5th ICCRTS, October 2000, Canberra, ACT, Australia. http://www.dodccrp.org/2000ICCRTS/index.htm

Law, A. M. and W. D. Kelton. 2000. *Simulation Modeling and Analysis*. New York: McGraw-Hill Book Company. 759 p.

Lazarsfeld, Paul. 1958. "Evidence and Inference in Social Research." *Daedalus* 87(4):99-130.

MacMillan, J., E. Entin, and P. Lenz. 1988. "Experiment Report: The Effects of Option Planning and Battle Workload on C2 Effectiveness." TR-368. ALPHATECH, Inc. 60 p.

MacMillan, J. and J. Shaw. 1990. "Experimental Evaluation of a Knowledge-based Air Strike Mission Planning Aid." Presented at Proceedings of the JDL BRG C2 Research Symposium, June 1990.

Mathieson, Graham L. 2001. "The Impact of Information on Command - Hard Evidence." Presented at 18th International Symposium on Military Operational Research (alternatively Dstl/Pub 01369), August 2001.

Mathieson, Graham L. and Lorraine Dodd. 2004. "A Conceptual Model of Organisational And Social Factors In Headquarters." Presented at The 9th ICCRTS at the House of Engineers in Copenhagen, Denmark, September 2004. http://www.dodccrp.org/events/2004/ICCRTS_Denmark/CD/papers/128.pdf

McClave, J. T. and F. H. Dietrich II. 1991. *Statistics*. San Francisco, CA: Dellen Publishing. 928 p.

McNemar, Quinn. 1962. *Psychological Statistics*. New York: Wiley. 451 p.

Myers, Gen Richard B. 2003. "Understanding Transformation." *Air & Space Power Journal* XVII(1):5-10. http://www.airpower.maxwell.af.mil/airchronicles/apj/apj03/spr03/myers.html http://www.airpower.maxwell.af.mil/airchronicles/apj/apj03/spr03/spr03.pdf

NATO. 2002. *NATO Code of Best Practice for C2 Assessment*. Washington, DC: CCRP. 302 p. http://dodccrp.org/html/pubs.html

Needalman, A., D. Mikaelian, E. Entin, and R. Tenney. 1988. "Contingency Planning in Headquarters." Presented at Proceedings of the JDL BRG C2 Research Symposium, June 1988.

Newton, LCol. S. J., Maj. M. M. Regush, Paul Comeau, G. H. Van Bavel, and Richard K. Bowes. 2003. "Experiment Report 001/2003 (Quick Look): Pacific Littoral ISR Experiment - Part I." Experiment Report 001/2003. Canadian Forces Experimentation Centre. 17 p.

Owens, William A. ADM USN (Retired). 2002. "The Once and Future Revolution in Military Affairs." 61 p.

Pearl, Judea. 2001. *Causality*. Cambridge, UK: Cambridge University Press. 384 p.

Popper, Karl R. 1959. *The Logic of Scientific Discovery*. New York: Basic Books. 480 p.

Radder, H. 2003. *The Philosophy of Scientific Experimentation*. Pittsburgh, PA: University of Pittsburgh Press. 311 p.

Rosenbaum, Paul R. 2002. *Observational Studies*. New York: Springer-Verlag. 373 p.

Rosenthal, Robert. 2002. "Experimenter and Clinician Effects in Scientific Inquiry and Clinical Practice." *Prevention & Treatment* 5(38). http://www.journals.apa.org/prevention/volume5/pre0050038c.html

Ross, S.M. 1987. *Introduction to Probability and Statistics for Engineers and Scientists*. New York: John Wiley & Sons. 492 p.

Rossi, P.H. and H.E. Freeman. 1985. *Evaluation: A Systematic Approach*. Beverly Hills, CA: Sage Publications. 423 p.

Rumsfeld, Donald H. 2002. "Transforming the Military." 3 p.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin. 623 p.

Shorack, G. R. and J. A. Wellner. 1986. *Empirical Processes with Applications to Statistics*. New York: John Wiley & Sons. 938 p.

Simon, H.A. 1957. *Models of Man*. New York: Wiley. 259 p.

Snedecor, George W. and William G. Cochran. 1989. *Statistical Methods*. Ames, Iowa: Iowa State University Press. 503 p.

Thomke, Stephan H. 2003. *Experimentation Matters; Unlocking the Potential of New Technologies for Innovation*. Boston: Harvard Business School Press. 307 p.

University of Chicago Press. 2003. *The Chicago Manual of Style*. Chicago: University of Chicago Press. 956 p.

US Department of Defense. 2001. "Quadrennial Defense Review Report: 2001." 71 p. http://www.defenselink.mil/pubs/qdr2001.pdf

US Joint Staff. 2000. "Joint Vision 2020." US Government Printing Office.

Zachman, J.A. 1987. "A Framework for Information Architecture." *IBM System Journal* 26(3):276-292. http://www.research.ibm.com/journal/sj/263/ibmsj2603E.pdf

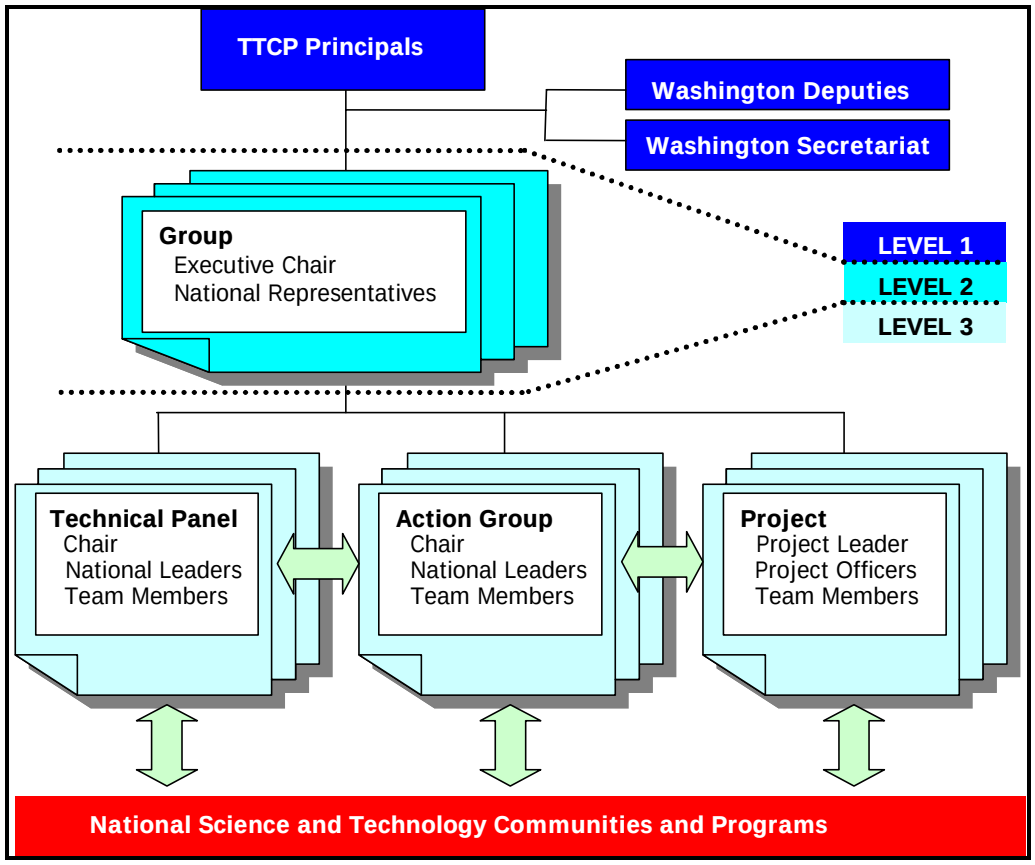
Annex D: Distribution List

All TTCP
Selected NATO Groups
CCRP
CD&E and OT&E Communities
Concerned Allied Organizations

Index

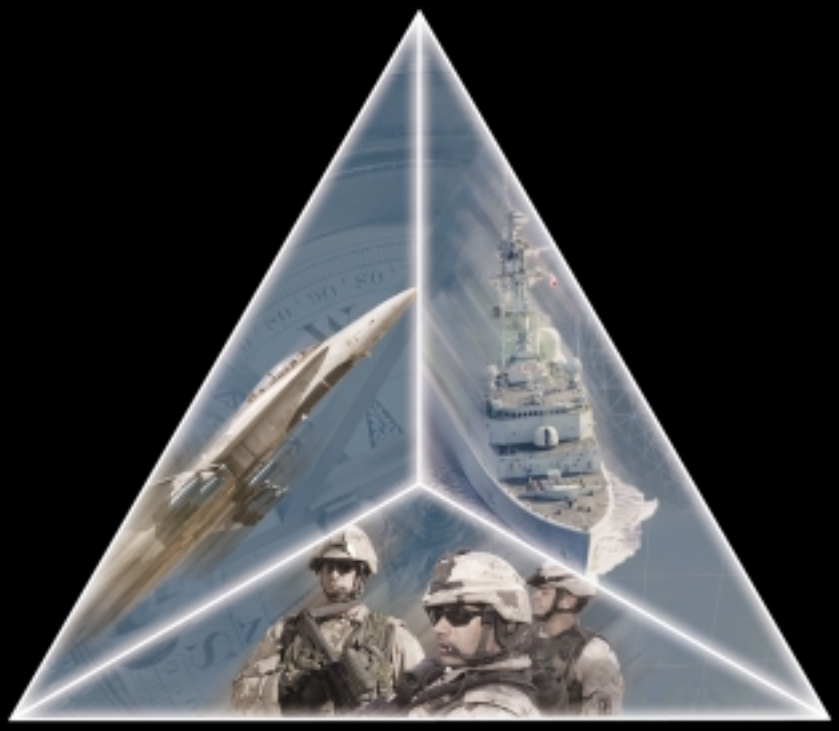
TTCP Document Feedback

The aim of TTCP is to foster cooperation within the science and technology areas needed for conventional (*i.e.*, non-atomic) national defense. The purpose is to enhance national defense and reduce costs. To do this, it provides a formal framework that scientists and technologists can use to share information among one another in a quick and easy fashion. Its structure is illustrated below:



More information on TTCP can be found on its public Website at http://www.dtic.mil/ttcp/

For the purpose of maintaining and updating TTCP unlimited distribution documents (publications that, due to their value to the academic, scientific and technological communities, are widely distributed) readers and users of these documents are invited to email their appreciation, comments and suggestions for future editions to **ttcp_docfeedback@dtic.mil** ttcp_docfeedback@dtic.mil This address is administered by the TTCP Washington Staff, who will pass feedback onto the appropriate document point of contact. For more information on TTCP document feedback, please see the TTCP guidance document 'POPNAMRAD', which can be found on the public website.



The thesis of GUIDEx is that robust experimentation methods from the sciences can be adapted and applied to military experimentation and will provide the basis for advancements in military effectiveness in the transformation process.