CRM D0010079.A3/1Rev
June 2004

The Art of Military Experimentation

Brian McCue

CNA

4825 Mark Center Drive • Alexandria, Virginia 22311-1850

Approved for distribution: **June** 2004

Steven W Klein

Helen T. Karppi, Director (Acting) for
Expeditionary Systems & Support Team
Integrated Systems and Operations Division

This document represents the best opinion of CNA at the time of issue.
It does not necessarily represent the opinion of the Department of the Navy.

Approved for Public Release; Distribution Unlimited. Specific authority: N0001 4-00-D-0700.
For copies of this document call: CNA Document Control and Distribution Section at 703-824-2123.

Contents

This page intentionally left blank.

Preface

> I am sending you a gift which, even though it may not match the obligations I have to you, is, without a doubt, the best that [I] can send to you; in it I have expressed all I know and all that I have learned from long experience and continuous study of worldly affairs.
>
> —Niccolò Machiavelli

The present document is based in part on my own four-year intellectual journey as an analyst at the Marine Corps Warfighting Laboratory (MCWL), during which I learned a great deal from the Marine officers and enlistees assigned there or attached during experiments. Part of the Lab's mission was to learn how to do military experiments, and the examples cited herein—positive and negative—are selected because they are instructive in this regard. The discussion of errors is meant only to improve future experimentation, and not to detract from the work of MCWL and those who have served there.

An *intellectual* is somebody who is *excited by ideas.* In the course of my four-year assignment to MCWL, I met Marines of every rank (unless I missed one of the levels of warrant officer), from newly joined privates to four-star generals. One thing I noticed was that *every Marine is an intellectual.*

This page intentionally left blank.

Introduction

> If, instead of sending the observations of able seamen to able mathematicians on land, the land would send able mathematicians to sea, it would signify much more.
>
> —Isaac Newton

This paper is part of CNA's project on military experimentation. The project's products are:

- *The Art of Military Experimentation (this document)*
- *The Practice of Military Experimentation,* and
- *Wotan's Workshop: Military Experiments Before World War II.*

The different products are intended to serve different readers' purposes. The newly assigned civilian analyst might want to start here, with *The Art of Military Experimentation*. The military officer (active duty or otherwise) newly assigned to an organization devoted to military experimentation, is advised to start by reading *The Practice of Military Experimentation*. Either should then read *Wotan's Workshop*, and then the other's starting point.

This document, *The Art of Military Experimentation,* looks at the thinking behind such experiments and explores the question of how to conceive of and plan experiments that will be fruitful of information, taking the execution of the experiments somewhat for granted. The companion *The Practice of Military Experimentation*, by contrast, focuses more on the execution of the experiments, with the intellectual underpinnings examined only to the (considerable) extent that is necessary for good execution. *Wotan's Workshop* presents pre-World War II examples of American and German experiments as case studies.

A conceptually related, albeit organizationally separate, effort has produced

- Analysis Planning for a Domestic Weapon-of-Mass-Destruction Exercise.

It stands alone, and the person working on such a project need not read any of the other products. However, the methodology-fancier who likes *The Art of Military Experimentation* might want to look at it as well.

The Practice and *The Art* embody considerable parallelism; certain themes sound in both documents, but in markedly different ways that reflect the difference in their intended audiences: that of *The Practice* is the major or lieutenant commander newly assigned to an outfit devoted to military experimentation, whereas that of *The Art* is the newly assigned analyst. For example, this document presents, below, “the key to the art of military experimentation,” whereas *The Practice* presents something else at the corresponding point, declaring it to be “the key to the practice of military experimentation.”

The Art, having been written second, contains specific references to *The Practice*, whereas *The Practice*’s references to *The Art* are not specific because *The Art* had yet to be written when *The Practice* was published. To remedy this lack of specificity, the appendix of *The Art* lists where references made in *The Practice* may be found in *The Art.*

The various products share some examples, but use them to different ends. In fact, examples are central to the entire project, which was undertaken partly because the existing how-to works on military experimentation are largely devoid of examples. *The Art of Military Experimentation* refers to about three dozen separate military experiments. About half of those are in the personal experience of the author from his time at MCWL. The remaining examples come from history.

The key to the art of military experimentation

The Practice of Military Experimentation explained that the key to the *practice* of military experimentation is that an experiment is not an exercise; the key to the *art* of military experimentation is that an experiment benefits from having a firm grounding in *theory*.

Such theory is usually separate from any hypothesis that the experiment is intended to test. The parallel is to the test of a physical piece of equipment: the hypothesis is that the equipment will work (or perhaps that it will fail), a point on which people might differ, but the test is designed according to a physical theory (e.g., Newtonian physics, or Maxwell's equations) whose validity is not at issue.

The plan of the work

Pursuant to the above, *The Art of Military Experimentation* is structured around a set of discussions of theory.

The initial chapter (following this introduction) is a discussion of experimentation, drawn largely on familiar examples from physics but with an emphasis on *thought experiments*. The idea of the "operational experiment" is introduced and underscored with an example from social science—Stanley Milgram's exploration of what he called the "small world," now better known as the supposed "six degrees of separation" between randomly chosen Americans.

The second chapter is a review of the taxonomy of military experimentation as it is practiced today. This review touches on the "Limited Objective Experiment" and the "Advanced Warfighting Experiment." It also discusses two forms of military experiment not always mentioned: the military thought experiment, and the experiment undertaken during actual military operations. Examples are given throughout. This chapter's structure is identical to the corresponding chapter of *The Practice of Military Experimentation*, but the content is quite different. In particular, the previously introduced notion of "operational experiment" is invoked where applicable, and the examples are different.

The third chapter, "Models, reality, theory, realism, garbage, and truth," covers some of the same ground as the sixth chapter of *The Practice* ("Accuracy, realism, fidelity, reality, truth, and cheating"), but—because of the difference in intended audience—from a wholly different standpoint.

Useful military theory being scant, the next chapter describes a family of theories that have been useful in experimentation, and to how the analyst might set about creating or finding other bodies of theory if those listed prove inadequate to support a particular experiment. A pattern emerges: well-developed theories lead to true measures of effectiveness (MOEs), vice the measures of performance (MOPs) that are all too often used in MOEs' place.

There follows a chapter on methodology, despite the fact that the entire work is, in some sense, about methodology. This chapter, too, has an analogue in *The Practice,* but that version is shorter and devoted to "methods." The distinction between "methods" and "methodology" is deliberate, and mirrors the distinction between the intended audiences of the two chapters.

A chapter-length example of a military experiment and its analysis is then given, to illustrate the points made so far.

The next chapter, "Why Military Experimentation Is So Hard," has some thematic overlap with the chapter "Obstacles to Successful Military Experimentation" in *The Practice*, but is different in content.

The final two chapters are unabashedly prescriptive. The first is on how to write reports about military experiments; the second is about how to organize a command devoted to military experimentation.

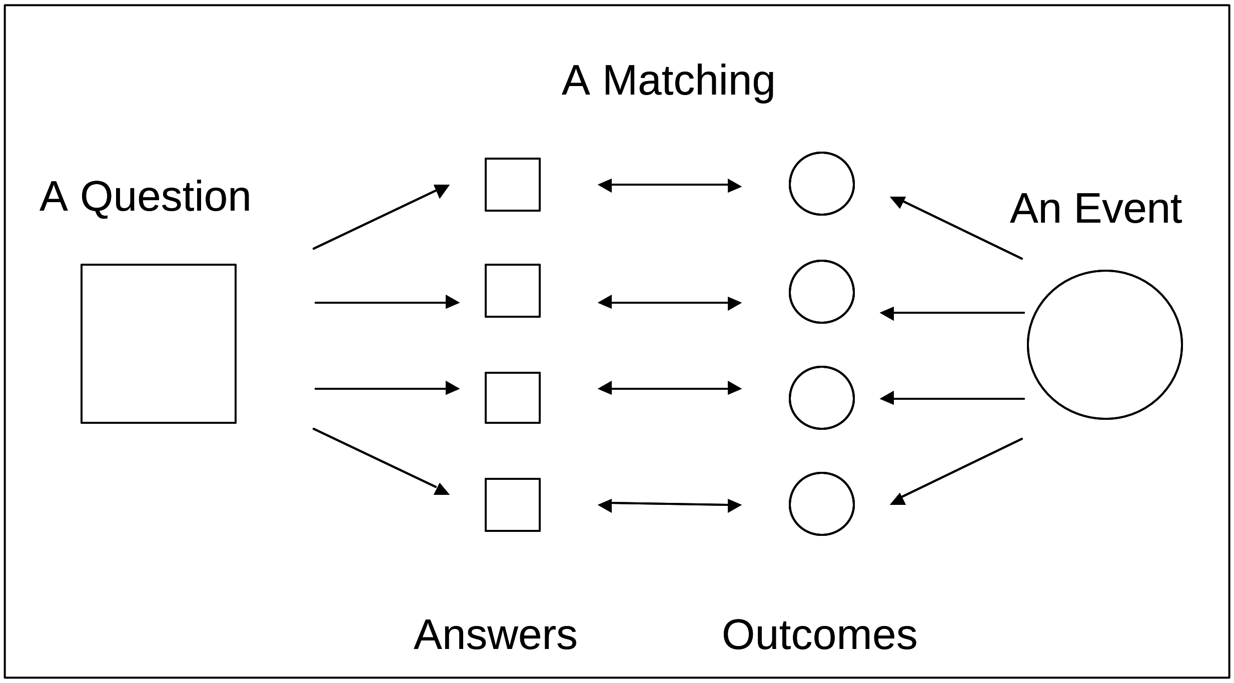
Experiments

> This idea of operational experiments, performed primarily not for training but for obtaining a quantitative insight into the operation itself, is a new one and is capable of important results.
>
> —Philip Morse and George Kimball

As shown in figure 1, an experiment consists of:

- An *event* that could turn out in any one of several ways,
- A *question* that could have any one of several answers, and
- A *matching*, normally pre-stated, between the outcomes of the event and the answers to the question.

Figure 1. Schema of an experiment



A familiar example is the use of litmus paper to test the pH of a sample. The *event* is that the litmus paper is dipped into the sample and turns color. The multiple outcomes are that it can turn either of two colors. The *question* is, “Is the sample an acid or a base?” The *pre-stated matching* is that the color red indicates an acid whereas the color blue indicates a base. This matching determines the answer.

Note that this account of experimentation does not require an experiment to have a hypothesis, a control group, a statistically valid number of trials, or any of the other trappings sometimes associated with experiments. An experiment *may* have some or all of these things, but if it does, they are part of the definition of the set of outcomes, and the matching of the outcomes to the answers.

Given this scientific outlook, one might wonder why the title of this paper refers to the “art” of military experimentation—if it’s so scientific, why is it an art?

The reason is that in military experimentation[1] a large number of real-world influences act on the experiment, preventing the experimenter from doing exactly what he or she would like. Therefore the problem must be worked from both ends: the experiment must be designed to fit the question, but the question may also have to be adjusted so as to fit the experiment.

In this process, two important traits must be retained:

- There are multiple possible outcomes, not just a single, guaranteed outcome.
- The matching between event outcomes and answers to the question is pre-assigned.

If there is only one outcome, or if there are multiple outcomes but they are indistinguishable, the event is a *demonstration*, not an experiment. If the meaning of the outcome is determined only after

[1] And in most other kinds as well, except *perhaps* the most “scientific” *and* well funded.

the experiment is over, it is an *exploration*, not an experiment. Demonstrations and explorations can be of value, but they are not experiments.

Sorensen points out that the "answers" could themselves be questions—an important point, though he does so in the context of a definition somewhat different from the one presented here. His definition of an experiment is "a procedure for answering or raising a question about the relationship between variables by varying one (or more) of them and tracking any response by the other or others" [1].

Thought-experiments

Thought-experiments date back to ancient times—examples can be found in the works of Plato—but the term was coined by the physicist-philosopher Ernst Mach. Sorensen (p. 205) defines the term:

> A thought-experiment is an experiment that purports to achieve its aim without benefit of execution.

It is natural to question the idea that anybody could learn anything about the world purely by reflection. In terms of our definition of "experiment" (or almost any other, except perhaps Sorensen's own, above) this definition seems almost to be a contradiction in terms: without the "activity" cited in our definition, how can there be any outcome, and thus any indication of a particular answer?

But the "activity" can, in fact, be purely mental. For example, a thought-experiment can be a useful way to cast doubt on a theory.

A thought-experiment of Galileo's

We are told that Galileo disproved Aristotle's theory of gravitation—in which heavier objects fell proportionately faster—by simultaneously dropping two cannonballs, of different weights, from the tower of Pisa and noting that they hit the ground at the same time. Yet he cast doubt on the theory, at the very least, via a *thought-experiment*. Given three cannonballs of light, medium, and heavy weight, suppose that the lightest and heaviest are connected via a strand of thread: now forming a single object, how fast are they expected to fall? [2] The possible outcomes would seem to be that the assembly of balls:

1. falls as fast as the light and heavy balls of which it is composed;
2. comes apart because its constituent balls fall at different speeds; or
3. falls even faster than the heavy ball, because the sum of the light and heavy balls' weights is greater than that of the large ball alone.

But in the first case, Aristotle is proven wrong because objects of different weight are falling at the same speed. The second outcome is contrary to experience because it says, in effect, that objects made of parts with different weights come apart while falling, which we know not to be true: many things have parts of different weights and remain intact while falling. And the third outcome, which is what Aristotle's theory would, strictly speaking, predict, now seems nonsensical because we cannot bring ourselves to believe that a slender thread connecting the two balls could have this effect.

Even though the activity takes place purely in our minds, our understanding changes as a result: we decide that Aristotle's theory, plausible though it sounds initially, simply cannot be true.

Thought-experiments' larger role

Thought-experiments see nearly constant, if unstated, use throughout science, because of the role they play in the planning of other experiments: when planning an experiment, one imagines how it might turn out and what these different eventualities would mean. This entire process amounts to a thought-experiment, in that it is a finding, by pure ratiocination, that the planned physical experiment will be of use [3].

Thought-experiments are also crucial to the generalization of results once an experiment has been done: on Wednesday, we will rely on the results of Tuesday's measurement of the speed of light because we do a thought-experiment to persuade ourselves that the day of the week does not influence the speed of light, or our measurement thereof.

How is it that can we learn without doing? One answer is that a thought-experiment is simply a form of mathematical proof. As in a mathematical proof, we can possess all the ingredients, and yet require some time and effort to assemble them in a new way, whereupon we learn something that we did not previously know.

Physical experiments

Inasmuch as military experiments are often considered to be analogous to the experiments of physical science, let us look in detail at some famous experiments in physics. These will be used as points of reference later on.

Galileo's weight-dropping experiment: proving a point

According to legend, at least, Galileo demonstrated a falsehood of Aristotelian physics by dropping balls of unequal weight from the tower of Pisa: released at the same time, these hit the plaza below at the same time, falsifying the belief that objects fall at speeds proportional to their weight.

This experiment is instructive in its simplicity: it has no quantitative measurement, no statistics, and no identification of a "control" or "baseline" case and an "experimental" case. It is simply a comparison, set up so as to prove a point.

The Cavendish experiment: a measurement

Isaac Newton's posited Law of Universal Gravitation held that the force holding celestial bodies in their orbits and the force that makes terrestrial objects (e.g., apples) fall down were one and the same. Newton had also shown that such a force, with a strength proportional to the product of the bodies' masses and to the inverse square of the distance separating them, would explain the observed motion of the Moon about the Earth. But, lacking an independent measurement of the masses of the Earth and Moon, he could not establish the constant of proportionality.

In 1798, more than 100 years after Newton recorded his experiments, Lord Henry Cavendish measured the gravitational constant directly (or, more poetically and not incorrectly, “weighed the Earth”), by measuring the attraction between masses in his laboratory. He did so by assembling two baseball-sized lead spheres into a dumbbell-like configuration, hanging it horizontally by its center, and measuring the torque created by bringing two basketball-sized lead spheres nearby.

With this apparatus and a great deal of care, Cavendish was able to make a measurement of the gravitational constant that came within a few percent of the presently accepted value.

The Michelson-Morley experiment: a negative result

By the 19th century, a great number of similarities had been noted between the behavior of light and the behavior of waves. (For example, light, like waves, can be reflected, refracted, diffracted, and made to create interference patterns; also, there were strong experimental reasons to think of brightness as analogous to the amplitude (i.e., height) of a wave, and to think of color as corresponding to wavelength.)

Waves, however, propagate in some medium, such as air or water, and so the theoreticians posited the existence of “ether” as an all-pervading fixed medium in which light waves propagated. Even the vacuum of space somehow supported this ether, since starlight shines through it. The remaining piece of unfinished business was to detect the ether itself. Michelson reasoned that the Earth, by moving through space, must experience an “ether wind” like the wind felt when extending a hand out the window of a moving vehicle.

To detect this wind, and thereby the ether, Michelson simultaneously measured the speed of light along two perpendicular axes. He later explained the reasoning behind this method by means of a thought-experiment:

> Suppose we have a river of width *w* (say, 100 feet), and two swimmers who both swim at the same speed *v* feet per second (say, 5 feet per second). The river is flowing at a steady rate, say 3 feet per second. The swimmers race in the following way: they both start at the same point on one bank. One swims directly across the river to the closest point on the opposite bank, then turns around and swims back. The other stays on one side of the river, swimming upstream a distance (measured along the bank) exactly equal to the width of the river, then swims back to the start. Who wins? [4]

The first swimmer will win, as can be verified by drawing a vector diagram. With his partner Morley, Michelson set up an experiment to make light "race" in the supposed ether along perpendicular axes, but could detect no difference in the speed of light along the two axes. They therefore concluded that there exists no ether after all. This negative finding was important in leading Einstein to dispose of the ether-provided fixed and absolute frame of reference along with the ether itself, and to think of light as possibly not being completely wavelike after all [5].

Rutherford discovers the nucleus: an instance of serendipity

In 1911, Ernest Rutherford and his associates shot alpha particles at sheets of gold foil, and other foils. Based on earlier experiments with sheets of mica, they expected deflections of a degree or two. But some particles seemed to be undergoing much larger deflections and, after considering various forms of error that might have caused these, Rutherford instructed his assistant to set up an experiment that would look for particles bouncing back off the sheet instead of penetrating. There were some. As Rutherford later explained by using a naval metaphor,

> It was quite the most incredible event that has ever happened to me in my life. It was almost as incredible as if you fired a 15-inch shell at a piece of tissue paper and it came back and hit you. On consideration, I realized that this

> scattering backwards must be the result of a single collision, and when I made calculations I saw that it was impossible to get anything of that order of magnitude unless you took a system in which the greatest part of the mass of an atom was concentrated in a minute nucleus [6].

It was known that atoms include electrons, which carry negative electric charge and might therefore interact with the positively charged alpha particle, but since each electron only weighs about one 8,000th as much as an alpha particle, electrons could not be expected to deflect alpha particles very much. It was reasonable to conclude that the reflections were being caused by interaction with a positively charged region of the gold atoms. But to explain the sharp angles of reflection, this region would have to be very small, very massive, and very charged. Rutherford therefore, as described above, dubbed it the "nucleus": only a thousandth the radius of the atom, it contains almost all the mass, and all of the positive charge. This notion was contrary to the pre-existing notion (based on the views of J.J. Thomson), which held that the positive and negative charges were each spread throughout the entirety of the atom.

Rutherford's discovery was an instance of *serendipity*; he found something for which he was not looking.[2] But the discovery was not wholly accidental, either. Rutherford was bombarding gold and looking for backwards-scattered alpha particles because he had found wide (but not backward) scattering in earlier tests on gold, and he was bombarding gold because he had seen 2-degree scattering when bombarding mica.

Operational experiments

The meaning of the term "operations," as used in the phrase "operations research" (or that of the term "operational," as used in the British term for the same thing, "operational analysis") has little in

[2] "Serendipity," the quality of finding important things for which one is not looking, is named after Serendip (now known as Sri Lanka), because it figures prominently in the folk tale "The Three Princes of Serendip."

common with the military meaning, which pertains to the level of war that lies between the tactical and the strategic.

As Morse and Kimball write, when attempting to define "operations research" at the beginning of their book *Methods of Operations Research*,

> The word "operations," in the definition, itself requires definition. Its use in military terminology is quite specific, but this usage differs somewhat from that current in industrial or other activities.[3]

Regrettably, Morse and Kimball go on to say, "A specific definition will not be attempted this early in the text," and continue the discussion without supplying a definition.

The physicist Percy Bridgeman defined the term "operational" as denoting that which can be described in terms of physical acts and measurements: this is the meaning of the term that is used in the familiar phrase "operational definition," and it is certainly closer to the operations-research meaning of the term than is the military definition.

For the purposes of defining "operations research," we may define an "operation" as an event in which one physical entity acts upon another. "Operations research," therefore, is the scientific study of these acts, as distinct from the study of the entities, or even the events.

As an example, consider the early operations researchers' work on the sighting of submarines from aircraft. These workers put little effort into the study of aircraft or submarines, but they put a great deal of effort into the study of the sightings: the distances at which they were made, the relative bearing of the submarine from the aircraft and vice versa, the altitude at which the aircraft was flying, and so on.

[3] Morse and Kimball, page 2.

We may define an "operational experiment" as one that seeks to establish some fact, or measure some quantity, by subjecting well-understood objects to well-understood operations. This use of the word "operational" is not the same as that used by the military in referring to, e.g., "the operational level of war."

One example is a meteorologist's experiment to measure the velocity of winds aloft: a small balloon is released, allowed to rise, and tracked by theodolite, or radar. Despite appearances, this is not an experiment about the flight of lighter-than-air balloons, and in fact it presupposes a good deal of knowledge regarding this topic because the meteorologist needs to allow for the buoyancy of the balloon before drawing any conclusions about updrafts or downdrafts.

Before probability theory was well developed, but after the development of card-game-based gambling, professional gamblers did *operational experiments* in which thousands of hands of cards were dealt and the distribution of various configurations (e.g., "full house"—two of one kind and three of another) were tallied. These were not experiments about cards or about dealing—they were experiments about hands and deals, and therefore about probability, done using cards. They would have been just as applicable to card games had they been done with, say, sets of balls rather than cards.

Milgram measures the "social distance"

As an example of an operational experiment, we can consider the 1967 experiment by which Stanley Milgram sought to measure the "social distance" between randomly chosen individuals. In it, persons received a package and a short description of the intended recipient. If they knew the recipient, they were to mail him the package; if they did not know him, they were to mail the package to somebody who might. At each stage, a postcard was to be separated from the package and mailed to Milgram so that he could track the package.

Milgram's objective in this experiment was to measure the "social distance" between randomly chosen individuals, now familiarly called the "degrees of separation," an expression Milgram did not use. His interpretation of his result was that strangers in America are separated by an average of six degrees of separation. Later arguments over his methodology (e.g., his non-treatment of cases in which the package never arrived, and countervailingly, of the idea that the package might not follow the shortest path) need not concern us yet, though we shall return to this topic later.

For us, the point of interest is that although Milgram's experiment used the mail and packages, Milgram was not investigating the U.S. mail system, and, although his experiment asked people to perform a task, he was not investigating how well they could do so—indeed, he was assuming that they could do so perfectly, given their sets of friends. Rather, he was using and measuring these mailing operations so as to investigate how people's networks of acquaintances are interconnected.

Artificial worlds

Experiments with simulation-like "agent-based," "artificial life," or "artificial world" computer programs, such as those of Andrew Ilachinski, [7] Joshua Epstein and Robert Axtell, [8] or the Marine Corps Combat Development Command (MCCDC) Project Albert, are operational experiments. At a simpler level, so is the sociological chessboard-and-coins experiment of Thomas Schelling, [9] in which pennies and nickels on a chessboard are moved via a chance mechanism that is slightly biased towards adjacency with like coins. Even the slightest bias leads to a segregated chessboard in remarkably short order, with stark implications for the goal of integrated neighborhoods. I call these experiments "simulation-like" because they have the trappings of simulations (e.g., representation of time, space, and entities, and in some cases a random element), but they are not attempts to reproduce any particular real-world situation in detail. The goal of these experiments is not to find out about the computer or the chessboard, or even the program or rules governing what happens in them, but to create results that are extensible beyond the particular computers and programs (or chessboards and rules) that the experimenters use. This aspect may explain the difficulty experienced by some in understanding the point of these experiments.

Some of the problems addressed by these computer programs are "computationally irreducible." Of course, faster computers or cleverer programming may be able to speed the process as measured by the clock on the wall, but the problems are “irreducible” in the sense that the final state cannot be calculated by any approach other than step-by-step, simulation-like unfolding of a series of states according to a set of rules; for example, there is no equation that will give the answer. Remarkably, the presence of this characteristic in some problems can be proven mathematically; it is not just a supposition made upon giving up on finding the equation. Other problems may not be truly computationally irreducible, but it is easier to treat them as if they were, setting up and running the program rather than struggling to find the equation that will give the answer in one step. Those who investigated probability by tallying the results of thousands of card-deals were treating a computationally reducible problem as if it were irreducible, because it was irreducible with the mathematics of the time.

Yet a considerable amount of reduction typically goes into formulating these “computationally irreducible” problems: such problems should really be termed “not further reducible,” rather than “irreducible.”

Operational irreducibility

Operational experiments—e.g., the meteorologist’s measurement of the wind with a balloon, or Milgram’s measurement of “degrees of separation” by package-mailing—typically address problems that are *operationally irreducible* in the sense of being *operationally not further reducible*. Part of the skill in doing such experiments lies in perceiving a convenient and practical set of operations that will shed light on the topic of interest, e.g., by Schelling when he reduced race relations to a chessboard, some coins, and some made-up probabilities.

Seen in this light, the thought-experiment is simply the limiting case of operational reduction, in which the point of irreducibility is not reached until *all* physical operations have been reduced away and all that remains is the thought process.

Military operational experiments

Most military experiments are operational experiments, in the sense introduced in the previous section. The exceptions are Limited Technical Assessments (LTAs), which, like field tests of equipment, are devoted to measurement. Note again that this usage of the term "operational" reflects the style of experimentation and has no connection with the same word as used in the term, "operational level of war."

In the case of military experiments, one aspect of reduction is the size of the scenario. One example is seen in the experiments of MCWL's Urban Warrior, the second of the three phases that made up MCWL's 1995-2001 "Sea Dragon," experimentation plan to improve Naval Expeditionary capabilities. Urban Warrior had only a company of "Blue" Marines, even though a real Marine operation would involve at least a battalion from a MEU and very possibly more. The thought was that we could validly reduce the scope to a single company and generalize from there, but that further reduction would threaten the validity of the experiment.

Another aspect of reduction is to eliminate processes that are extraneous to those of interest. In MCWL's urban experiments, the utility of supporting fires was of interest, but the process by which the fires were targeted was not.[4] Therefore the only two aspects of fires that were represented were the call for fire, which was done by radio to Experiment Control, and the dispersion with which the fires would arrive, which was simulated by a dice-throwing process done in Experiment Control. Later, in MCWL's Capable Warrior, the third phase of Sea Dragon experimentation, the process of targeting and coordinating fires *was* of interest and was represented accordingly, but the effects of dispersion were not of interest and the dice-rolling method of dispersion was accordingly not needed.

MCWL's Hunter Warrior AWE (Advanced Warfighting Experiment), the first phase of Sea Dragon, was an operational experiment with a radically new style of expeditionary fighting: squad-

[4] There was interest in "squad leader call for fire," but this topic was purged from experimentation, as recounted elsewhere in this document.

sized teams, operating independently, would attack enemy concentrations by calling in precision strikes of artillery, air support, and/or naval surface fire support.[5] As such, it had the squad-sized teams, the enemy concentrations, and a command-and-control system by which the strikes could be called in, but the strikes were purely notional and the equipment in the command-and-control system was built by combining commercial off-the-shelf (COTS) radio, personal data assistant (PDA, then in its infancy) and Global Positioning System (GPS) devices. Many skeptics looked at the hardware and complained that it was too delicate and non-secure and that no experiment was needed for them to see this; however, but they missed the point that the hardware was a mere surrogate, used to facilitate the *operation* that was the experiment's true subject.[6]

Serendipity in military experimentation

Because "serendipity" is the quality of discovering things that one does not expect to find, it is not clear that it can be pursued as a conscious strategy—but military experimenters often attempt to do so. They look upon experimentation as "going out and fooling around," and expect that by doing so they will discover something.

Although serendipitous discoveries have been made in this way, it is not a good recipe for experimentation. At best, it is inefficient, because there is no guarantee that anything remarkable will happen, in which case no conclusion will be possible: in an experiment planned along the lines presented earlier, each outcome, even the non-remarkable one(s), points to some answer to the organizers' question. At worst, it creates results that mislead because they are, unbeknownst to the experimenters, results of the experiment's artificialities.

[5] This style of fighting, though widely decried as nonsensical at the time, strongly resembles what was later done in Afghanistan during Operation Enduring Freedom.

[6] Surrogates are discussed at greater length in The Practice of Military Experimentation.

A hierarchy of military experiments

> A battle is won in a certain way; it might very well have been lost. It might have been won or lost in a million ways. … To assess which weapons are better requires in the first place experiment, and little else but that. Battle records don't seem to me to hold the answer.
>
> —Solly Zuckerman [10]

This chapter reviews the current taxonomy of military experimentation, upon which there is remarkably wide agreement; Joint and Service-specific experimentation projects work within a fairly consistent conceptual structure of military experiments, ranging from the LTA, through the LOE, to the AWE. Most would, at least informally, add the war game to the lower end of this hierarchy; we will do so here, and also go so far as to add the *military thought-experiment* at an even more basic level.

Organizations devoted to military experimentation generally see this hierarchy as being not only conceptual, but also procedural: a given topic can be expected to be the subject of an LTA, then to appear in a LOE (along with other topics that have been treated in war games or LTAs), and finally to appear in an AWE.

The Practice of Military Experimentation contains (pp 15-21) a parallel treatment of the topics addressed in this chapter.

Military thought-experiments

Overall, most military thought-experiments have probably consisted of trying to visualize a coming battle and assess, via judgment, experience, or just "feel," which side will win. Based on the fact that most battles have had at least one loser, we must conclude that this form of military thought experiment is not very reliable.

However, thought-experiments can valuably play other roles in military experimentation.

One way in which they are used is, as in traditional scientific experimentation, to plan and then to generalize other experiments.

Other uses come closer to the thought-experiment as discussed in the previous chapter. During the interwar period, for example, the Marine Corps established the Fleet Marine Force as an entity with a charter to engage in expeditionary warfare, and set about the creation of a usable manual of landing operations. Absent any authorities on the topic, the latter effort was accomplished by dint of what today would be termed "brainstorming:" Quantico's schools for officers were devoted entirely to the project, in which each Marine prepared a chronological account of a landing operation. These lists were then subjected to a multistage winnowing process at the hands of ever-more senior officers. [11] This process is remarkable for its bootstrap nature: one of the Marines wrote that the group

> ...approached its subject... about the same as every other committee, with a lantern in one hand and a candle in the other—but neither of these seemed to throw much light on the subject, so we wound up by hiding our lights under a bushel and using the imagination that God gave us to use for this particular purpose. [12]

The result was the famous *Tentative Manual for Landing Operations*, published in 1934.

It is important for analysts to recognize that different types of person think in different ways. For example, while analysts and Marine Corps officers are intellectually similar in some ways—e.g., curiosity, and a compulsion to explain anything they know to anybody who will listen—they are different in others. The creation of a physical set-up seems to benefit officers in ways that do not apply to analysts. Either the analysts can imagine the set-up so well that they can don't need to create it physically, or the officers see something in the physical creation that the analysts don't; in either case, the fact remains that the military officers will get a great deal out of creating a physical set-up when analysts just don't see why it would make a difference.

Perhaps the process of creating the set-up may lead the officers into what is, in effect, a thought-experiment that they would not otherwise do.

For example, MCWL personnel for a time discussed "sensor-to-shooter" fires. The idea came up rather often, but was not developed, or even really defined. Then, the topic appeared as an item to be executed in a live-fire experiment. Faced with the prospect of actually needing to set up and execute "sensor-to-shooter" artillery fire, a group of MCWL officers embarked upon a wide-ranging and yet detailed discussion of the topic, which, over the course of two or three hours, resulted in a definite plan for doing "sensor-to-shooter" fires. Of course, one reason for this success was the obligation to put on the event but an analyst who attended the meeting, and analysts who monitored the shooting from various positions, observed that the officers derived a definite intellectual benefit from the process, though the analysts did not.

This example shows a possible benefit of experimentation, and it also shows that the analyst must refrain from minimizing the importance of these physical set-ups; the military officers find them quite valuable, and their viewpoint must be respected.

The Practice of Military Experimentation gives additional examples of military thought-experiments [13].

A question for discussion

A scientist working for the Air Force in the early 1960s told his young son that that his laboratory was working on computers that could play games, as a prelude to building computers that could make tactical decisions regarding air combat. Already, he said, they had developed a computer that could play tic-tac-toe: it was so good that it could always win if it had the first move, and it could always be assured of getting at least a draw if it had the second move.

"That's impossible!" the small boy exclaimed. How did he know this?

War games

War gaming has a long history; it has been suggested that chess has its origins in some kind of military game. War games can usefully be divided into tabletop war games, their computerized cousins, and seminar war games.

War games are kin to thought-experiments in that they do not involve actual equipment, terrain, or military units. The key difference is that the war game, being a game, embodies more thoughts than those of a single person, either by bringing several people together or, more recently, by adding the "thoughts" of a computer. The war game may also embody considerably more formalism, though the "seminar war game," discussed below, is an exception. War games are necessarily *operational experiments* in the sense discussed in the previous chapter, because they explore no unknowns in equipment or personnel performance.

Tabletop war games began to be used for the training of staffs in the 19th century. In a tabletop war game, the playing surface of the table serves as the game board, and is a military map; the playing pieces represent military units. The scale of the mapboard and of the units is chosen to suit the action being war gamed; the board could show an entire continent or a single village, or anything in between, and the pieces could correspondingly be corps, individual troops, or anything in between. In naval war gaming, the units are almost always individual ships. Game rules, in some cases quite complex, govern the movement of the units and their ability to destroy one another. The players' roles, or at least their perspectives, are those of the top one or two levels of command on each side. Several books usefully trace the history of this practice well into the 20th century, e.g., those of Allen, Perla, and Wilson. It is interesting to note that some people play such games for fun.

The war games of Admiral Dönitz

Germany's Admiral Karl Dönitz developed, during the inter-war period when Germany was forbidden by the Versailles Treaty from having submarines, the new theory of submarine operation that later became known as "wolf-pack tactics." He used a chart-based naval war game in planning his Second World War U-boat campaign against Allied convoys. Later, he wrote:

> In the winter of 1938-39 I held a war game to examine, with special reference to operations in the open Atlantic, the whole question of group tactics—command and organization, location of enemy convoys and the massing of further U-boats for the final attack. No restrictions were placed on either side and the officer in change of convoys had the whole Atlantic at his disposal and was at liberty to select the courses followed by his various convoys.

The points that emerged from this war game can be summarized as follows:

> 1. If, as I presumed, the enemy organized his merchantmen in escorted convoys, we should require at least 300 operational U-boats in order to successfully wage war against his shipping.
>
> 2. Complete control of the U-boats in the theatre of operations and the conduct of their joint operations by the Officer Commanding U-boats from his command post ashore did not seem feasible. Furthermore, I felt that his "on-the-spot" knowledge particularly as regards the degree of enemy resistance and the wind and weather conditions prevailing would be altogether too meager. I accordingly came to the conclusion that the broad operational and tactical organization of the U-boats in their search for convoys should be directed by the Officer Commanding U-boats, but that the command of the actual operation should be delegated to a subordinate commander in a U-boat situated at some distance from the enemy and remaining as far as possible on the surface. I therefore insisted that a certain number of U-boats under construction should be equipped with particularly efficient means of communication which would enable them to be used as command boats.
>
> 3. [The programmed force of U-boats would be inadequate.] [14]

Dönitz recounts that his belief that his adversaries would use convoys “was not generally held, [15] but this belief and the results summarized above were borne out later as the Second World War unfolded in the Atlantic.[7]

Hector Bywater

Honan’s *Visions of Infamy* describes how Hector Bywater seems to have, in a hybrid of war gaming and thought-experimentation, staged mock battles with wind-up ship models as part of the research for his prophetic book *The Great Pacific War*, written in 1925. Much of the story, resembled the ensuing Pacific campaigns, Japanese and American alike, of the Second World War. Perhaps Bywater was an example of a person who benefited from the creation of physical set-ups.

Example of a war game

Nowadays, the concepts of tabletop war gaming have largely been transported into the realm of the computer game, moving the considerable administrative burden of the game onto the computer and allowing flexible graphical displays. As such, the use of this kind of war game merges with the use of *modeling and simulation*, the topic of a later chapter.

During his work at the Operations Research Office, the physicist George Gamow invented a simple war game to be played by analysts. The rules are given by Page:

[7] Dönitz had earlier estimated that in a war with Britain, his submarines would have to sink 2/3rds of a million tons of shipping per month. Germany started the Atlantic phase of the war in 1939 with slightly fewer than 60 ocean-going submarines, but with more coming, and sank an average of about 1/6th of a million tons of shipping per month through the end of 1941. Therefore each U-boat sank a long-run average of (1/6 million)/60 tons per month, and so to sink 2/3rds of a million tons of shipping per month, 240 submarines would be needed—not at all far off from 300.

> [The] game is played with three identical boards, one for each of the players and one for a referee.
>
> The board … represents a tank battlefield by a lattice of hexagons, … some of which are hatched to represent wooded areas of low visibility. The white hexagons represent open fields, and the size of a hexagon represents the "radius of action" of a tank in battle.
>
> Each player starts with ten markers representing tanks at his back line, and "a move" consists in displacing any number of tanks into any of the adjacent hexagons. Each player sees his board only and must infer from the play where his opponent's tanks are located.
>
> If two opposing tanks arrive on adjoining white hexagons, "a battle" is announced by the referee, who spins a coin to decide which tank is eliminated. When a moving tank comes into contact with two enemy tanks simultaneously, it must "shoot it out" first with one of them, and then, if victorious, with the other.
>
> A tank in the woods obtains a clear kill on any tank which moves into an adjacent white hexagon; a coin is flipped to determine the survivor if another tank moves into the same hexagon in the woods. The objective of the game is to kill off all the opposing tanks, retaining the maximum of one's own tanks.[8]

This game is notable for having been designed for analysts rather than for military men (military board games had been in use for training officers for about a hundred years) or for hobbyists (H.G. Wells, Fletcher Pratt, and others had created war games to be played for fun). Gamow's idea was that the manual game would lead to a computerized version.[9]

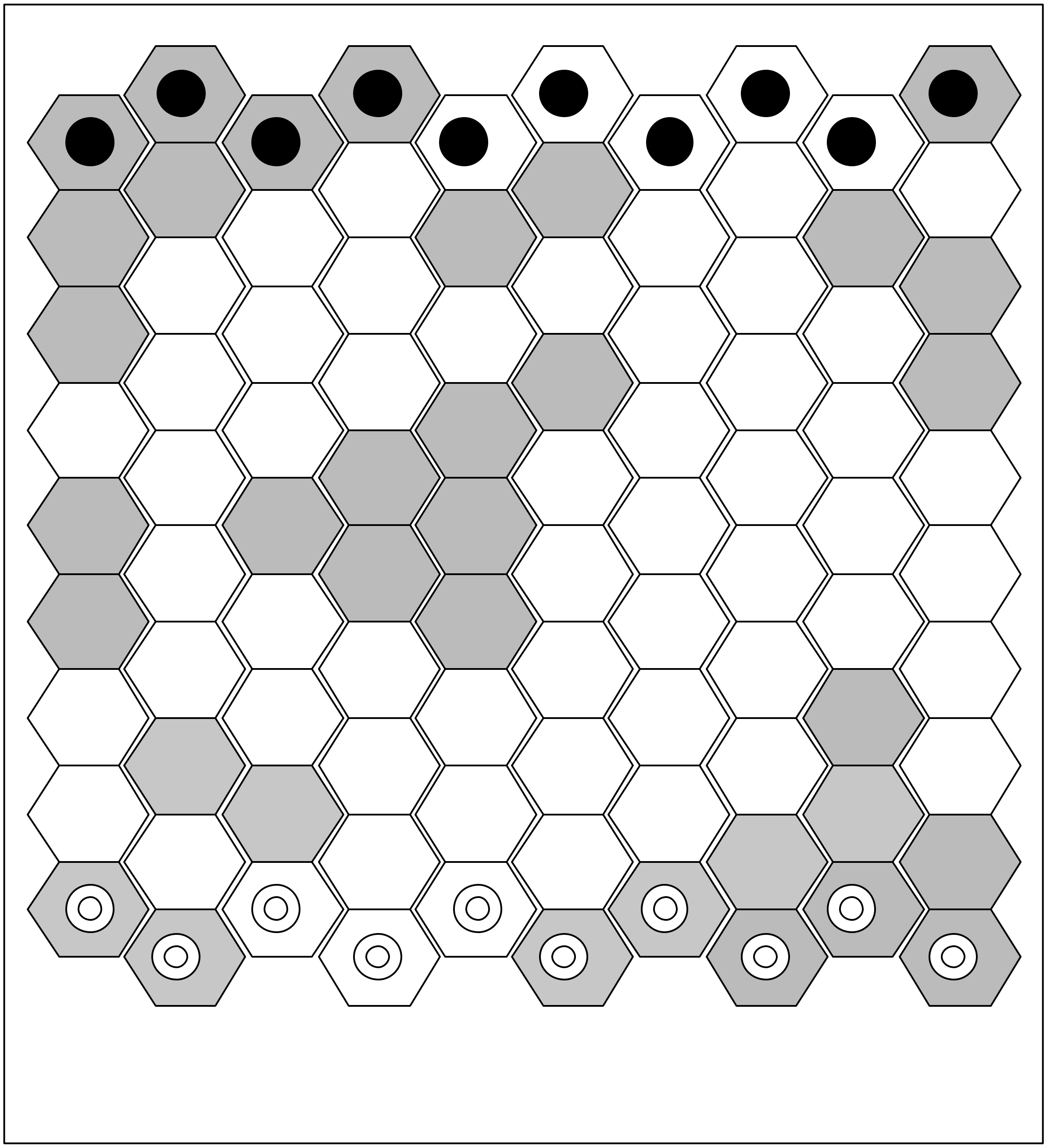
Many war games, computerized or otherwise, are hideously complex, so it is well worth noticing the simplicity of this game, with its 225 words of rules, and a simple diagram to to serve as the terrain. An important aspect of war game design is the choice of scale, and

[8] Page, Thornton. "A Tank Battle Game." Journal of the Operations Research Society of America volume 1, (1952) pages 85-86.

[9] Mirowski, page 362.

much of the simplicity of this game stems from its scale: the distance from hexagon to hexagon is the tank's gun range, and the length of a turn is the time that a tank takes to move this distance (see figure 2).

Figure 2. Tank battle game board, with ten black tanks (top) and ten white tanks (bottom) in starting position—Page op. cit.



Various questions can be posed regarding this simple game, such as:

- The rules seem to take for granted that the “clear kill” mentioned in the final paragraph of rules is automatic. Could there ever be a case in which an advantage could be gained by passing up a “clear kill” opportunity?
- The rules do not specify whether, in the case of the “clear kill,” the victim is told the hexagon from which the clear kill came, and sometimes there will be more than one possibility. Suppose that White is to be given this information and Black is not—what difference does that make?
- Using a six-sided die to give White’s tanks a two-thirds chance (v. the original rules’ one-half chance resulting from flipping a coin) of prevailing in each individual combat (and Black’s a corresponding one-third chance), what reduction in the starting strength can White withstand and still have an even chance of winning the game?
- Suppose that on the referee’s board (only), three tanks on each side are marked as “aces:” whenever an ace tank encounters a normal tank other than in a “clear kill” situation, the ace tank has a two-thirds chance of victory rather than a one-half chance. Knowing this rule, but not knowing which tanks are the aces, ought the players to play differently? What if the aces win five-sixths of the time?
- Suppose that the White side consists of two players, who cannot see each other’s boards (much less those of the referee and the opponent). With how many tanks must each White player start in order for the game to be even?
- Suppose that the White side consists of ten players, each with his or her own board and one tank. With how few tanks can Black start and still have the game be even? How much does the answer change if the White side can confer before the game begins? How much does the answer change if the White players can received advice from an eleventh player, to whom the referee provides a view of the board as it was two turns earlier?

These questions have the interesting property of being impossible to answer on the basis of pure consideration; probably even experienced players of the game would have have trouble answering these questions, and different players would give different answers. Nor are the questions subject to mathematical analysis via game theory or the like: the game seems to be "operationally irreducible" in the sense defined earlier. The only way to answer these or similar questions is by direct experimentation in repeated playings of the game under each of the variant circumstances.

Surely real war is no simpler. The difficulty of using opinion, analysis, or even experience to answer the above questions regarding a simple game suggests the greater difficulty, or impossibility, of using these methods to answer similar questions regarding real combat.

Seminar war games

Seminar war games usually concentrate most of the players onto the "friendly" side, thus moving the focus of the game away from the playing board and into the deliberations of the staff. In such games, the functions of the "enemy" side and of umpiring are merged and the governing rules greatly simplified, with the goal of the seminar game being—as the name suggests—the furtherance of a good discussion as opposed to the generation of detailed force movements. Typically, a few hours of discussion result in a single "move." Its results are resolved rather rapidly and returned to the staff for consideration in formulating their next move.

In Marine Corps parlance, "war game" has come to refer to the kind of structured discussion that would occur in a seminar war game, and Marine Corps war games are designed so that the turn-resolution step occurs only once or twice per game, if at all. Inasmuch as the value of a seminar war game lies in the discussions it causes the participants to have, the Marines have succeeded in operational reduction, stripping the seminar war game to its essence and focusing all their energy on that—*assuming* that the quality of the discussions is unimpaired by the absence of the turn-resolution steps.

Limited technical assessments

LTAs are similar to field tests, but with greater flexibility of procedure. Being technical in nature, they are naturally equipment oriented.

In a traditional test, the personnel are likely to be intimately familiar with the equipment, whereas in an LTA, the personnel are usually Service people who have just been through a day or two of training. Their performance is likely to be more similar to that of the actual users than would be the performance of professional testers or the people who built the equipment.[10]

However, the biggest difference is in the conduct of the experiment. In a traditional test, the goal is to conduct a fair evaluation. To ensure fairness, the test will proceed in a pre-determined way almost regardless of how it is going, with the only exceptions being safety-related. In an LTA, the goal is to learn as much as possible, and if the test article fails in each of the first 15 attempts, there is no point in putting it through another 85: the LTA will be halted, something will be changed, and then the LTA will resume.

Historically, General William “Billy” Mitchell’s famous 1921 ship-bombing experiment was equivalent to an LTA. In it, the decommissioned ex-German battleship *Ostfriesland* was bombed by U.S. Navy and U.S. Army airplanes. The ship sank, Mitchell declared battleships to have been made obsolete by airpower, and the impression stuck. But *Ostfriesland* was dead in the water (i.e., stationary), and thus presumptively easier to hit than a moving target would be, and because she was unmanned, there was no damage control. The latter point was probably quite important inasmuch as the bombing took place over two days, and leaks that started on the first day admitted water unchecked all through the night, leaving the target quite low in the water on the beginning of the second day. Thus Mitchell arguably pursued the process of operational reduction too

[10] Herman Kahn cites an extreme example, in which the German testing of an anti-aircraft gun showed that one in four rounds might be expected to hit; the wartime average was one in 5,000. Kahn ascribes the difference in large part to the test personnel, whom he characterized as “athletes with Ph.D.s in physics.” See also McCue, *Wotan’s Workshop*.

far, reducing beyond the minimum and discarding needed “details” such as damage control and the difficulty of hitting a moving target. [16]

As of January 2004, the MCWL archives contain a few dozen LTA reports.

Limited objective experiments

LOEs outwardly resemble military exercises, but are in fact experiments because their outcomes are not predetermined, and because they are structured so that the possible outcomes will indicate answers to one or more questions.

The LOE is defined by the presence of an Opposing Force, and by free play on the part of at least one side. LOEs can address equipment, tactics, or organization. Because of their size, most LOE events serve more than one experimental goal. It can be difficult to disentangle the sub-experiments from one another—for example, if Blue’s performance improves when it is given a number of futuristic technologies and some new tactics, what made the difference? There may be no way to tell, but there is still value in the finding that the new technologies and the new tactics, taken together, were of benefit. In fact, this finding may be more important than any finding regarding a single item in isolation, because innovation of equipment, tactics, or organization will be used in a future that also contains other innovations, so they are best tested together.

LOEs are *operational experiments* in the sense discussed in the previous chapter. LTA-like sub-experiments may enter into them, but it is important to notice if an LTA is part of an LOE, there will be outcomes of the LTA (e.g., that the equipment fails) that will preclude having any outcome at all from the LOE.

Some LOEs are large enough to encompass combined-arms activity.

As of January 2004, the MCWL archives contain about 20 LOE reports.

Advanced warfighting experiments

AWEs have an opposing force, and at least one-sided (if not two-sided) free play, and are enough larger than LOEs that they certainly encompass combined-arms activity, and may include joint or coalition participants.

AWEs, like LOEs are *operational experiments* in the sense discussed in the previous chapter, though—again—LTA-like sub-experiments may enter into them.

In practice, AWEs are widely found to be beyond the point of diminishing returns to scale in terms of experimental value: MCWL and other organizations have had trouble deriving experimentation benefit from AWEs because the large number of participants, VIPs, media personnel, joint and coalition partners, etc., get in the way of experimentation. An experiment is based on an event that can turn out in more than one way, but the AWE's goals of training, and of maintaining good relations with the VIPs, the public, the media, and the Joint and Coalition partners all militate in favor of having an event that can turn out only one way.

On the other hand, an AWE *can* accomplish experimental goals: MCWL's Hunter Warrior AWE did so.

As of January 2004, the MCWL archives contain reports on four or five AWEs.

Experimentation in real-world operations

Deliberate experimentation can and does occur in the context of real-world operations.

Morse and Kimball give an example regarding a hypothetical (or ostensibly hypothetical) air-launched antiship missile whose success rate has started to decline. Enemy jamming is suspected. Morse and Kimball outline the process of trying the missile with and without an add-on antijam module, noting the proportion of successes in each case, and applying a statistical test to determine whether the difference between the proportions is significant. However, a subtlety arises: assuming that the enemy has been found to be using jamming, experimentation ought to continue in order to be sure that the antijam modules (whose use doubtless imposes a variety of costs) are still needed. Morse and Kimball suggest some formulas for the proportion of non-antijam missiles to use, though these appear to be somewhat ad hoc, and to embody some unstated assumptions.

The general problem of experimenting in real-world operations, military or otherwise, has been identified as the "two-armed bandit problem," in which a gambler is confronted with a slot machine that has one arm on each side instead of the traditional single arm that gives the slot machine its "one-armed bandit" nickname. The two arms presumptively correspond to different payoff probabilities, and the gambler faces a problem of conflicting goals: gaining new knowledge, and capitalizing on the knowledge already gained. Even some greatly simplified versions of the two-armed bandit problem remain unsolved, and the preceding paragraph's missile problem is a good example of how two-armed bandit problems can arise in military situations. Their existence points, in fact, to an important reason for controlled military experimentation: to separate the beneficial process of learning from the painful and costly making of mistakes in combat.

In real-world experimentation, an even greater difficulty arises if one contemplates the possibility that the enemy knows that one is varying one's tactics, and can try to confuse the picture by varying his own tactics—subject, of course, to the costs imposed by the fact that he faces a two-armed bandit problem of his own while doing so. The result leads straight into game theory, and the discussion in Morse and Kimball appears odd until one realizes that, having been written in 1946, it pre-dates most work in game theory as we now know it, having "only" von Neumann and Morgenstern's seminal 1944 *Theory of Games and Economic Behavior* to go on. It is also worth noting that the game-theoretic treatment amounts to a thought-experiment, inasmuch as it is a systematic exploration of "If we do *this* and they do *that* …."

The Second World War operations researcher Solly Zuckerman dealt with bombardment problems in which such considerations as countermeasures and deception did not arise. Interestingly, Zuckerman characterized his Second World War field work in planning and monitoring the bombing of the fortifications on Pantelleria as experimentation:

> The "Professor of Anatomy," as they all knew me, had been offered an opportunity to show how he thought a bombing plan should be designed, and the plan had worked. From my point of view the operation had been an experiment, essentially because it had been possible to check daily whether certain arbitrary but necessary criteria to measure operational results had been achieved, and because it had then become possible to make a direct check of the whole operation. To that extent, Pantelleria was an experiment.[11]

Zuckerman's criteria were, in fact, far from arbitrary. He had, back in England, made estimates of the radii of destruction of various weights of bomb, based on field tests and the results of German bombing. He had used pictures from photographic reconnaissance to make realistic estimates of the accuracy with which bombs could be delivered. In Tunisia, he had a chance to verify these estimates by examining Tripoli first-hand, it having been bombed, and then occupied, by the British. The antiaircraft batteries of the Mediterranean island of Pantelleria were to be neutralized by bombing, and Zuckerman used his estimates of the bombs' destructive radius and accuracy (each much more pessimistic than the beliefs of the Royal Air Force officers, which theretofore had been the only guide) and the Poisson distribution to estimate the required number of bombs. He had daily photographic cover, allowing him to do what we would, today, call "Bomb Damage Assessment," monitoring the whole process and—in particular—updating his estimate of bombing accuracy. As he commented, "Today, this kind of planning would be regarded as elementary. At the time it was entirely novel."[12]

On several occasions, MCWL has sent pieces of equipment with operating Marine Expeditionary Units. The intent has been not so much for the unit to perform formal experimentation in real-world operations as that it would simply try out the gear and report back.

[11] See Zuckerman, page 196.

[12] See Zuckerman, page 187.

Initially, results were disappointing, because the units didn't send in any observations. The coming of e-mail to the Fleet suddenly made communication easier for the deployed Marines, and MCWL started getting the desired feedback. Observations of this system have produced the following guidelines: Each project needs to have its own e-mail address at the MCWL end, so that Fleet Marines can write to it without having to keep track of the possible rotation of the particular person who is reading the messages. Also, feedback will be better on devices that work, because these will be re-used. A device that fails in its first Fleet use will probably not be re-used, so at most one feedback report will be forthcoming. Finally, it must be recognized that the trial-use equipment sent out with operating units is not connected to the supply system, so it has no source of spare parts or maintenance: anything that breaks will simply be discarded.

Models, reality, theory, realism, garbage, and truth

> All models are wrong; some models are useful.
>
> —George Box

As aspects of military experimentation or otherwise, the items enumerated in the title of this chapter may seem quite disparate, but in fact they are intimately connected, as this chapter will show.

Models

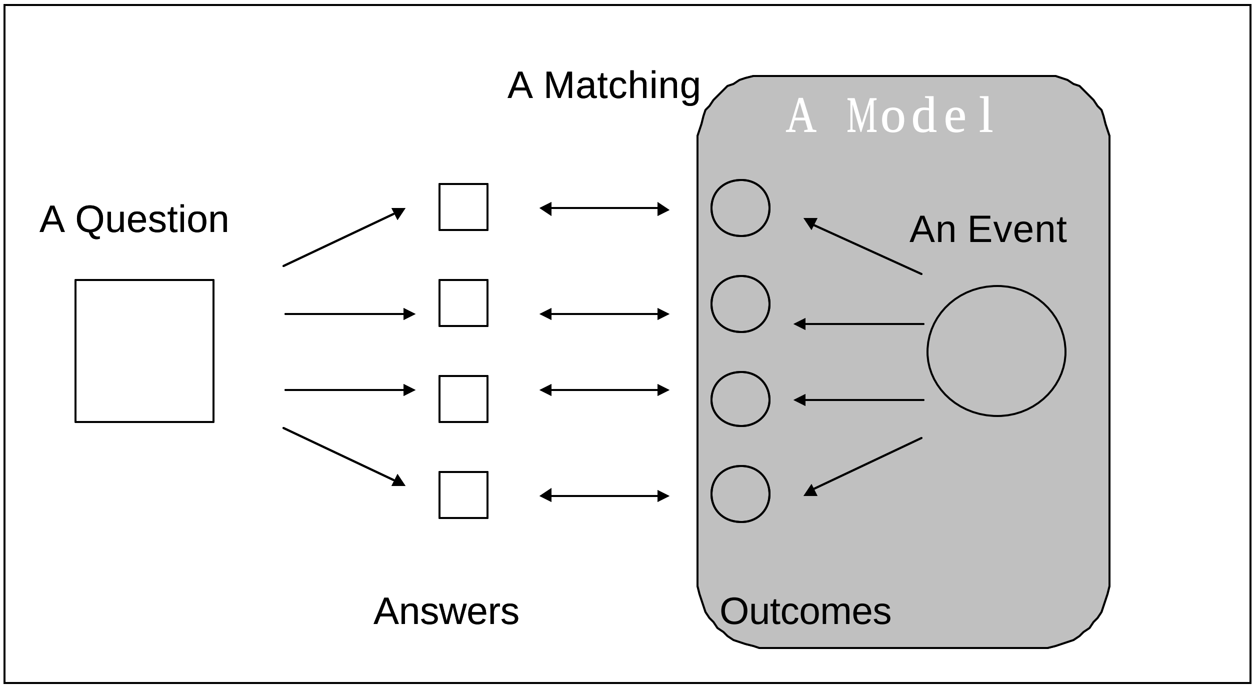
Nowadays, “model” is generally taken to be synonymous with “computer model,” and many tend to look on computer models of warfare with skepticism. If they consider the matter at all, workers in military experimentation see their live experiments as an alternative to the distrusted modeling, when in fact the exercise-like live experiment is a combat model as well—just not a *computer* model. It is important to realize that the activities undertaken in the field, at sea, or in the air are themselves warfare models, albeit not resident in a computer. Like a computer model, this model should be examined critically, and judged on factors other than appearance.

One might say that Galileo, Cavendish, Michelson and Morley, and Rutherford experimented directly on the matter, light, and atoms in which they were (respectively) interested, whereas in military experiments the object of actual experimentation, pursuant to the argument advanced in the preceding paragraph, is a mere *model* of reality, not reality itself. Therefore, an extra layer of inference lies between the military experimenter and the answer to his or her question.

This interpretation would be partially correct, but also partially misleading: a layer of inference does separate the experimental apparatus from reality, but this layer exists in the case of physical science as well. Galileo and Cavendish, for example, were experimenting with gravity and cannonballs because they were interested in finding out about the gravitational forces exerted by the Sun, the Earth, and the Moon. Rutherford was using real atoms and particles, and though one can make the point that Rutherford was interested in finding out about all atoms and particles, not just the ones upon which he was experimenting, that is undeniably a short leap: perhaps the reputation of particle physics as the purest of sciences stems from the fact that its experiments' results need so little generalization before they can be applied. Yet Rutherford also created a model: he suspended one magnet from a long cable and put another one at a fixed point, so that when the suspended magnet swung past the fixed magnet, the repulsion modeled the deflection of the alpha particles by the nucleus. [17]

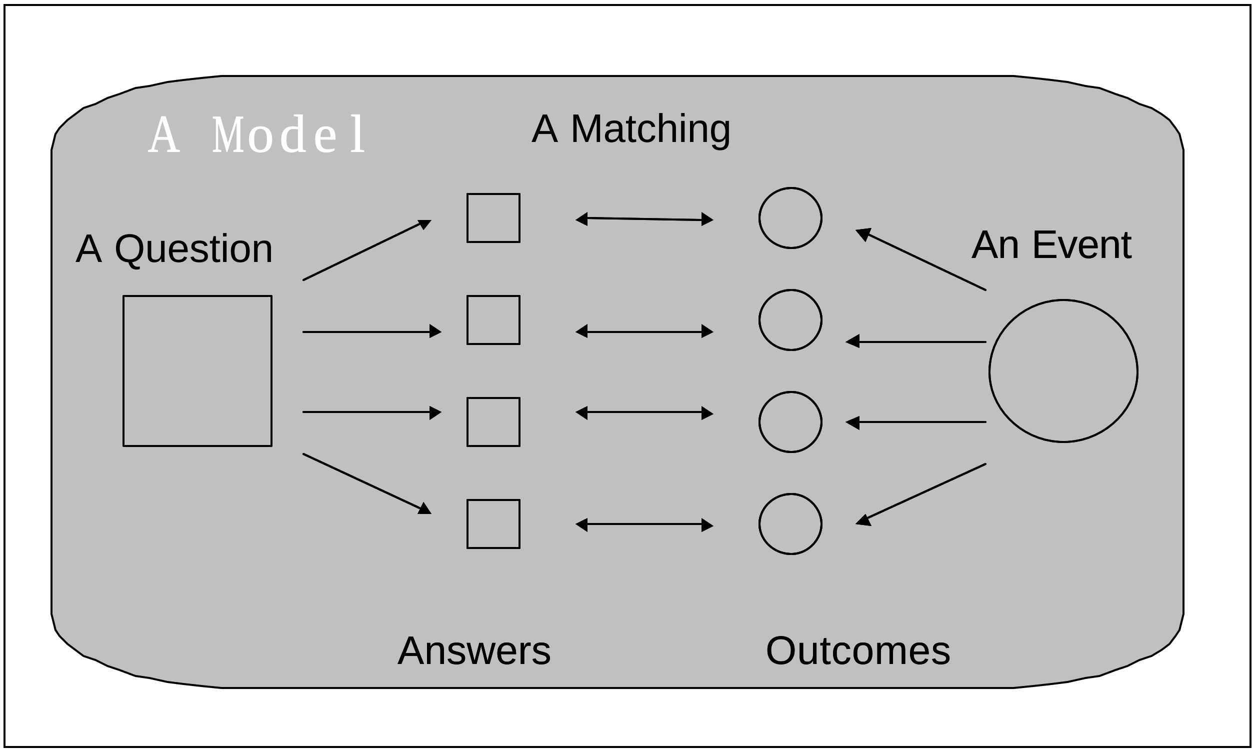
Figure 3 shows how to experiment using a model; the model provides the event, and the outcomes.

Figure 3. Experiment using a model



Of course, one must take care not to become so attentive to one's model that it becomes an object of experimentation in itself. Figure 4 depicts this condition (of which computer-using combat modelers are, not always unjustly, constantly accused of partaking): the question and the answers, as well as the event and the outcomes, are all inside the model.

Figure 4. How NOT to experiment using a model



Even field-tests or LTAs can be considered to use a model, in that the test-range set-up is a model (infamous for optimism) of the real-world battlefield. Only experiments like those of Zuckerman, undertaken in action against the enemy, can be considered model-free.

The preceding discussion, however, neglects another inferential layer: that which, in the case of military experiments as well as those of physical science, often separates the question addressed in the experiment from the question about which the experimenter is really wondering.[13] Just as the experimental apparatus is a particularization of the real subject of interest, the experimental question and its answers are particularizations of the real question of interest. Consider, for example, Billy Mitchell's ship-bombing experiment: Mitchell was not interested in establishing the vulnerability of *Helgoland*-class battleships to 1,000-pound bombs dropped by particular types of Navy and Army aircraft. He was interested in establishing the vulnerability of big warships in general to aerial bombing in general.

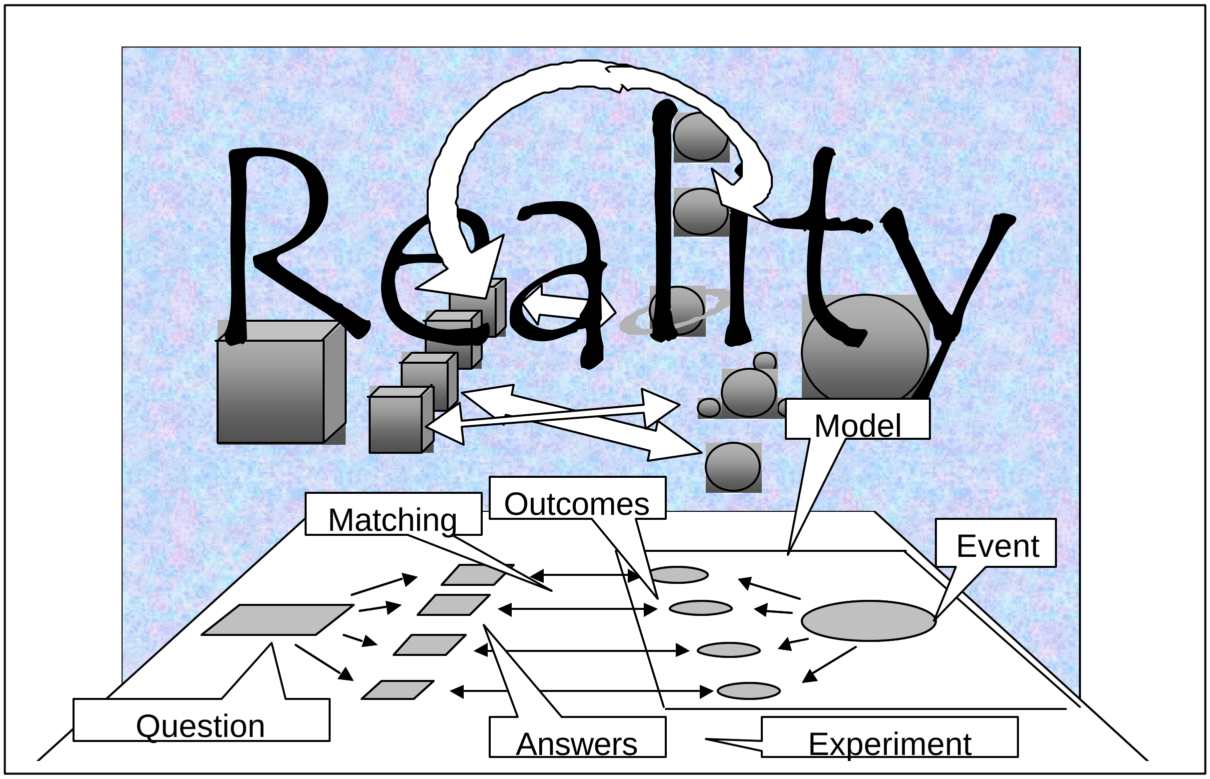
The experimenter experiments upon his or her apparatus, but wonders about reality.

Reality

Any discussion of reality must start with Plato's famous parable of the cave, [18] in which some unfortunate captives have been chained in a cave all their lives, and can see nothing of the outside world (or, indeed, of anything) except shadows cast on the back wall of the cave. These they believe to be real things, because—having seen only shadows all their lives—they do not know any better. Plato held that all people are in a similar condition, seeing and interacting only with images of real things, not the real things themselves. I will leave the philosophy to the philosophers, but borrow the metaphor: our experimental set-ups are mere images of a richer reality, and our experiments' questions are therefore necessarily projections of our real questions onto the world of the experimental set-up. This state of affairs is illustrated by figure 5, in which the experimental schema appearing in several preceding figures is seen as a surface on which the experiment's question, possible answers, activity, possible outcomes, and even the matchings, are the shadows of a far richer and more complexly multifarious reality.

[13] In an engineering test, there may be no such layer because the question being answered does in fact center on the article in question: under what load will it break, or how accurately can it shoot? This narrowness of focus may be considered the defining characateristic of a *test*, as opposed to other experiments.

Figure 5. The model and the experiment in relation to multifarious reality



The experimenter must then ensure that three aspects of correspondence are correct:

- The correspondence of the experiment's event to the events of reality
- The correspondence of the outcomes of the event to answers to the experiment's question, called a "matching" when discussed above, and
- The correspondence of the experiment's question to the genuine question regarding reality.

The correctness of these correspondences is to be judged according to how they compare with the way in which reality's events match up with the answers to the genuine question. That is to say, in terms of figure 4 above, that the arrangement of the two-dimensional figures (including the two-headed "matching" arrows) in the experiment is judged according to how well it corresponds to the arrangement of the three-dimensional figures (again, including the two-headed arrows) in the part of the figure labeled “reality.”

To accomplish this, however, we need an understanding of reality other than that embodied in our experiment, because we propose to check the latter against the former. Such an understanding is called a “theory.”

Theory

The word “theory” has a variety of meanings. It is sometimes used:

- As if synonymous with “hypothesis,” or even “speculation,” as in, “I have a theory.”
- As the antonym of “practice,” as in “That’s all very well in theory, but it would never work in practice.”
- To mean “systematically organized knowledge applicable in a wide variety of circumstances, especially a system of assumptions, accepted principles, and rules of procedure devised to analyze, predict, or otherwise explain…,” [19] as in “music theory,” “game theory,” and “the kinetic theory of heat.”

The role of theory in experimentation is to create an understanding of reality that can be used to create and connect the experiment’s event and its question in a way that reflects the connection of reality’s events to the questions about which the experimenter is really interested.

One important function of theory is simply to enforce consistency. Even a flawed theory can be of significant benefit in this regard, e.g., by fostering a standard terminology and a tendency to stick to a single set of assumptions. Let’s consider the physics examples given in the previous chapter. Each one, in its own way, shows how theory does this.

Theory in physical experiments

Galileo was trying to prove a point about the speeds at which falling bodies fell, but he could not measure these speeds directly and he was hard pressed even to measure short spans of time. But he could tell, with some assurance, whether two things happened at nearly the same time. Therefore he applied what little kinematic theory existed in his day,

$$rate = distance/time,$$

simultaneously releasing balls and allowing them to drop the same distance, and deduced that their (average) rates of fall were equal because the impacts occurred simultaneously.

Absent Newton's theory of gravitation, with its idea that gravitational attraction is proportional to mass, Cavendish would not have known that the objects in his experimental activity had to correspond to reality only in terms of mass (that is, that he could use simple lead balls to answer a question about the Earth and the Moon, with no need to make the balls out of rock, or to paint continents and oceans on one and craters on another). Nor would he have known how to translate the displacements of the balls into a measurement of the gravitational constant or the "weight of the Earth."

Michelson and Morley had a fairly accurate figure for the speed of light and had an idea of the speed of the Earth's supposed passage through the ether. Based on these, they knew that the difference between the two light beams' times in the "race" was going to be very small indeed. To make their set-up capable of measuring such a small difference, they used the existing theory regarding light, according to which two streams of light waves that differed only in being offset from one another would cancel, creating dark bands of "interference fringe." By this effect, the experimenters would be able to measure a difference in speed even if it was so slight as to put one stream of light only a fraction of a wavelength (i.e., less than the thickness of a soap bubble) ahead of the other. [20]

In establishing the existence of the atomic nucleus, based on the deflections of alpha particles passing through gold foil, Rutherford was able to avail himself of two pieces of theory—Coulomb's Law (according to which charges attract or repel as the (signed) product of the charges, and inversely as the square of the distance between them), and Newton's Second Law (force equals the product of mass and acceleration)—and thereby find the equation of motion of a positively charged alpha particle as it passed a deflecting positive charge in an atom of the gold foil. Solving this equation for an observed sizable deflection, he was able to show that the alpha particle, to have experienced such a deflection, must have approached an entire gold atom's worth of positive charge to within 1 percent of the gold atom's radius: this calculation established the existence of a small "nucleus" containing all the positive charge in the atom.

Billy Mitchell's LTA-like experiment, on the other hand, left out too much: a theory of naval combat would have included damage control, or at least the realization that the night-long unattended leaking of water into the ship needed to be recognized as part of the event, rather than simply pretending that the second day's bombing was an immediate continuation of the first day's bombing. Even if there wasn't anything that Billy Mitchell could do to make *Ostfriesland* a moving target or to provide some semblance of damage control, he or others could have recognized, in the interpretation of the outcome, the part that the artificialities played in determining the outcome, and corrected accordingly when drawing conclusions.[14]

Theory in operational experiments

Morse and Kimball address directly the need for theory as part of the basis for experimentation:

[14] Artificialities are discussed at length in *The Practice of Military Experimentation*.

> [An] important requirement is that one should have some general theory of the operation before the experiment is started. This requirement is in common with other scientific experiments; one does not usually blindly measure anything and everything concerned with a test, one usually knows enough about the phenomena to be able to say that such and such variables are the crucial ones, and that effect of others is less important. One should know approximately where the errors are likely to be the largest, and should be able to get the range of the variables over which the greatest number of measurements must be made. It is not necessary that the theory be completely correct, for the theory merely provides a framework for planning the experiments. If the measurements turn out to disagree with the theory, this will be almost as helpful as if they agreed. In fact, an investigation of the disagreement between the measurements and the preliminary theory sometimes provides the most fruitful results of the whole experiment. [21]

Let us consider Admiral Dönitz's experimentation in this light. Admiral Dönitz's theory of U-boat warfare viewed the U-boats of the day not so much as undersea ships as submersible ships. Dönitz felt that Germany's First World War employment of U-boats as individual raiders had failed to take into account not only their near-total the other side would form its ships into convoys.[15] Dönitz also noted an under-used asset of U-boats: their respectable speed when running on the surface, and that surface operation actually represented a countermeasure to the anti-U-boat sonars of the day. Based on this inventory of U-boats' strengths and weaknesses, and anticipation of the adoption of convoy as a countermeasure, he conceived of "wolf pack" tactics: a dozen or more U-boats would form a long line at right angles to the expected track of the convoy, spreading out as

[15] Based on thought-experimentation, Dönitz seems to have appreciated that even without escorts, the convoy is a countermeasure to the individual submarine, because the close grouping of the merchant vessels reduces the size of the region from which any are spotted. Because the submarine is limited in its ability to attack, the convoying side is more than willing to trade a large chance that one vessel will be sighted for a slim chance that many vessels will be sighted all at once. Escorts, if present, provide the added advantage of repulsing the attack of a single submarine, or at the very least preventing a re-attack.

far as possible without creating a gap through which the convoy might pass. When a U-boat saw the convoy, it would send a signal to higher headquarters, which would mastermind the convergence of the U-boats at a point farther along the convoy's route, where they would submerge and lie in wait, and attack the large number of ships with a large number of U-boats. Any escorts would be overwhelmed by the U-boats' tactic of attacking all at once—on the surface, and at night, if possible. [22]

Dönitz had a number of questions about his idea. He later enumerated:

> a. *The exercise of control.* How far is it possible to exercise command over a number of U-boats? Is it possible during the actual attack, or only as far as to ensure co-ordinated action before the attack? What is the ideal balance between the exercise of overall command and giving the U-boat its independence of action? Must command be exercised by a person actually at sea? In a U-boat? Or in a surface vessel? Is it, anyway, possible to exercise command from a U-boat? Can command be exercised wholly or partially from land?
>
> b. *Communications.* How can a U-boat be contacted when it is surfaced, when it is at periscope depth, when it is completely submerged, from another U-boat, from a surface ship and from a land station? ... The whole question of transmitting, receiving, and reporting beacon signals. ...
>
> c. *Tactical.* How should the U-boats, operating together, act? ...[23]

To answer these and other questions, he resorted to experimentation. The essential point to notice is that the experiment began with a *theory* of U-boat warfare (as distinct from any *hypothesis* about it), which led to a set of definite questions, some of which could be answered only by at-sea experimentation.

Sources disagree as to when this experimentation began. Some date it as early as the first part of the 1920s, when Germany had not yet violated the Versailles Treaty ban on submarines. In this view, torpedo boat exercises in tactical development, undertaken in that period, were in fact exercises in submarine tactical development, with the torpedo boats being used as surrogates [24] in what would be an excellent example of *operational reductionism*. It is certainly possible: the above-cited questions all refer to the part of the plan during which the submarines would be on the surface, and Dönitz was in a torpedo-boat flotilla at the time.

In 1935, Dönitz was given command of the Third Reich's first U-boat flotilla, and started work on wolf-pack tactics right away. Whether or not the torpedo-boat evolutions had been intended as U-boat experiments, they were used as a source of insight into future U-boat operations. One of Dönitz's subordinates wrote:

> The end of 1935, then, saw the birth of those wolf-pack tactics which were later to be perfected in so masterly a manner. But between anticipation and perfection there were many stages. For reconnaissance and screening duties we adopted the old torpedo-boat tactics as our godparent....[25]

Later, in 1937, Dönitz began to experiment with actual submarines: a wolf pack of some 20 submarines located and successfully "attacked" a convoy of armed transports sailing from East Prussia to Swinemunde, in the Baltic Sea, with Dönitz exercising command by radio from a surface ship at Kiel. Subsequent experiments in the Baltic and elsewhere were supplemented by real-world experience in the Spanish Civil War. [26]

The Warfighting Laboratory's Hunter Warrior experiment, to take another example, dealt with a proposed style of warfare in which supporting fires (to include CAS) took on predominant importance. Therefore the Hunter Warrior experiment was designed around observation and fires-calling, with little provision for direct-fire small-arms engagements.

The perils of inadequate theory

Milgram's small-world experiment has recently received considerable criticism, most of which can be traced to lack of sufficient theoretical underpinning for the experiment. In fact, Milgram undertook his experiment[16] in a virtual vacuum of theory, much to its detriment. For example, the experiment's set-up makes an implicit assumption that people will route the package via the shortest possible path, but no reason is advanced as to why they would do so. Conversely, most packages never got to the destination person at all, and Milgram dealt with these simply by ignoring them altogether and using only successfully received packages in computing the famous average of six links.

Nor does Milgram appear to have taken into account the results of Solomonoff and Rapoport on the "Connectivity of Random Nets," which would indicate that for any reasonable estimate of people's numbers of friends, six steps is actually *more* than one might expect for a country the size of the United States, and thus indicates the prevalence of insular cliques rather than amazing connectedness.

The lack of applicable theory is as serious a problem for many military experiments as it was for Milgram's. In the military, the widespread derogatory use of the term "theory" in the first two senses of the three given earlier (see the bullets in the beginning of the Theory subsection) has probably not only detracted from its use in the third sense, but perhaps even deterred some people from the activity described therein. With a few exceptions, military theory is woefully underdeveloped by any standard. In fact, much of what passes for "military theory" is platitudinous and without empirical foundation. [27] Absent an applicable, well-developed theory, experimentation all too often turns into aimless departures from normal activity, undertaken with the hope that something interesting will emerge, and/or mere demonstrations of the hardware capabilities of new equipment.

[16] His small world experiment, i.e., the only experiment of his that is discussed in this paper. (This note is necessary because "the Milgram experiment" is a term normally used to mean something else.)

Realism

Suppose that some left-over alchemist or magician had objected to Cavendish's result on the basis that it was *unrealistic*: the cannonballs aren't the color of the Moon, lack craters, etc. Cavendish's construction of his experiment was guided by theory: it was important to get the mass right, but not the color or texture. Cavendish could perform a high-fidelity gravitation experiment with unpainted metal spheres because the theory of gravity said that gravity stemmed from object's mass, not its color, texture, or composition. Absent a detailed theory of what he was measuring, Cavendish would have been obliged to reproduce the color, texture, composition, etc., of the Earth or Moon when making his torsion balance.

There *have* been a few successful theories applicable to warfare. CNA's predecessor organizations, the wartime ASWORG (Anti-Submarine Warfare Operations Research Group) and the postwar OEG (Operations Evaluation Group), developed a mathematical theory of "search and screening" [28] that fit the third definition, and proved quite useful. It, and some other formal theories, will be presented in the next chapter. But for now we will take "theory" to be a somewhat more rough-and-ready body of "systematically organized knowledge" than what would be needed to make a mathematician happy.

Results of Warfighting Lab experiments have often been dismissed on the grounds that the experiment wasn't "real." But not all of the experiment needs to be real, only certain parts, and these parts can be identified through the use of *theory*.

Example: penetration, thrust, and swarm

The first LOE of Urban Warrior was devoted to the exploration of three urban tactics: *Penetration, Thrust,* and *Swarm* (PT&S).

- *Penetration* was usually defined as “a raid without a withdrawal:" the force would move to the objective with rear security, but would not maintain a permanent hold on the ground over which it had passed. Once at the objective, it would accomplish whatever was to be done there (very possibly just occupying a valuable installation such as a water purification plant) and expect to be extracted later by a larger force.

- *Thrust* meant that the force would hold open a narrow corridor all the way from the line of departure to the objective. Some wags observed that the most obvious reason to do a thrust would be to rescue Marines who had done a Penetration.

- *Swarm* featured multiple mobile forces that operated independently and met only on the objective. Swarm was often presented as applying when in a reactive, if not defensive, situation while holding a lot of ground: in a peacekeeping situation, for example, small forces would constantly be on the move in a largely friendly environment, but they could swarm to any trouble spot that might develop.

These tactics were new, and were subjected to multiple parallel seminar-type war games in preparation for LOE 1. The Marine officers in these games expressed skepticism regarding the tactics. To some, Penetration looked like a “How Not To” example from the Advanced Warfighting School, or “Mogadishu II.” Thrust seemed to be nothing new, and in fact looked like exactly the kind of thing that has given city fighting a bad name. The Marines didn’t like Swarm either: it sounded non-proactive, chaotic, and uncontrolled, and when a controlled version was posited, they pointed out that it was identical to what patrols do already.

Some theory would have helped. For example, consider the following:

> You can tell from the resemblance of PT&S to How Not To examples that PT&S are *logically* sound: that’s why people have been tempted to use them. They used to fail not because of any conceptual weakness, but simply because of weapon lethality was insufficient to support them. With today’s weapons (and even more so with tomorrow’s), and in

> the urban environment where everything is more lethal because of the short ranges at which engagements take place, lethality will be so high that PT&S will be the tactics of choice. Notice that on a much vaster scale, the Soviets used Penetration and Thrust, at least, during WW II. Now we can use these tactics on the MEU level because a modern-day or near-future MEU—at urban engagement distances—packs the same lethality as a WW II Soviet corps.

This may have been the theory behind PT&S; however, it was never stated, because the tactics were formulated by a taciturn, intuitive genius, with whom discussion of theory was impossible.

If the above had been acknowledged as the theory behind the proposed tactics, the effect on experiment planning would have been practical and immediate. For example, those planning the adjudication procedures were concerned about the proposed procedures for adjudicating shots with M203 grenade launchers, SMAW rockets, and SAWs.[17] If the theory underlying these tactics had said they were supposed to work because Marines could suppress adjacent buildings with M203s and SAWs, then adjudication of shots with these weapons would have been seen as important in assessing the new tactics and effort would be applied to improving the adjudication system. If, on the other hand, Penetration, Thrust, and Swarm had been held to be promising for other reasons, e.g., that obscurants work especially well in an urban environment, then the designers of the experiment would have known they could tolerate a weak M203 adjudication procedure but would need to have sufficient HC smoke grenades on hand.

The perils of serendipity

One of the many problems of relying on serendipity as a strategy for military experimentation is that one cannot design the experiment so that the artificialities do not threaten the goal if one does not know what the goal is.

[17] Adjudication is defined and discussed in *The Practice of Military Experimentation*.

After the experiment has taken place and the alleged serendipitous discovery identified, the analyst may have difficulty knowing whether it is a true finding, or spurious result stemming from some *artificiality* of the experiment. In one of MCWL's Urban Warrior experiments, for example, it was observed that "helicopter-mounted Hellfires and 20mm guns proved remarkably effective." This seemed to be a serendipitous—and important—finding of the experiment until it was realized that there was no provision for scattering the fire of helicopter-mounted Hellfires and 20mm guns. Consequently, these were always adjudicated as hits, and their seeming effectiveness was entirely spurious.

However, truly serendipitous results *can* occur, and they can be highly valuable. Very possibly, the most important result of MCWL's Urban Warrior series of experiments was the serendipitous finding of systematic flaws (later rectified by Project Metropolis) in the Marines' training for urban combat. When an experiment seems to have produced a serendipitous finding, the first order of business must be to establish the validity of the finding via an experiment expressly designed to do so. This second experiment serves to rule out the possibility that the serendipitous finding is an illusion.

Garbage in, garbage out

A computer-age aphorism holds, "Garbage in, garbage out;" this concept even has an acronym, GIGO. [29] A related aphorism holds that if one adds a tablespoon of fine wine to a barrel of garbage, the barrel still contains garbage, but if one adds a tablespoon of garbage to a barrel of fine wine, a conversion does occur, and the contents become a barrel of garbage.

One approach to military modeling—including not only computer models and simulations, but all the "models" used in military experiments, including exercise-like live action—might be termed TITO: "truth in, truth out." The idea is that if all of the pieces and parts are completely realistic, then the whole will be completely realistic, and its results can be trusted. One problem with this approach is that nothing can ever be perfect, and—in line with the aphorisms—any amount of garbage in the input turns the entire input, and thus the output, into garbage. Another problem is that even if all the inputs are true, they may not, and in fact cannot, be the whole truth. This is the Achilles Heel of physics-up combat models such as Janus; they may not, and in fact cannot, include everything. Even the National Training Center at Fort Irwin, the ultimate in physics-up models in that it uses real people and some real equipment, suffers from some of the same problems as entity-level physics-up models, because even it does not start with the whole physical truth.

So, what is needed is some form of GITO—Garbage In, Truth Out. Strange as it may seem, this is actually possible.

Truth

Some have gone so far as to argue that a model is worthwhile *only* if it can perform GITO. After all, if you can't get out anything other than what you put in, what's the point?

Recognizing that even true assumptions can lead a model to false outputs, the economist Milton Friedman seeks the validation of models in their predictive power, not in the truth of their assumptions. He points out that a full set of true assumptions is not even a necessary condition for a valid model, in that false assumptions can still lead to true predictions. (In particular, Friedman is interested in assumptions about human behavior; a true set of predicates for human behavior would be impossibly large and complex, so Friedman seeks a set which, though necessarily false, will give correct answers to his questions. Paul Samuelson called this the "F-twist.") In our terms, he actually rejects TITO!

Friedman, in what Samuelson would later term the "extreme version of the F-twist," states that models whose true predictions come from false assumptions have special merit. Neither Samuelson nor Blaug (from whose book this account is taken [30]) can see any compelling reason to espouse the extreme version of the F-twist. But surely there are practical reasons as well as philosophical ones for preferring a model that can make a silk purse from a sow's ear to one that requires a silk purse as an input.

The resolution of this paradox is that although the model may be (and in fact almost certainly is) “a wrong model” in the sense that it is wrong about some particulars, it can nonetheless be “the right model” in that it answers the questions under study correctly.

The process of modeling has been compared to the art of cartooning, on the basis that the skill lies in knowing what to leave out. A more radical simile would be impressionist painting: the whole picture evokes the whole subject, even though—when viewed from close up—no part of the picture resembles any part of the subject.

To return to Plato and the cave, the model and the real world are both shadows of the same ideal. The respects in which the model is “wrong” are respects in which it particularizes the ideal differently from the way in which the real world does, but the correct use of the model is not adversely affected by these.

Consider, for example, an architect’s model of a complicated mansion. It differs from the actual mansion in numerous respects—it is the wrong size, it is made from different materials, etc., etc., and in fact its points of resemblance to the real thing are few, *but* it suffices for the intended purpose: it can be used to answer such a question as, “Can one see the back door of the garage from the solarium?”—before either the garage or the solarium has been built.

To turn to a military example, consider the military officer who disdains computer models as inaccurate and says that his wisdom is based on his study of history. To be sure, the tanks of El Alamein differ from today’s tanks to an even greater degree than do the tanks in the disdained computer models, and the elephants of Zama are more different still. But the student of *broad general truths* may find that the tanks of the 1940s and the elephants of the BC era are shadows of some great truth of which today’s tanks are also a shadow, and he may thus benefit from his reading, even though it fails to meet the standard by which he rejected the computer models.[18]

In some cases, the paradox doesn't really even exist:

- Sometimes the purpose of the experiment is to demonstrate an "existence theorem," i.e., a statement that something could possibly exist.
- Sometimes the fact that the experiment may have given the wrong answer is not important, because the goal was to find questions, not answers.

It is tempting to define a "model" as "a producer of truths by a process other than logical deduction or induction."

Recent work in "complexity" has addressed a phenomenon called "emergence," in computer programs that deal with the interaction of agents. [31] At the group level, one can observe properties whose rootedness in individuals' traits is perceived, if at all, only with difficulty and hindsight. For example, a few simple rules can make artificial agents *flock*, in two or three dimensions, very much as do birds—but without any overall control and with each "bird" executing only local rules that govern only its behavior. Some of the complaints about combat models, particularly about the ways in which they treat human behavior, relate to emergent properties. Cohesion and suppression, for example, are often cited as poorly treated in combat models. They are properties of groups, not of individuals, yet they are clearly emergent properties because the groups consist only of individuals. Similarly, routs, breakthroughs, last stands, and the like are rarely predicted by traditional models, but they happen routinely in agent-based ones. These are emergent behaviors. A modeling effort that succeeded in reproducing and predicting emergent properties and behaviors would be a clear success, because something would have come out that was not put in. It would also qualify as GITO, showing a clear profit in the sense that what-

[18] The modern reader may be helped by thinking of the Platonic "shadows" as particulars, and the "ideal" as an abstraction of them: philosophers say that Plato didn't think of it that way, but we said above that philosophy would be left to the philosophers.

ever its assumptions about people, they would be far simpler than the truth.

In physical science, a *theory* that was known to be based on false assumptions would be discounted immediately, or at least restricted to regimes in which its assumptions were sufficiently close to being true for practical purposes. (For example, classical physics is retained for many purposes.) However, physical science offers many *models* as explanations, heuristics, or aids to calculation, without pretending that they are true per se. Examples include the nuclear cross-sections and the wave and particle interpretations of light, as well as Faraday's "lines of force" and the whimsical terms used in quantum mechanics, in which not only English words (e.g., "spin," "color") are appropriated, but also the structure of English, in which, for example, "up" is not a color.

To be sure, physical scientists use the term "model" far less than do their self-appointed sideline commentators. But when a scientist does refer, explicitly or otherwise, to a "model," she means something that is wrong, but in a useful way.

Some theories with military applicability

There is nothing so practical as a good theory.

—Kurt Lewin

One purpose of the preceding chapter having been to persuade the reader of the importance of theories, the present chapter will acquaint the reader with some theories (i.e., bodies of "systematically organized knowledge ... devised to analyze, predict, or otherwise explain," not hypotheses or speculations) that can be of use in considering military matters. Most of these have literatures of their own, to which references are given.

I believe all of these theories to be of use, but to have limited scope; the temptation to believe the over-generalizers must be avoided. On the other hand, the whole can be greater than the sum of the parts: the analyst who understands all of these theories will possess a set of structures that will enable him or her to conceive of useful abstractions in new situations. Part of the appeal of these theories lies their *portability*: as we will see in some of the examples below, the same theory can, be applied in a wide variety of different—or even disparate—situations.

Obviously some basic physical theories, such as the theory of gravitation, have military applicability because they apply to nearly all areas of endeavor. Such theories will not be addressed here.

In 1935, G.F. Gause, an early evolutionary biologist, remarked, "Apparently every serious thought on the process of competition obliges one to consider it as a whole, and this leads inevitably to mathematics." [32] The same could be said of almost any type of process, if considered as a whole: an important point held in common by all of the theories discussed in this section is that they are inescapably mathematical, and in particular probabilistic. This mathematical treatment, in turn, will be seen to lead to the formulation of *Measures of Effectiveness*, which are highly desired in experimentation.

Probability theory, and statistics

Probability theory is the basis of the more particularized theories to be discussed in the succeeding sections, but it also can be applied in raw form to military experiments. It is too large and well known a topic to be the subject of the kind of précis that will be presented regarding the theories treated in the succeeding sections, and in any case the reader is assumed to have been exposed to it in school.[19]

As an example of the application of probability theory in raw form, let us analyze MCWL's data on which parts of Marines' bodies received hits in the Project Metropolis series of experiments in urban warfare.

An early MCWL-sponsored study voiced a hypothesis regarding wounds in urban combat: "Wounds of head, neck, and chest caused by small arms will increase." [33] It went on, however, to admit, "there appears to be no single or even collected work that supports this assertion, though it seems likely that this would be the case."

Later MCWL field experiments with urban fighting used a surrogate for small arms fire: Simunitions®. These are 9mm paint rounds, fired by a reduced charge from specially adapted M-16s,[20] and the paint allows subsequent observation of where, on the participants, the rounds hit.

[19] The person interested in the art of military experimentation (or, indeed, experimentation of any kind) would do well to read the recent *Probability Theory: The Logic of Science*, by E.T. Jaynes. Though the book is new, most of the content has withstood the test of time, since it consists of articles written and published over a period of decades. Jaynes provides an exceptionally clear account of the conflict between the Bayesians and their opponents, which will therefore not be discussed here.

[20] Simunition® is a registered trademark of SNC Technologies Incorporated. A special-purpose barrel and upper receiver adapt the standard-issue M-16 to fire these; for safety's sake, this adapter is unable to fire standard non-paint 9mm ammunition.

Table 1 shows data on hit locations recorded during MCWL's Project Metropolis series of experiments, compared with single-shot rifle and machinegun wounds in the U.S. Army's Bougainville campaign during the Second World War. [34]

Table 1. MCWL urban data compared to Army jungle data

	Urban	Jungle	Totals
Head	62	171	233
Torso	154	138	292
Arms	101	120	221
Legs	74	112	186
Total	391	541	932

If there were no real difference between the Project Metropolis experience and that of Bougainville, one would expect the data to be close to that of the "null hypothesis" shown in table 2 .

Table 2. Urban/jungle null hypothesis

	Urban	Jungle	Total
Head	98	135	233
Torso	123	169	292
Arms	93	128	221
Legs	78	108	186
Total	391	541	932

The null hypothesis takes as given the totals of the rows and columns, i.e., it takes as given the overall numbers of wounds sustained in the MCWL experiments and at Bougainville, and the overall total hits in the head, torso, arms and legs. But under the null hypothesis, there is no real difference between the two cases, so one would expect that the particular entries would be proportional to their row and column subtotals. For example the number of head wounds sustained at MCWL would be expected to be

$$233 \text{ x } 391 \ / \ 932 = 98.$$

The null hypothesis is, at any rate, different from the observed data, so the question then becomes, Is it "close," and how ought we to measure "closeness"? A reasonable answer to the second part of the question is that we should ask, "Under the null hypothesis, how likely is it that we would observe what we observed?" We can then apply the chi-squared test[21] to answer this question, and the test says that if the urban and jungle wound-site frequencies were in fact drawn from the same distribution, the observed data would be highly improbable—they are, in this sense, very far from the null hypothesis. Inspecting the data table and comparing it to the null hypothesis table, we can see that while the hits on the limbs are about what might be expected, the head and torso hits are quite different: the urban head hits are much fewer (36 fewer) than expected, while the urban torso hits are much more (31 more) numerous than would be expected. This finding is at odds with the hypothesis, which maintained that the proportions of head, neck, and chest wounds would increase in urban warfare.

Search theory

The quantitative theory of search was developed by early operations researchers in the United States and the United Kingdom during the Second World War. The standard work is Koopman's *Search and Screening*, but Morse and Kimball's *Methods of Operations Research* and Sternhell and Thorndike's *Antisubmarine Warfare in World War II* might be better places to start. Several other books and a large number of articles address one aspect or another of the topic. Washburn's *Search and Detection* may be of particular interest to the military experimenter: Washburn performed a number of experiments to compare mathematical results to the performance of human searchers (and evaders).

The essence of search theory is really just the proposition that search is a process that can be addressed in mathematical terms. The objects of the search are seen as points in some appropriate space such as a two-dimensional sea surface, a three-dimensional volume of air or water, or a one-dimensional interval of time.

[21] The details of the test are standard textbook material, and so will not be recapitulated here, but it is worth commenting that the black-box nature of the chi-squared tests, and of statistical tests in general, derives from the fact that they are approximations (albeit often very close ones) of probability calculations, with the meaning-obscuring recourse to tables having been necessitated by the absence of today's quick and easy computing.

Searchers detect these targets according to a probability function that usually includes among its arguments the distance of the searcher's closest approach to the target, but may well also include other variables expressing characteristics of the target or the environment.

The existence of closed-form solutions to many of the important equations of search theory has led to a quite a number of results regarding how to construct searches that are "optimal" in one respect or another, and such results predominate the current literature. These quantify and solve tradeoffs of the kind faced by the individual who wanted to look for his keys where he thought he might have dropped them, but also saw an advantage in looking nearby, where they were less likely to be but the light was better.

Treatments of search theory occasionally make reference to *detection theory* (addressed below); Koopman and Washburn do so, for example. Likewise, links have been forged between search theory and *information theory* (also addressed below).

Search theory was applied with great success in the Allies' effort to counteract German U-boats in the Second World War, as recounted in the above-cited books as well as in the book by Waddington, who says that perhaps the most important effect of the use of search theory was to highlight the importance of the concept of "target density," i.e., the number of targets divided by the area (or other appropriate measure) of the region they occupied.

The concept of target density led to an interesting definition: the *operational search rate* is an MOE obtained by dividing the number of targets sighted by the target density and by the time spent in searching. In a typical Second World War submarine-hunting case, for example, 5,000 hours of aerial search in a region infested with one submarine per 100,000 sq n.mi. might result in ten sightings. Thus, the operational search rate would be

$$\frac{10}{\left[{}^{1}\!/_{100,000}\right]\cdot 5,000} = 200 \text{ sq n.mi./hour.}$$

This results-based estimate of the search rate was typically (albeit not always) less than the rate that would result from a calculation based on the airplane's speed and the distance that observers could see to either side. As such, it was instructive because it reflected the effects of the observers' fatigue, the camouflage (or submergence) of the submarines, and so on. Thus it could be, and was, used to predict the number of future sightings. [35]

Such successful MOEs are hard to come by, especially in a real-world setting: the wartime operations researchers were uncommonly fortunate to know the density of targets by a means (signals intelligence) other than the aerial search effort whose effectiveness they were trying to measure.

Today's emphasis on sensors in warfare, e.g., those carried by Unmanned Aerial Vehicles (UAVs), ensures that search theory will be as important in the future as it has been for the last several decades of military analysis. The operational search rate could be determined in military experiments (in which the target density would be known), and measurements of the operational search rate would provide a valuable counterpoint to the search rate claims made by proponents, which are invariably the result of multiplying the UAV's speed by the width of its field of view. In one MCWL experiment, the operational search rate for a particular UAV was found to be *zero*, no targets having been noticed by the operators. In the real world, one would not know whether targets had been present, but the experiment was instrumented and analysts played back the UAV's flight and saw that it came into the proximity of targets. They then examined the taped output of the UAV and verified that in fact the targets had been in the field of view—but so fleetingly and indistinctly, and after such a long period, that the real-time user did not notice them.

The idea that a target could be present and yet not register, leads into the next theory to be discussed, detection theory.

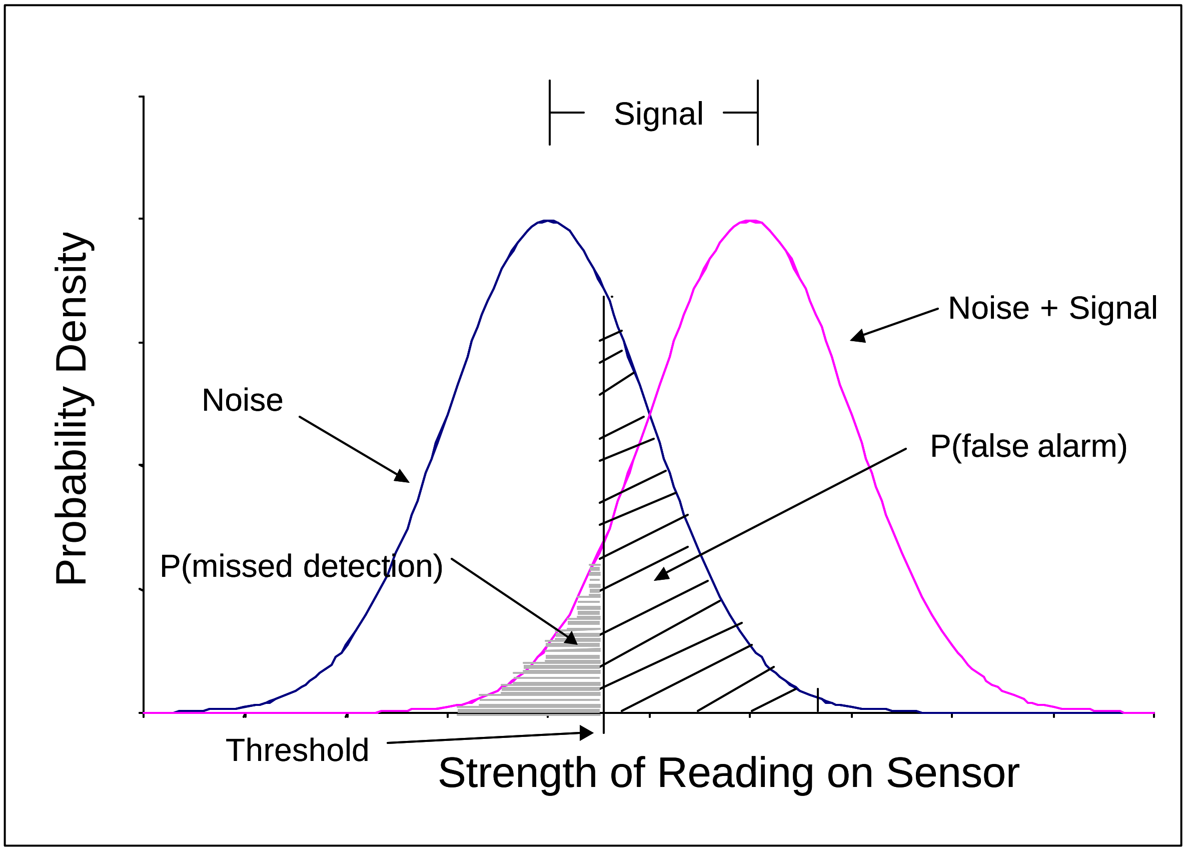
Detection theory

Detection theory may be considered to be a highly specialized branch of statistics that addresses the question of confirming the presence or absence of a *signal* that appears amid a level of *noise* that is high enough to make the problem non-trivial. As the terms suggest, the signal might be an acoustic signal, or it might be nearly any other form of signal—electro-magnetic (i.e., light, or radio), seismic, chemical, radiological, etc. Selin's *Detection Theory* provides a useful introduction, and treatments also appear in the 1980 edition of Koopman's *Search and Screening*, and in Washburn's *Search and Detection.*

A principal result of detection theory is the derivation of the *receiver operating characteristic curve*, as follows. The variation in the background noise level can be expressed as a probability density function, showing the probability that the noise is at any given level. In each individual "look," the receiver hears this noise, and the signal if there is one, and the operator faces the problem of setting a threshold above which a detection will be reported. The trouble is that there is no level that is not sometime attained by pure noise, absent the signal, but, conversely, sometimes the noise is much lower and an overly high threshold will result in non-detection even when the signal is present.

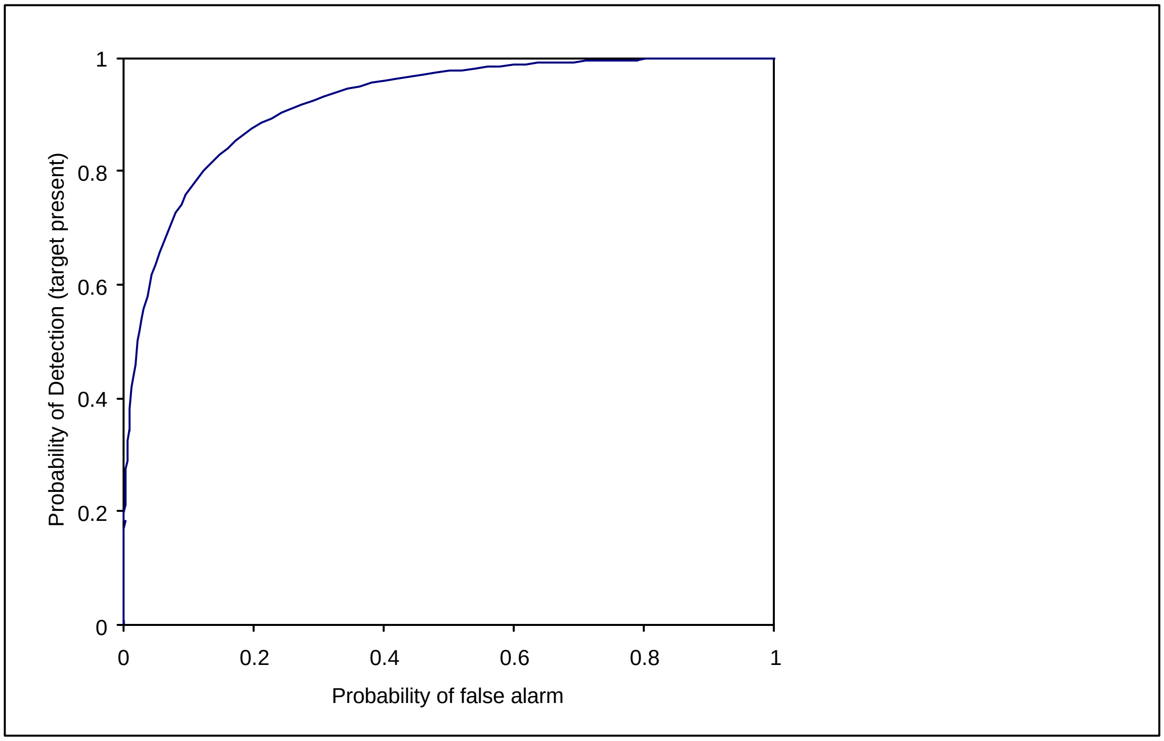
This situation is depicted in figure 6, which shows the distribution of pure noise to the left, and, on the right, the shifted distribution that results when the signal is added to this noise. At most or all threshold settings, there is, as shown, some probability of missing the signal when it is present, and some probability of reporting a false alarm when no signal is present.

Figure 6. The basic idea of detection theory



The result is a tradeoff, usually expressed dismally as the tradeoff between two bad things: the probability of missing the signal when it is present, and the probability of having a false alarm, i.e., of reporting that the signal is present when it is not. The inescapable existence of this tradeoff and the operational importance of false alarms—whose occurrence and harm are neglected or ignored during the development stage of many systems—are key points in military experimentation, if only because an assessment of a new system's propensity for false alarms can be had only through a realistic operational experiment. This tradeoff can be plotted by varying the threshold parametrically and creating a plot of detection probability versus false alarm probability, as shown in figure 7. This is the Receiver Operating Characteristic curve, or ROC curve. Figure 7 shows the ROC curve for a signal that is twice the standard deviation of the noise: by varying the detection threshold, the operator can operate the receiver at any point on the curve. For example, a 98 percent chance of detecting the target (if one is present) can be had—at the cost of having a 60 percent chance, on any given reading of the sensor, of registering a false alarm if no target is present. In most situations, the false alarm probability must be held to a low level, e.g., 1 percent or even much less. But doing so with this receiver would give only a 4 percent chance of detecting a target if it is present.

Figure 7. Receiver operating characteristic (ROC) curve



Of course, the situation shown here and in the previous figure is, for the sake of illustration, worse than anybody would ever want to have. The trouble is that the signal is so small in comparison to the variation in the background noise. False alarms and misses can be traded, but the only way to lessen both at once is to move to a situation in which the curves in the first figure have less overlap, i.e., the ratio of signal to standard deviation is increased. This quantity or, (for reasons having to do with the definition of electrical power) its square, the ratio of the squared signal to the variance in the noise, is therefore a fundamental MOE for a receiver.[22] This is the famous "signal-to-noise" ratio, an MOE of such importance that it has passed into the language as a metaphor.

Increasing it would push the curve in the second figure farther into the upper-left corner, reducing the false alarm probability associated with a given probability of detecting the target if it is present.

Often, those who propose or design sensor systems fail to take into account the trade-off between misses and false alarms. (The author was once asked, in incredulous tones, "You mean the *same sensor* can fail in two different ways?")

In many settings, "looks" occur nearly continuously and targets are not normally present, so the single-"look" false alarm probability must be kept very low so as to have an acceptably low false alarm *rate*. We will see an example of this in a later chapter.

Detection theory usefully shapes one's thinking by pointing out that the false alarm rate is just that—a rate—and that false alarms therefore occur independently of detection opportunities. This realization, moreover, resolves the seeming paradox that arises if one thinks of the quality of the sensor in terms of the ratio of false alarms to true detections: operators, and the recipients of their warnings, are often surprised to find that the sensor works less well—in the sense that a seeming detection is more likely to be a false alarm—when targets are few.

The "detection opportunity," readily enough defined, is notoriously difficult to identify in practice when working with operating forces in the real world. Military experimentation offers a way to assess the performance of sensors in a realistic environment, yet one in which the detection opportunities are in fact known. Detection theory would come to the forefront of any military experiment that dealt with sensors. It also appears explicitly in the development of some of the other theories addressed in this section.

[22] Electrical engineers sometimes use the term "figure of merit" to mean "measure of effectiveness."

Information theory

Originally formulated by Claude Shannon in the 1940s, the mathematical theory of information addressed certain problems that had bothered communications engineers for some time, such as how to calculate the information-transmission capacity of a given telegraph or telephone line, and how to predict the degree to which that capacity would be degraded by extraneous "noise" present in the line, e.g., that caused by lightning. Important parts of Shannon's trail had been blazed by the physicist Leo Szilard and the communications engineers Harry Nyquist and R.V.L. Hartley. [36]

For whatever reason, perhaps the innate beauty of the topic, the light it sheds on parts of our everyday experience such as the use of language, or both, there exist a number of brilliant explications of information theory, aimed at the beginner but not stinting on mathematical content. These include John R. Pierce's *An Introduction to Information Theory: Symbols, Signals and Noise* (originally entitled *Symbols, Signals and Noise: An Introduction to Information Theory*), Gordon Raïsbeck's *Information Theory: An Introduction for Scientists and Engineers*, and *The Mathematical Theory of Communication*, in which Claude Shannon's original paper and a slightly later paper by Warren Weaver are reproduced. E.T. Jayne devotes a chapter of his *Probability Theory: The Language of Science* to information theory. The mathematically stout of heart should seek out Satosi Watanabe's masterwork, *Knowing and Guessing: A Formal and Quantitative Study*.

In connection with the present purpose, Pierce's book is notable for devoting its first chapter to a discussion of "the world and theories," as a prelude to the book's presentation of information theory.

The utility of information theory for the military analyst is that it forges the link between the supply side, (channel capacity) and the demand side (messages or sensor output). The familiar concepts of optimal encoding, compression, and error correction all come from information theory.

The principal contribution of information theory is to define a *metric* of information, the "bit." Raïsbeck's presentation, and others, start with some simple desiderata, and end up with an entropy-like definition of information, due to Shannon. This, in turn, leads to theorems regarding what is and is not possible to do within a given channel capacity, and pointers to optimal encoding of such messages as English text or digitized pictures, in which not only do all characters not appear with the same frequency, but also the stream of characters contains substantial autocorrelation.

In its mathematical treatment of signal processing, information theory verges near detection theory in explaining why detectors are better able to detect some signals than others, and in explaining and quantifying the trade-off between the probability of detection and the false alarm rate.

Near the beginning of most older presentations of information theory is a short digression on mathematical definitions and their bearing on the real-world referents denoted by the same terms, [37] occasioned by the upcoming definition and use of the word "information" in a narrow technical sense. The authors clearly seek to squelch any objection to the seemingly reductive quantitative treatment of what the reader might consider an ineffable marvel, information. Today, the personal computer industry has made many people comfortable with measuring quantities of information in "bytes," each corresponding to a letter, number, or other typographic symbol and each composed of eight "bits," the irreducible 0-or-1 binary digits sometimes (plausibly, but inaccurately) termed "the smallest possible amounts of information." Because of this familiarity, however, an opposite warning may be in order: the bits, bytes, and megabytes in which we and our household appliances traffic nowadays represent a greater step in reductionism than is made by the theory of information. The mathematical definition of “information” contains more of the ordinary-language meaning of the term than is implicit in our present-day use (which more closely corresponds to a definition and quantification of *data*), and perhaps recaptures some of the marvel as well.

A basic theorem holds that a signal of duration T and bandwidth B can be completely characterized by $2TB$ samples, and that—conversely—$2TB$ samples completely specify the signal insofar as it is contained in the given bandwidth. This fact leads to the idea that the information-carrying capacity of the channel is proportional to the bandwidth, an idea that is now so ingrained that “bandwidth” and “information-carrying capacity” are used interchangeably, and established bandwidth issued as an MOE.

Queueing theory

Queueing theory was originated in the early 20th century by Erlang, in his study of the management of telephone systems. Today the standard work on the subject is Thomas L. Saaty's, *Elements of Queueing Theory, With Applications*, but many others, and a large literature of articles on particular aspects, exist.

The situation addressed by queueing theory is familiar from visits to the bank: customers arrive at random intervals, and receive service that takes randomly distributed amounts of time. Sometimes customers have to wait before they can begin service, because somebody else is being served. If, over the long run, customers arrive at a rate faster than the average rate of service, disaster of course ensues. But queueing theory explains a phenomenon that is observable in everyday life: even if customers arrive at an average rate that is less than the average service rate, finite queues can form.

For example, if customers arrive at an average rate of λ persons per minute and are served at an average rate of μ persons per minute when any persons are present to be served, and if arrival and service are both Poisson processes with $\lambda < \mu$, then the expected value for the number of people in line will be

$$L_q = (\lambda/\mu)^2/(1-\lambda/\mu),$$

and the average waiting time in line will be L_q/λ.

Queueing theory treats mathematically the extent of these queues and how they form in under various conditions—e.g., the presence of multiple servers, and the existence of customers who leave the queue if they have been in it too long or who will not join it if it contains too many people already.

Actual queues, formed by people or things awaiting service, can be of great importance to the military insofar as they arise in the study of logistics. A variety of MOEs suggest themselves: the length of time spent in the queue; the length of time spent in the system as a whole (i.e., in the queue, and subsequently while being served), the probability of, upon arriving, not having to wait at all; and so on. Most of these can be expressed in equations in the case of Poisson arrivals and service, and of course all can be computed numerically from any given distributions of arrivals and service times.

Some CNAC experiments have addressed queueing in emergency medical clinics.[23] The clinics consist of multiple queueing situations: patients arrive, enter a queue, complete part of the process, and enter another queue. Some parts of the process consist of diagnosing the patient and deciding which of two or more alternate queues he or she should next enter. In one clinic, the design reflected a belief that if the service rate equaled the arrival rate, all would be well. We knew from the above equation that (accepting for the moment the assumption of Poisson rates), if the arrival rate was even as high as 95 percent of the service rate, space would be needed to accommodate an average of 18 people in line, and often more. If the arrival rate were 99 percent of the service rate (which, in the event, was closer to the truth), a line of almost 100 people would form, and consume an enormous amount of clinic space. We also knew from queueing theory that the *shapes* of the distributions of arrival and service times—not just their mean—would make a difference in how the lengths of the queue fluctuated, and therefore in the maximum length that a given queue might be expected to attain. This knowledge impelled us to record the length of each service time, rather than just keeping track of how long the clinic was in operation and the total number of people served. In fact, we kept track of each patient's entry into each queue, inception of service, and completion of service. With these data and the conception of queueing theory, we were able to reconstruct where each person was at any given time, and then to write a short computer model that replicated the observed behavior of the clinic and could be used to explore excursions.

The idea of queueing, though, has also been used advantageously as an approach to considering such situations as air defense in which the "customers" are oncoming bombers and "service" consists of shooting them down. Ralph Klingbeil and Keith Sullivan present *A Proposed Framework for Network-Centric Maritime Warfare Analysis,* [38] in which warfare processes as disparate as anti-submarine warfare,

[23] The setting was that of preparedness for domestic bioterrorism, a near-military topic at least.

(air) strike warfare, maritime interdiction operations, and others are cast and analyzed in terms of queueing theory.

Queueing theory's slight generalization covering "birth-and-death" or other "renewal" processes has also been used extensively to investigate military matters. Sakitt, for example, used it in his study of hypothetical Cold War submarine warfare under the arctic ice. [39]

Game theory

Of the theories discussed here, game theory is doubtless the one with which the most readers will be familiar. A popular introduction is available in J.D. Williams's charming *The Compleat Strategyst*, and more detailed treatments are presented by Anatol Rapoport in his two books, *Two-Person Game Theory* and *N-Person Game Theory*. Perhaps the most useful treatment from the military standpoint is that by Melvin Dresher, in his *The Mathematics of Games of Strategy*; for tactical problems, Isaacs' *Differential Games* remains important, despite its age and the huge gap between its results and the needs of the real world. The field's seminal work, von Neumann and Morgenstern's *The Theory of Games and Economic Behavior* is still of value, and a wide (and, in many instances, mathematically deep) literature continues to grow. Recent developments have addressed cooperative games (such as might involve multiple parties on the same side of a military conflict), and Axelrod's *The Evolution of Cooperation* provides a fascinating and accessible introduction to the intriguing concept of the Nash Equilibrium:[24] The work of Thomas Schelling (e.g., *Arms and Influence*), which was of enormous influence in shaping nuclear strategy, stands apart because so much of it is devoted to games in which the players' attention is focussed on the worst possible outcome, not the best or the most likely one.

Except in a few particular situations, game theory has proven problematic in military settings because military decision-makers do not have the precisely quantified "payoffs" that the theory presumes exist.

[24] John Nash is the main character in the recent movie *A Beautiful Mind*, and in Silvia Nasar's preceding book of the same name.

For the military experimenter, game theory is more useful for its concepts than for any particular result, much less any MOEs. The concepts include the perception that a wide variety of situations are in fact “games,” not in the sense that they are frivolous behaviors, but in the sense that they are competitions conducted according to well-defined (albeit usually not agreed-upon or contrived) rules, with the participants being “players” in the sense that they conceive of strategies and use these to try to outwit their opponents.

Military experiments in command and control can in some cases benefit from being thought of as games, and some—e.g., the “Scud Hunt” game of Peter Perla et al.—are explicitly cast as games. Likewise, search-and-evasion situations can be considered as games: Washburn’s experiments in this regard have already been mentioned, and Isaacs, following game theory’s tradition of fanciful paradigms, casts such a game in terms of “The Princess and the Monster.”

“Decision theory,” an alternative to traditional statistics, treats decisions as games against nature. [40]

Lanchester’s attrition theory

Frederick William Lanchester advanced the proposition that opposing military forces, e.g., collections of armed men or machines, can be seen as eroding one another, each shrinking at a rate proportional to the other’s size. [41] If we let *A* and *B* be the sizes of the two forces and *a* and *b* be constants representing their quality or efficiency, then Lanchester’s proposition may be written mathematically as a pair of coupled differential equations:

$$A'(\mathrm{t}) = -b \cdot B(t)$$
$$B'(\mathrm{t}) = -a \cdot A(t),$$

in which *A(t)* and *B(t)* are the numerical strengths (i.e., the number of units—e.g., men, airplanes, or tanks) on each side, *A’(t)* and *B’(t)* are the respective rates of change of these quantities, and *a* and *b* represent the rates at which one unit (on side A or B, respectively) can kill or otherwise neutralize units on the other side.

Lanchester's 1916 publication of this idea is now known to have been anticipated by others, but Lanchester's name is the one that stuck. Lanchester did not publish the solution of this system of equations, but he did make a key observation regarding it: that the quantity

$$a \cdot (A(t))^2 - b \cdot (B(t))^2$$

is a *constant*, determined by the initial conditions: the ongoing decrease in A(t) is accompanied by a decrease in B(t) that keeps the value of the entire expression constant

Lanchester's interesting presentation includes several intriguing applications of this idea. One is that it leads to a statement regarding the relationship between quality and quantity: two forces are equally matched if

$$a \cdot (A(0))^2 = b \cdot (B(0))^2.$$

Another is that if one is in possession of the beginning and ending numbers of combatants on the two sides, one can use the fact that

$$a \cdot (A(0))^2 - b \cdot (B(0))^2 = a \cdot (A(final))^2 - b \cdot (B(final))^2$$

to find the ratio a/b, a (relative) MOE for the quality of the forces on the two sides, apart from their quantity:

$$\frac{a}{b} = \frac{B(0)^2 - B(final)^2}{A(0)^2 - A(final)^2}.$$

Lanchester's original book *Aircraft in Warfare* is hard to find, but most of the relevant section is contained in the extract provided by James Newman in volume IV of *The World of Mathematics*. A short treatment is given by Körner in *The Pleasures of Counting*; a rigorous (even more so than the original) treatment is given by Epstein in *Non-Linear Dynamics, Mathematical Biology, and Social Science*, whose later section on stability can be applied to the Lanchester equations.

There is a lot to be said for thinking of military forces as engaged in mutual erosion, and Lanchester's original idea has been the object of continued use and study since it was re-discovered by the WW II-era operations researchers. Numerous attempts have been made to validate it against data from historical battles, and greatly elaborated versions of Lanchester's equations now form the basis of many a computer simulation.

The present author has numerous reservations regarding the theory, e.g., that the line of reasoning presented above ignores the fact that the system of equations is unstable; that in modern situations the profusion of weapons systems prevents any single number from representing the numerical strength of each side, and prevents any single constant from characterizing the effectiveness of one side against the other, since targeting doctrine must be taken into account (which has as of yet proved impossible) and that through maneuver each side tries to bring the entirety of its strength against a portion of the enemy's.

Several analysts have had the idea of applying the Lanchester equations to data from historical battles, to see if they "fit." Multiple difficulties arise:

- Getting good data. It is hard to get good data, *comparably collected* (i.e., with the same counting rules for who is a combatant at the beginning or end of the battle) for both sides of a historical battle.
- Inhomogeneous forces. The Lanchester approach assumes that although the two sides may have troops of different qualities, each individual side is composed of substantially identical combatants.
- Inhomogeneous fighting: The Lanchester approach assumes that the fighting is "collective," in the sense that each combatant can always find an enemy combatant at whom to shoot.

These difficulties are difficult to surmount, though some successes have been claimed. [42]

In the Marine Corps Warfighting Laboratory's urban combat work, excellent data were available for the two sides, and in most battles each side consisted almost entirely of riflemen. The fighting was arguably not entirely homogeneous, but just for the sake of example, let us explore the use of Lanchester theory in attempting to solve a problem that arose in MCWL work. *If* there is any land-battle circumstance to which Lanchester theory can be applied, MCWL's urban combats are it.

For some time, MCWL's experimentation focused on urban warfare. Initial experiments at Camp Lejeune indicated, inter alia, that the Marines did badly in urban fighting and that part of the problem was their training. Later on, a set of improved urban tactics, and a training package to inculcate them, was developed and a new set of experiments was done at Victorville, in the housing area of the former George Air Force Base, now the Southern California Logistics Airport. The Blue force tended to do well in these, but there was no particular effort to match the later experiments to the earlier ones.

The difficulty was that the two sets of experiments had widely varying initial conditions as well as outcomes. In particular, the latter experiments were done with much smaller opposing forces. Yet MCWL wanted to reach a conclusion regarding the new training, not simply a self-evident conclusion that it is better to outnumber the enemy. Specifically, MCWL asked: Were the greater margins of victory attributable to the improved training, or were they simply the result of the greater ratios of initial forces?

Table 3 shows data from events with comparable tactical situations—attacks against a single objective, done using the "Penetration" tactic—from the initial and later phases of the MCWL effort, undertaken at Camp Lejeune and Victorville, respectively.

Table 3: Troop data from Lejeune and Victorville “penetration” scenarios [43]

	Start		Finish		Percentage losses		Blue/Red quality ratio
	Blue	Red	Blue	Red	Blue	Red	
Lejeune	65	40	19	17	71%	58%	0.34
	66	37	52	18	21%	51%	0.63
	64	38	36	13	44%	66%	0.46
	104	101	29	10	72%	90%	1.01
	83	90	50	12	40%	87%	1.81
	69	27	49	6	29%	78%	0.29
	96	89	55	30	43%	66%	1.13
	60	93	4	43	93%	54%	1.90
Victorville	91	20	62	14	32%	30%	0.05
	84	20	72	11	14%	45%	0.15
	88	27	74	14	16%	48%	0.24
	87	20	57	4	34%	80%	0.09

Given the starting and ending numbers of troops and the last of the equations above, one can compute, for each battle, the ratio of the Blue quality to the Red quality according to Lanchester theory. This quantity is shown in the rightmost column of the table. The cases show considerable variation, but although the Victorville cases appear favorable in terms of the percentage of Blue forces lost, the Blue/Red quality ratio in *all* of the Lejeune cases is higher than those in *any* of the Victorville cases, arguing against the case that the new training was better than the old.

This finding regarding the “improved” training is not dispositive, if only because the Lanchester approach assumes that everybody on each side is always able to find somebody on the other side at whom to shoot, and this might not have been true in these battles. Moreover, one could argue against the above result on the basis that the urban terrain at Camp Lejeune was different from that at Victorville. This is undoubtedly true, but it undermines the original contention that the Lejeune experiments constituted a control case for those at Victorville.

General comments on the theories

For military experimenters, the preceding theories are probably a good example of the Einstein quote, “Education is what you have left when you have forgotten everything you learned in school.” The brief presentations are not intended to make the reader an expert on any of the theories; rather, the purpose is to showcase certain theories that have proven useful, and thereby to show what a useful theory can be like.

The theories are quantitative and, in particular, probabilistic. Two probability distributions, the exponential and the Poisson, arise repeatedly. Early operations analysts therefore viewed them as fundamental; we might argue that these distributions saw frequent use because of their tendency to lead our computerless predecessors to closed-form solutions, but their success cannot be gainsaid.

Quite importantly, the theories' mathematical formulations lead naturally to measures of effectiveness, MOEs. They are explicated in published literatures, and they lend themselves well to analysis of military situations.[25] (They are, however, applied more widely than just to military problems—even Lanchester's attrition theory, or something quite akin to it, has been applied by naturalists to predator-prey interactions.)

It is interesting to note the appearance of a trade-off in the theories: in the above account, the best-developed and most useful (e.g., queueing theory) are also the narrowest in applicability. Yet all—even queueing theory—have applicability beyond the purposes for which they were originally invented, and beyond the domains suggested by their names.

It is also interesting to note that some of the theories—e.g., those of queueing and search—address operations that are familiar to us from daily life; for many in today's world of high-bandwidth consumer choices of all kinds, the same could almost be said of information theory. Thus it should come as no surprise that they are applicable in military affairs, though the author has seen cases in which the seemingly everyday nature of information, queueing, and search has led physical scientists to doubt that there could possibly be meaningful quantitative treatments of these topics.

[25] This characteristic is not to be taken for granted; there are similar well-developed mathematical theories, e.g., that of branching processes (see Harris), that have yet to see military application as far as the author knows.

While knowledge of the various theories described above will not guarantee the analyst success in his or her military experiments, such knowledge cannot hurt, and ignorance can. In addition to the occasional payoff of being able to apply an equation or result directly, the process of having learned to think in terms of these theories will benefit the analyst by leading him or her to the habit of contemplating all operational observations in theoretical terms. Of course, this habit is presumably also inculcated by the analyst's prior scientific training, but the theories cited above, with their operational orientation and explicit inclusion of people in the objects of study, may point the way even more clearly.

Devising new theories

Having seen a number of successful theories and then read a statement that these will probably not suffice to support experimentation, the reader is doubtless wondering if she or he is going to have to create a new theory, and wondering how one would set about doing so.

The theories cited as examples are all quantitative, so the analyst trying to create a new theory would doubtless identify all the measurable or countable quantities available, and measure or count them. Then what?

Two systematic precepts are suggested in the literature of operations analysis: dimensional analysis (and therefore scaling), and variational analysis. Unsurprisingly, each is borrowed from physical science.

Dimensional analysis and scaling

The first systematic precept is to examine the quantities and consider the *dimensions*—length, mass, time, etc.—in which they are measured.

One way to use these dimensions is to use them to find out how to combine the quantities via multiplication and division so as to create *dimensionless constants*. In many cases, these will simply be ratios of pairs of variables having the same dimension. Notice that many of the theories above hinge on dimensionless quantities, such as the signal/noise ratio in information theory, the ratio λ/μ of service rate to arrival rate in queueing theory, and the quality ratio a/b in Lanchester theory. There are mathematical reasons which practically guarantee that such ratios will be important, and their use can be considered to be under-treated in undergraduate physical science education.[26] H.E. Huntley's 1967 book, *Dimensional Analysis*, shows many examples of how dimensional analysis can be used to solve undergraduate physics problems.

An instructive example of the use of dimensionless variables as touchstones in analysis is given by Morse and Kimball in their study of U-boat "circulation" in the North Atlantic during the Second World War. The problem was non-trivial because the *rate* at which U-boats could be repaired (at ports in occupied France) depended upon the *number* present: the more boats there were in-port, the more that could be worked on, but space was scant and there arose a noticeable clogging effect, as a result of which the repair rate suffered from diminishing returns to scale. Accordingly, the analysts saw the important variables as

B = the number of U-boats at the base(s),

C = the average rate (in boats/month) of U-boat repair in a lightly-filled base, and

M = the maximum per-month rate at which U-boats that could be repaired.

These combine into the dimensionless quantity CB/M, which the analysts used (in concert with their predilection for Poisson and exponential distributions) to create a formula for the rate at which boats could, for varying values of B, be repaired and returned to sea:

$$L = M(1 - e^{-CB/M}).$$

[26] Some have suggested that this is because if the students knew to search for dimensionless constants, elementary physics problems (e.g., the solution of the physical pendulum) would be too easy!

This rate and various expressions that amount to conservation-of-U-boats laws then became a U-boat circulation model. [44]

A related way to use the dimensions is to combine them according to geometrically-derived *scaling laws*. This approach was used by analysts at MCWL to create an a priori estimate of whether an airborne video camera would be able to detect vehicles or dismounted troops. The analysts compared the size of the camera's field of view in pixels to the size of the region of ground it covered, in feet, and applied a rule-of-thumb to the effect that an object must occupy at least 20 pixels in order to become recognizable. They concluded that from the altitude at which the camera was being flown, it might barely detect vehicles and would be unlikely to detect dismounted troops.

Variational analysis

The second systematic precept is to analyze the data so as to identify the results of variations. This advice is given by Blackett and by Waddington, [45] who suggest considering a military situation in terms of a "yield," Y, and causes X_1, X_2, X_3, … X_n. The response of Y to changes in the various causes can be observed, and then summarized as the total derivative

$$\Delta Y = \frac{\partial Y}{\partial X_1}\Delta X_1 + \frac{\partial Y}{\partial X_2}\Delta X_2 + \frac{\partial Y}{\partial X_3}\Delta X_3 + \ldots \frac{\partial Y}{\partial X_n}\Delta X_n .$$

Obviously, this approach pre-supposes the availability of a large amount of data. The struggle against the U-boats having resulted in a great amount of data from a large number of cases that were essentially comparable, Blackett, too, gives a U-boat example. The situation is one of aerial attacks on U-boats that have been sighted on the surface. The aircraft can detect the U-boats from quite a distance away, but as the airplane approaches the U-boat, the U-boat's crew might well also sight the airplane and begin to submerge.

Once the diving process starts, it takes 45 seconds to complete.

A change in the color of the aircraft had been proposed, and test-range experimentation had suggested that it would reduce by about 20 percent the range at which the U-boat crew can sight the aircraft. The question was, how much good would that do?

At first glance, the problem seems nearly imponderable: reducing the range at which the U-boat can detect the airplane can't hurt, but how can one estimate the amount by which a 20 percent reduction will help when the range is not known?

The "yield," Y, is in this case the probability that the U-boat is not yet submerged by the time that the airplane is in position to release its bombs or depth charges. Blackett points out that no amount of a priori reasoning can derive the distribution of Y; it must be gained from experience. Blackett had an ongoing war from which to gather such data, but in peacetime one might resort to experimentation. Table 4 shows the lengths of time that U-boats had been observed to be submerged as of the arrival of the airplane at the point where it would release ordnance.

Table 4: Distribution of U-boats' submergence times [46]

U-boat state when aircraft reaches drop point	Percentage of occurrence
Not yet submerged	34
Submerged less than 15 seconds	27
Submerged 15-30 seconds	15
Submerged 30-60 seconds	12
Submerged more than 60 seconds	11

Blackett supplies a solution in words, which we may translate into the format of the variational method. [47] Let us define *f(t)* as the density of the probability that the U-boat submerges *t* seconds before the arrival of the aircraft at the drop point. Figure 8 depicts this function in such a way as to reflect how little is known about it: it is known only through the fact that it must be decreasing (because, once submerged, the sub remains submerged) and through the values of its five cumulants *F(t)* shown in the table above. Note that, in particular, we do not know the value of –T, where the function really begins.

The "yield," *Y*, is the probability that the submarine will still be on the surface as of the arrival of the aircraft. Letting *x* be the time that the airplane is visible to the submarine, we can write *Y(x)* in terms of the figure:

$$Y(x) = \int_{-T}^{x} f(t)dt = F(x) - F(-T).$$

Then

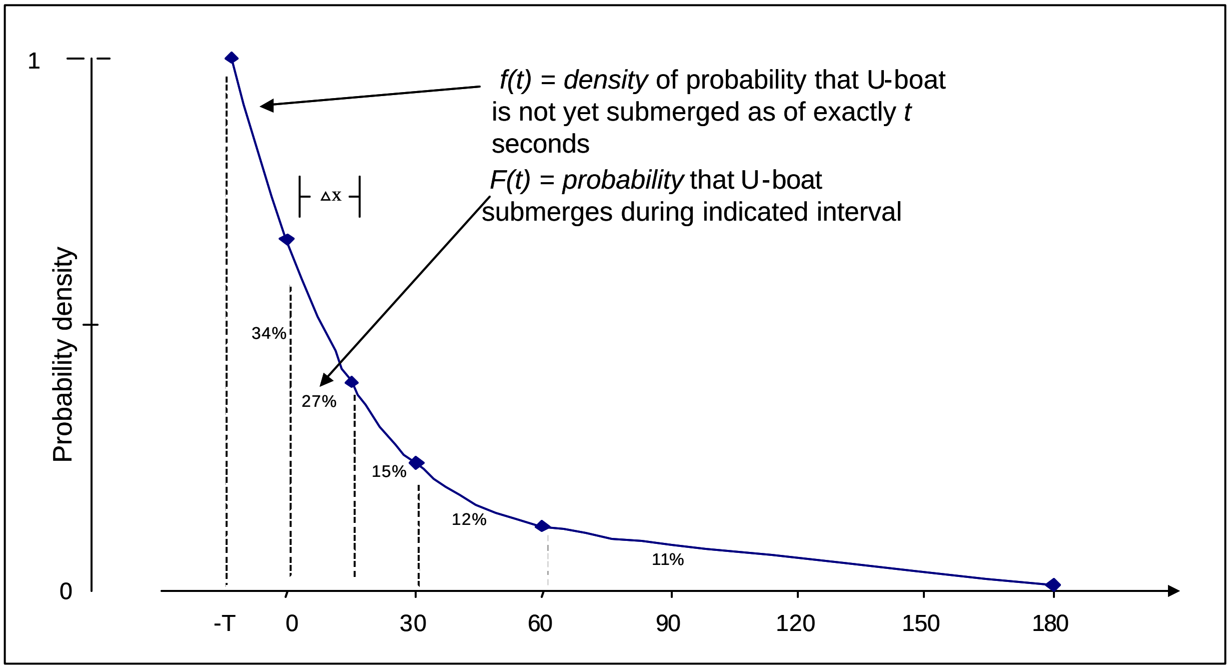
$$\frac{\partial Y}{\partial x} = f(x)$$

and

$$\Delta Y = \frac{\partial Y}{\partial x}\Delta x = f(x)\Delta x,$$

rendering the unknown *T* irrelevant.

Figure 8. Probability of U-boat submergence during air attack



Because the submergence process is assumed always to take 45 seconds and the boats are in widely varying states of submergence when the aircraft passes over, there is clearly a great deal of variation in the range at which the aircraft is being sighted. The wartime analysts reasoned, however, that the 20 percent reduction can help only in the cases of the boats that are just recently submerged as of the passage of the aircraft. Suppose that such a boat has been submerged for $\triangle x$ seconds by the time that the airplane passes, but that a reduction of the approach time by 20 percent would cause the boat to be caught in the act of submerging, i.e., to have just exactly 45 seconds of warning. Then we have

$$0.8\ (\triangle\ x + 45) = 45,$$

So $\triangle x$ = 11.25 seconds. The improvement is $f(0)\ \triangle x$, and although we don't know $f(0)$, we can estimate $f(0)\ \triangle x$ by noticing that it should be somewhat less than the 27 percent that is the area under $f(t)$ between t = 0 and t = 15. (Note that this method of estimating $f(0)\ \triangle x$ somewhat compensates for the shortcoming of the linear approximation.) The modified camouflage can therefore be expected to add attacks on about 25 percent of the submarines, in addition to the 34 percent already receiving attacks.

This rather involved line of reasoning is a remarkable combination of formalism and deduction; one might well have thought that no solution was possible, given the scant data. The wartime analysts backed the change of camouflage, and a larger-than-predicted improvement occurred. [48]

Creativity and inspiration

On a higher plane than that of particular methods and examples, much has been written about the process of creating, replacing, proving, and disproving theories. As such, it has much more to do with "theories" that are "hypotheses" than with theories that are the "systematically organized knowledge … devised to analyze, predict, or otherwise explain." Although it could be scorned on the grounds of being "theorizing about theorizing" or worse, one would do well to read the centerpiece, Thomas Kuhn's *The Structure of Scientific Revolutions*. Mark Blaug's readable *The Methodology of Economics* also contains much that is of value to military operations analysts. Wilson's *An Introduction to Scientific Research* focuses much more on experimentation than the others; Kaplan's sizable *The Conduct of Inquiry* is written from the standpoint of the non-quantitative social scientist; and Jaynes's *Probability: The Logic of Science*, mentioned earlier in connection with its exposition of probability theory in terms of everyday thought, merits a converse mention here as an exposition of thought in terms of everyday probability.

Thomas Edison said that genius is 1 percent inspiration and 99 percent perspiration. As means of inventing theories, the method of dimensional analysis and the variational method place 100 percent reliance on perspiration, with no room for inspiration. Blackett writes discouragingly, if not disparagingly, of the "a priori" method, characterized as an "attempt to find general solutions to certain rather arbitrarily simplified problems" from "first principles," though he admits that "in times of peace, when up-to-date numerical data on war operations are not available, this method may alone be possible. [49] But surely there is room for some creativity and inspiration: indeed, the theories of search, detection, information, queueing, etc., could not have been created by the variational method, and very probably not by dimensional analysis either.

Indeed, one of Blackett's own great successes was a mixture of the variational method, dimensional analysis, thought experiment, and inspiration: the realization that what really mattered about convoys was the ratio of the number of merchant vessels in them to the number of escort vessels surrounding them, and that therefore convoys should be made as large as possible. [50]

Methodology[27]

> A fatal ambiguity surrounds the expression "the methodology of..." The term methodology is sometimes taken to mean the technical procedures of a discipline, being simply a more impressive-sounding synonym for methods. More frequently, however, it denotes an investigation of the concepts, theories, and basic principles of reasoning of a subject, and it is with the wider sense of the term that we are concerned ...
>
> —Mark Blaug

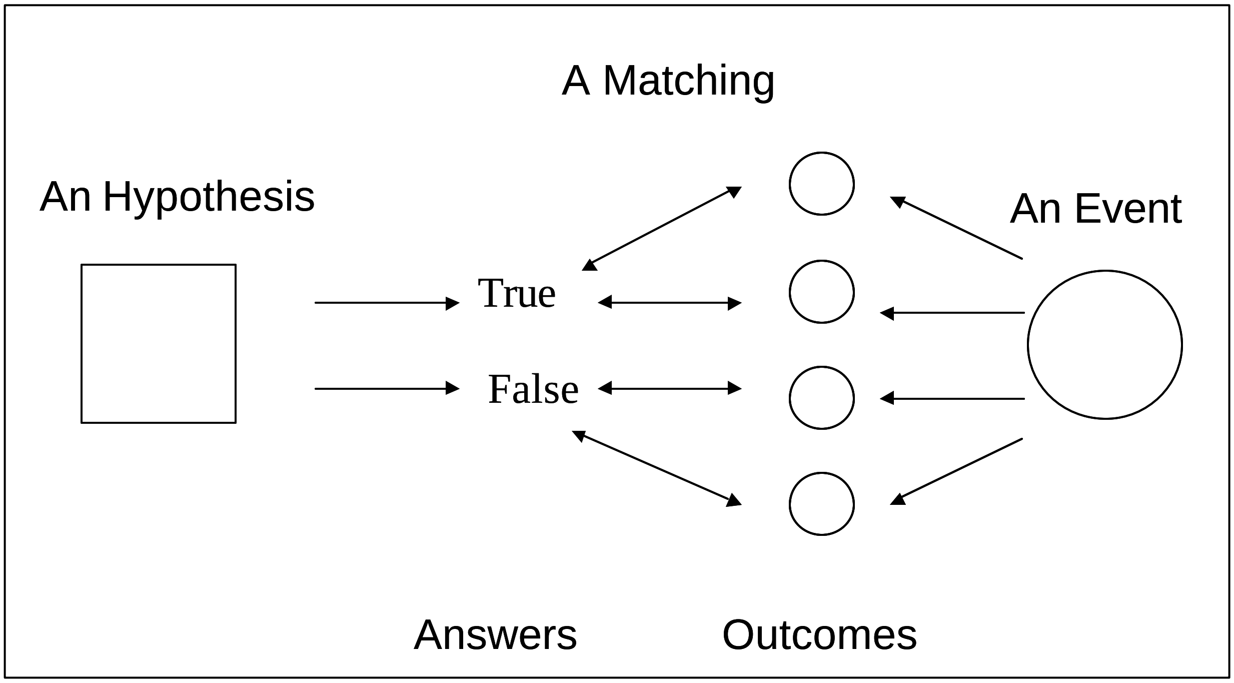
In a sense, of course, this entire paper is about methodology, but this chapter will address formalities of methodology such as hypotheses, measurement, the base case versus the experimental case, and the need for iteration.

Hypotheses

A *hypothesis* is a statement whose truth is to be tested by an experiment. As shown in figure 9, the possible outcomes of the experiment reduce to two: those that confirm the hypothesis, and those that show it to be false. Experiments (or sub-experiments) designed around hypothesis-testing are therefore easy to understand, but nonetheless they can be quite difficult to set up.

[27] See also the section entitled "Methods," in *The Practice of Military Experimentation*.

Figure 9. Schema of a hypothesis-testing experiment



A standard MCWL hypothesis was that some specified capability would be of benefit to an expeditionary Marine force. The trouble with this hypothesis was that the capability, as such, was almost certain to be of benefit, but the more important issue—whether it was of sufficient benefit to be worth doing—went untouched. In fact, it is difficult to imagine how the costs (in terms of not only money but also e.g., the additional space, maintenance, training, etc., that the proposed capability would entail) could be incorporated into a field experiment at all.

Analysts seem to be much better than others at determining whether a proposed experiment will actually shed any light on the hypothesis it is supposed to address. Also, it is the analyst who will eventually have to write the report about the experiment, for which reason alone the analyst should have the final say in whether the activity is well matched to the objectives.

Indicators and instances

In terms of figure 1, a basic hypothesis-testing experiment has as its question, "Is the hypothesis true?" and it has as its answers, "Yes," and "No."

The experiment's event could, of course, turn out in many possible ways, but with only two answers to match, these have to be grouped into those that confirm the hypothesis and those that deny it. Such *binary event outcomes* are called "indicators." A very simple experiment could have just one indicator, but it is preferable to have more than one, in recognition of their possible imperfections.

For example, one of MCWL's many experiments with handheld Common Tactical Picture (CTP) devices sought to test the hypothesis that such devices would help in night combat. The experiment consisted of two night attacks: one conducted without the handheld devices, and one with them. The comparison was to be made in terms of the following indicators (here simplified greatly from a multi-page description that lined them back to MCCRES Standards from Task 02H.01.06 (Conduct a Night Attack)):

- Did the Blue side win?
- Was there less fratricide when using the CTP devices?
- Were the attackers able to advance to the attack position and then cross the Line of Departure (LD) in a more timely way and with better light and noise discipline when using the CTP devices?

Being binary (yes-no), or based on a small number of possibilities, indicators are easier to observe and record than something that requires counting or measurement. Indicators are manifested in particular "instances," which would be what is actually observed and recorded. For example, particular fratricides, or Marines who got lost on the way to the LD, would be instances of the above indicators.

Capabilities versus technologies

Often at MCWL, the desire would arise to experiment with a particular technology—for example, knee and elbow padding sewn integrally into the uniform. To do so, a hypothesis—e.g., that Marines operating in urban terrain would benefit from having protected limb joints—would be created to fit the desired experiment.

These hypotheses suffered from multiple drawbacks:

- They usually appeared so self-evident as not to be worth testing,
- A person exposed only to the hypothesis would not have enough information to deduce the true nature of the experiment—there could be lots of ways of protecting Marines' joints, and the pads under test represent only one of them—and
- Worst of all, they opened the way for arguments as to whether the stated hypothesis would be well tested by the chosen equipment: "Perhaps padding *is* a good idea, but these particular uniforms are so poorly padded as to provide an inadequate test."

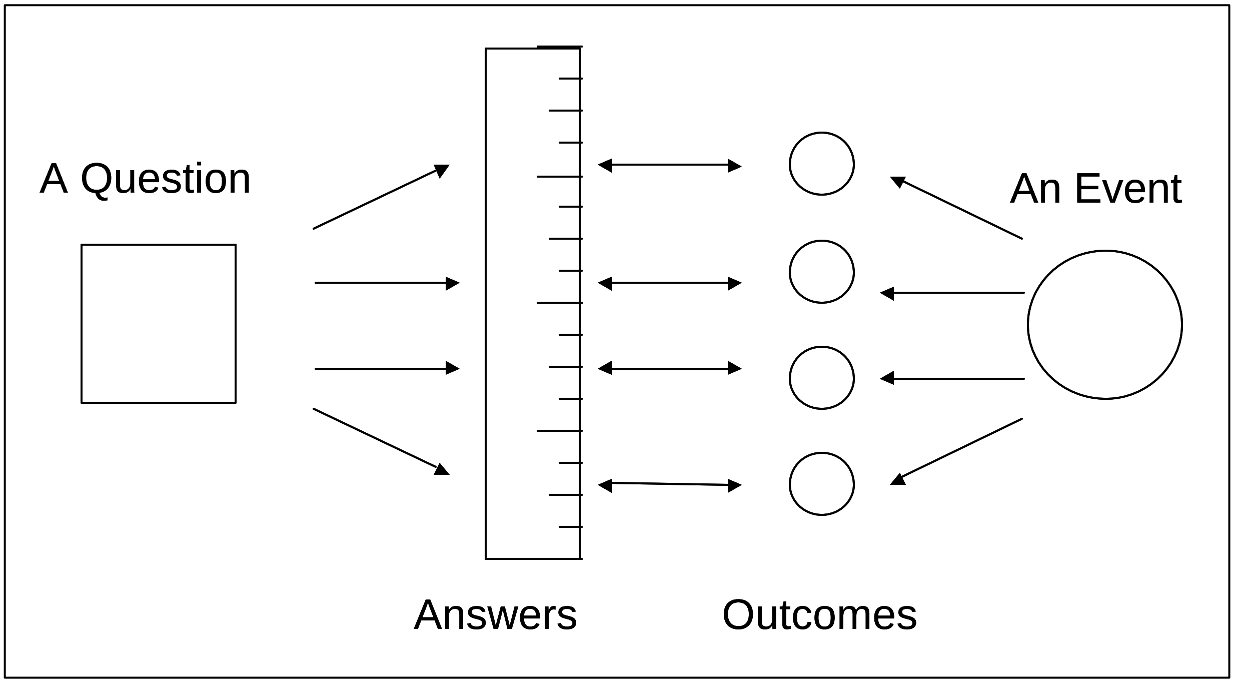
If the desire is to test a particular piece of equipment, one should simply form and test a hypothesis regarding that piece of equipment, not something more general. If what is desired is a test of the utility of a given capability or level of performance, one may need to resort to a surrogate, or an adjudication procedure, to replicate the pure capability, net of any implementation-specific traits.

Measurements

This section refers to physical measurements.

As shown in figure 10, an experiment devoted to *measurement* replaces the finite set of outcomes with a continuum of outcomes. For example, in an experiment devoted to measuring the accuracy of a mortar, the continuum would be the set of positive numbers, i.e., the distances, in meters that constitute possible results for the estimate of the CEP (median miss distance).[28]

Figure 10. Schema of a measurement experiment



The use of indicators might entail measurements, but the measurements are not the final answer to the question. For example, in the experiment with the night attack using the handheld CTP devices, one of the indicators was whether the troops reached the line of departure more quickly with the devices than without. To use this indicator requires measuring the time needed to reach the line in each case, but these times are not in themselves the answers to which the experiment points.

[28] See also McCue, *Estimation of the Circular Error Probable*. It is important for the analyst to realize that to a non-analyst, “estimation” means “guessing;” the resulting collision with statistical terminology can be averted by use of the term “measurement.”

MOEs and MOPs

Measures Of Effectiveness (MOEs) are a staple of military operations research. [51] Yet, despite a great deal of lip service during the preliminary stages, successful uses of true MOEs (as distinct from the indicators discussed above) in military experimentation are rare. The previous chapter has shown several instances in which MOEs have emerged as mathematical consequences of formulating a theory. It does not seem to be a great leap to conclude that the formulation of an MOE requires a formal system-level understanding, deeper than is usually available for military topics, especially when knowledge is still at so formative a stage as to call for continued experimentation.

Experimenters therefore sometimes resort to MOPs. These are quantitative expressions of something that seems to be good, but cannot be justifiably termed "effectiveness." For example, baseball's Runs Batted In (RBI) and Batting Average are (if only because of the confounding effect of the actions of team-mates on a given player's RBI, or on the utility of his being able to get to first base) MOPs and not MOEs: after a century of organized baseball, we do not understand it sufficiently well to know what truly constitutes offensive, let alone defensive, "effectiveness" in that game. An even more extreme example comes from football, in which the MOP of time-of-possession (of the ball) is carefully measured and reported, and then commentators can be heard to disagree on whether time-of-possession is a good thing or not!

LTAs often measure MOPs, such as the CEP, and for them to do so is appropriate.

A frequently used method of creating putative MOEs is to ask the participants for subjective grades, e.g., of the hardware they have been using. This method is fraught with peril: the participants' preferences do not *measure* effectiveness in the sense that arithmetic (e.g., averaging) can validly be performed on the results. The grades may not even be positively correlated with effectiveness: see *The Practice of Military Experimentation*, for several examples in which users' assessments went far awry. [52]

Base case versus experimental case

Experience has shown that if people remember only one thing from school about experiments, that one thing will be the idea variously known as a “base case,” “control,” or “baseline.”

Early MCWL experiments did not have any such feature, and some dismissed them out-of-hand on this basis. Yet an experiment need not have a base case: the Cavendish experiment, for example, did not. The rationale for structuring the early MCWL experiments without a base case was that the hypothesis was that something (e.g., a tactic, a set of technologies) would work or not, and in that context a base case was meaningless.

Later MCWL experiments benefited from the presence of a base case.

As a practical matter, it can help to *do the experimental case first*. One reason for this is that the participants will learn during the experiment, and if the experimental case is done second, it is possible that any improvement is ascribable to learning and not to the experimental tactics or technologies. Of course, if the base case is done second any learning will act to lessen the apparent improvement caused by the technology or tactics, perhaps lessening it to less than zero. Another reason to do the experimental case first is that if the technology fails catastrophically, the base case needn’t be done at all.

If there can be no base case, all is not lost. History, experience, or the wisdom embodied in field manuals can be taken as indicative of the baseline.

Resetting

Sometimes, the events in an experiment unfold in such a way that all value may be lost. The most obvious example is an early defeat of the Experimental Force at the hands of the Opposing Force. If this impends—or after it has occurred—there is really no choice but to stop, alter, and resume the experiment and hope that chance or learning will cause the Experimental Force to do better the second time.

However, there are two great dangers in this course of action.

The first is that the Opposing Force will conclude that the experiment will be repeated until it loses, and accordingly will decide to exert minimal effort the second time, the better to lose forthwith. The second is that afterwards the instance in which the Experimental Force was swiftly defeated will be viewed as an aberration that doesn't really count, and that only its ensuing victory on the second try will be remembered or used in analysis.

These dangers can be avoided if it is made clear to one and all that the first try, in which the Experimental Force was defeated, will be treated as no less valid than the second. The Opposing Force—tired after its efforts, and then frustrated upon seeing victory snatched away from it administratively—will need an especially clear, patient, and understanding explanation of this point. In attempting to give such an explanation, a MCWL analyst resorted to the science fiction concept of "branching time streams," and was surprised by how comfortable the young Marines were with this idea.

However, it is still all too easy to dismiss the disasters, on the grounds that "ExCon said that didn't really happen." Then we have an experimental set-up in which nothing really bad can ever happen to the Blue side, and all the technologies, TTPs, and other experimental innovations will check out as having performed well. Therefore the written report of the experiment will have to make very clear that the "road not taken" however disappointing, was at least as valid an outcome of the experiment as the one dictated by ExCon—more valid, in fact, because it resulted from force-on-force free play instead of from ExCon fiat.

The Japanese planning for Midway included preparation via tabletop war gaming. In the game, land-based American bombers flying off of Midway Island attacked Admiral Nagumo's carriers and sank two of them. The presiding officer reduced the score to one carrier sunk and one somewhat damaged, and then a second round of revision undid even this result, "refloating" the carrier for use in later operations against New Caledonia and Fiji. [53] In the actual battle, American dive bombers flying off of carriers sank four Japanese carriers. Some have suggested that the game (before the adjustments) had been a successful prediction of the American victory, but this view goes too far and neglects the fact that in the game the damage was done by land-based airplanes, whereas in real life it was done by carrier-borne planes. The true lesson of the war game was that carrier warfare is dominated by the offensive, and that an ambush attack can do a great deal of damage to a carrier force. In fact, Fuchida and Okumiya recount that as a result of the game,

> the question was raised as to what plan the [Nagumo] Force had in mind to meet the contingency that an enemy carrier task force might appear on its flank while it was executing its scheduled air attack on Midway. The reply given by the Nagumo Force staff officer present was so vague as to suggest that there was no such plan, and Rear Admiral Ugaki himself cautioned that greater consideration must be given to this possibility. Indeed, in the actual battle, this is precisely what happened. [54]

The real mistake in the Japanese gaming of Midway can be seen in this light. Having lost carriers to land-based aviation in the game, and then having decided (almost certainly correctly) that land-based bombers could not realistically inflict such damage, the Japanese undid the damage; however they lost track of the fact that in their game the Americans had had carriers in the vicinity of the island, and that a player had observed that the game pointed to the concern of ambushes in general.

Even if an experiment's written report is very clear about its having had two outcomes, the point remains that the events unfolding on the ground are more real to observers and participants than those they are told would have happened if experimentation had not been halted. Thus, again, the result is a biased view in which bad things never happen to the Blue side.

Statistics

In military experimentation, there is a constant struggle regarding sample size: the largest sample sizes that the experiment planners can imagine (or afford) are smaller than the smallest sizes that the analysts deem valid.

Statistics courses suggest that a sample size of 30 is about the smallest that can be countenanced, pointing out such cautionary calculations as:

> If we test a sample of 10 new missiles and all 10 fly properly, we may feel confident of the missile, but the 95 percent confidence limits for the proportion defective are 0 and 0.267. [55]

On the other hand, a retired Marine has argued:

> If I drop a rock on my foot, I don't have to do it ten times to be sure that it hurts.

Each argument seems persuasive in its own way. What is the difference?

The difference is that the inventory of missiles is seen as having such a property as "proportion defective," which might lie somewhere in between zero and unity, whereas we tend to think that given stimulus to a foot will either cause pain or not, with the same result every time. Thus any given stimulus needs to be tested at most a single time. This line of reasoning would apply to any test situation in which there is reason to think that the outcome, though unknown, is likely to be the same every time, or very nearly every time.

The cookbook-statistics advice on sample size may not be completely binding on the analyst of a military experiment. A lesser number of trials, so small as not to be very convincing by traditional statistical standards, may be enough for a military experiment because the military analyst's incentives to avoid error are structured differently from those of the traditional scientist. In normal science, confidence intervals are customarily expected to cover 90 percent (or in some cases, even 95 percent) of cases, and large sample sizes may be needed if these intervals are to be reasonably short; the reason is that the scientist is greatly averse to declaring an interval and then being wrong about what it contains. In military matters, a 50 percent confidence interval—e.g., a CEP—can be of great use. Similarly, the scientist is much more willing to run the risk of wrongly rejecting a true hypothesis than to run the risk of wrongly accepting a false one, while in military matters, the situation may be much more symmetric, with the two kinds of error roughly equally to be avoided. Therefore, whereas the usual scientific approach is to cast the antithesis of the hypothesis as a "null hypothesis," and then perform an experiment that aspires to "disconfirm the null hypothesis at the 95 percent level," the analyst of a military experiment may be content with an experiment offering disconfirmation at the 80 percent level, with a correspondingly smaller sample size.

More imaginatively, the military experimenter might, as the result of a preliminary test or expert opinion, be able to invert the relationship of hypothesis and null hypothesis, taking the view that the experiment's hypothesis is be accepted unless refuted by experimental evidence. This procedure would be anathema from the standpoint of classical statistics, but only because of that discipline's unwillingness to incorporate prior information.

Zuckerman's work can be interpreted in this light. After some gruesome tests on animals and examination of the effect that German bombing had on Britain early in the war, he formulated his idea of the effectiveness of bombing and used it to plan the bombing of Pantelleria, as described above; in his examination of the reconnaissance photographs of the island after each day's bombing, he was really only checking for *disconfirmation* of what he already believed.

The fictional detective Sherlock Holmes used similar two-stage reasoning in the story "Silver Blaze": suspecting that a horse was going to be intentionally lamed, Holmes sought *confirmatory* evidence: he looked for newly and inexplicably lame animals in general, on the grounds that the culprit-to-be would need to practice. Finding a number of such sheep, Holmes adopted the belief that the horse was going to be lamed.

In some cases, small samples can be enough to produce a useful result *given the initial state of knowledge*, especially if the initial state of knowledge is far from being correct. While we must assume that military people know their jobs and are correct about most topics, experiments do not address topics at random: experiments may well address topics specifically chosen because of a suspicion that the conventional wisdom is wrong, or is about to become wrong because of some change (e.g., a technological change), or because the experts disagree. An example of this kind of situation appears in the next chapter.

Finally, in military experimentation, samples of more than a modest size may be wasted. [56] The usual "cookbook statistics" account of error and sample size holds that the overall random variance in an experimental measurement is the per-trial random variance divided by the number of trials. [57] But random variance is not the only kind of error; there is also systematic error, or "bias," a form of error that (by definition) cannot be reduced by increasing the sample size and thus will be difficult or impossible to detect, especially if the experimenter lacks a good theory by which to know what to expect. Even with a theory, it is possible for the experimenter to suspect bias but have no idea of its sign or magnitude. For example, one of the artificialities of MCWL's Urban Warrior experimentation was that the MILES gear would not shoot through bulkheads ("walls"). MCWL analysts knew that this artificiality probably introduced error, but had no idea of its probable effect, e.g., whether it tended to favor the attacker or the defender.[29] So we have

total variance = (unknown bias)^2 + random variance / number of trials.

Note that the contribution of "unknown bias" is *not* divided by the square root of the number of trials, and that this bias is squared to make it a variance, like the other terms, and that therefore increasing the number of trials does not help once the unknown bias starts to be the dominant source of error.

Iteration

The term "iteration" can be used to describe the repetition needed to amass enough data for the use of statistics, as discussed in the previous section, but it also has a larger meaning: the process of performing multiple experiments on the same topic. See also the parallel section in *The Practice of Military Experimentation.*

[29] Analysts rejected the argument often made with respect to artificialities, especially this one, that since we were unsure of the direction of the effect, we ought to figure that it would "all even out." To do so would be like the person who, upon being handed a coin and, told that it was loaded, bet that it would come up heads 5 out of 10 times, on the rationale that since he didn't know which way it was loaded he should assume that it was fair.

Serendipity

A large military experiment can lead to an unexpected discovery. For example, MCWL's Urban Warrior experiments led to the unexpected discovery that the urban tactics being taught to Marines had basic flaws. In another example, some of MCWL's Capable Warrior experiments yielded such unexpected discoveries as the huge potential of "killer UAVs." So compelling were these findings, and so difficult was it to reach conclusions about the points that the experiments were supposed to elucidate, that some people understandably came to consider the serendipitous findings of the experiments to be their main product.

From that idea, it was but a short step to the idea that experimentation didn't really require "all that hypothesis stuff," and that simply fielding the ingredients of an experiment would suffice, because serendipitous findings would result. This notion is pernicious, and is accordingly decried at length in *The Practice of Military Experimentation*, under the heading "Discovery learning."

On the other hand, serendipitous discoveries undoubtedly arise. The discovery of the deficiency in urban tactical training was the greatest benefit of the Urban Warrior series of experiments, and it was completely serendipitous.

The role of modeling and simulation

In this section, the terms "modeling" and "simulation" will be used in the specific sense of "computer modeling," not in the more general sense used elsewhere in this document.

Validation

Skeptics often maintain that users of computer models of saying of their result, "It must be true; it came from a computer." Instances of such claims are difficult or impossible to find, however, and the skeptics' position is often little better, in that it boils down to a claim that anything coming from a computer must be false.

Some years ago, the modeling and simulation community finally accepted the need for validation of their models' results, though the standard of validity remains an open question. [58] The idea of validating models against historical battles turns out to be far more problematic—and less dispositive—than it might appear. Historical battles are seldom documented well enough to serve this purpose, and each one took place only once.[30] And in any case, much of the impetus for modeling comes from the desire to explore exactly the modern operations and interactions with which there is little or no historical experience, such as MANPADS-v.-CAS aircraft, the use of PGMs of all kinds, and the use of UAVs, especially in the search for high-value targets such as ballistic missiles.

Non-predictive uses of simulation models

Modeling and simulation can, however, be useful quite apart from any ability to predict future outcomes. Other possible uses include:

- Showing what *could* happen,
- Showing what's important,
- Facilitating exploration,
- Forcing systematic thought, and
- Demonstrating "existence propositions."

Of late, "agent-based" simulations have emerged as a topic of great interest in modeling. In combat modeling in particular, the work of CNA's Andrew Ilachinski has been seminal, especially in demonstrating the proposition that extremely orderly combat behavior is possible without any command hierarchy whatsoever.

[30] Some types of tactical encounter, e.g., that of fighters and bombers, or submarines and convoys have taken place enough times to create meaningful statistics, but there was only one landing at Tarawa, one Battle of the Bulge, and one pursuit of the *Bismarck*.

Using simulation models in military experimentation

One seldom-mentioned virtue of computerized models is their consistency. The computer program can act as a proxy for the otherwise-usually-absent theory of military operations, and—as we have seen—military experimentation benefits greatly from the presence of some theory.

It has sometimes been stated that MCCDC and MCWL have used a "model, test" protocol in developing tactics. According to these statements, MCCDC used Andrew Ilachinski's ISAACS (Irreducible Semi-Autonomous Adaptive Combat) artificial-world model to create tactical innovations that could be the subject of live experimentation by MCCDC's subsidiary MCWL.

MCWL *has* used the computer war game JCATS at an early stage of experimentation with, e.g., Ship-To-Objective Maneuver (STOM) topics.

Some military experimentation efforts, such as that of the Australian Army, have used a "model, test, model" approach. In this, a first experiment is done in a computerized combat model. After results have been obtained, one or more field tests are done with live troops, to validate the findings of the modeling effort. Any shortcomings of the model are rectified, and the model is used again in a final round of experimentation.

This method is seen as combining the best characteristics of *in vivo* and *in silico* experimentation: the use of the computer permits experimentation with large battles and allows for the number of repetitions needed for statistical validity and for experimenting under a wide variety of circumstances, while the field test acts to validate the computer findings.

This page intentionally left blank.

Design and analysis of a military experiment

> Truth emerges more readily from error than from confusion.
>
> —Francis Bacon

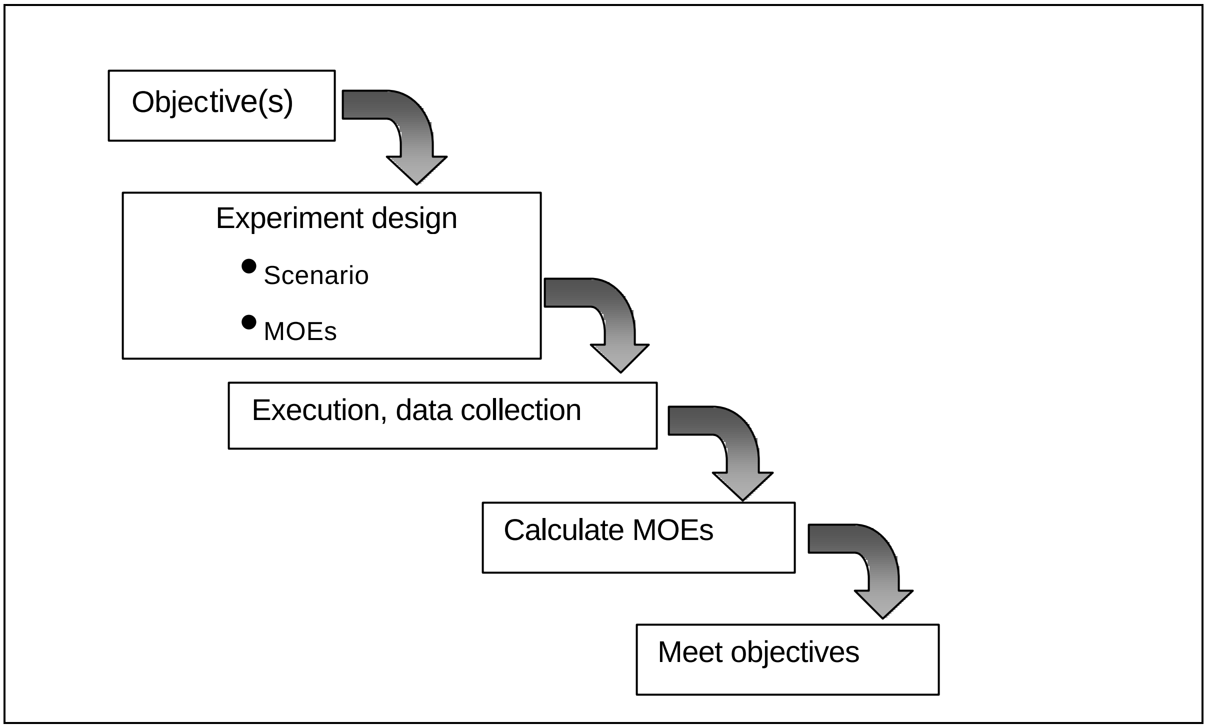
This chapter is constructed around the example of an experiment involving an unattended ground sensor, supposedly capable of detecting troops moving on foot via a nearby trail. The purpose of the experiment is to test this capability. This example is a composite of several experiments, some done at MCWL and some not, with details greatly altered so as to ensure releasability.

A basic view of the experimentation process

Figure 11 shows a basic view of the design and analysis of such an experiment. According to this view, the process is straightforward. The objective was to test the sensor's ability to detect troops. Therefore the scenario was that the sensor would be set up near the trail, with its output transmitted to a receiver elsewhere in a realistic manner. Nine teams, of four Marines each, would move down the trail at irregular intervals over the course of the test day, and the sensor operator would note the times at which the sensor indicated that troops were passing. The MOE was to be the probability of detection, measured by dividing the number of detections by nine, the number of detection opportunities.[31]

[31] As noted earlier, "detection opportunity" is a notorious concept in the analysis of real-world operations. Its meaning is clear, but the analyst usually has great difficulty in assessing the number of detection opportunities. Control over both sides, and therefore a firm grip on the number of detection opportunities, is one of the great benefits of experimentation.

Figure 11. Basic view of the experimentation process



Execution and data collection would then be simple matters, and the analyst's responsibility would be to add up the number of detections and thereby calculate the MOE. The performance of the sensor would then have been measured, and the objectives met.

A more sophisticated view of experimentation

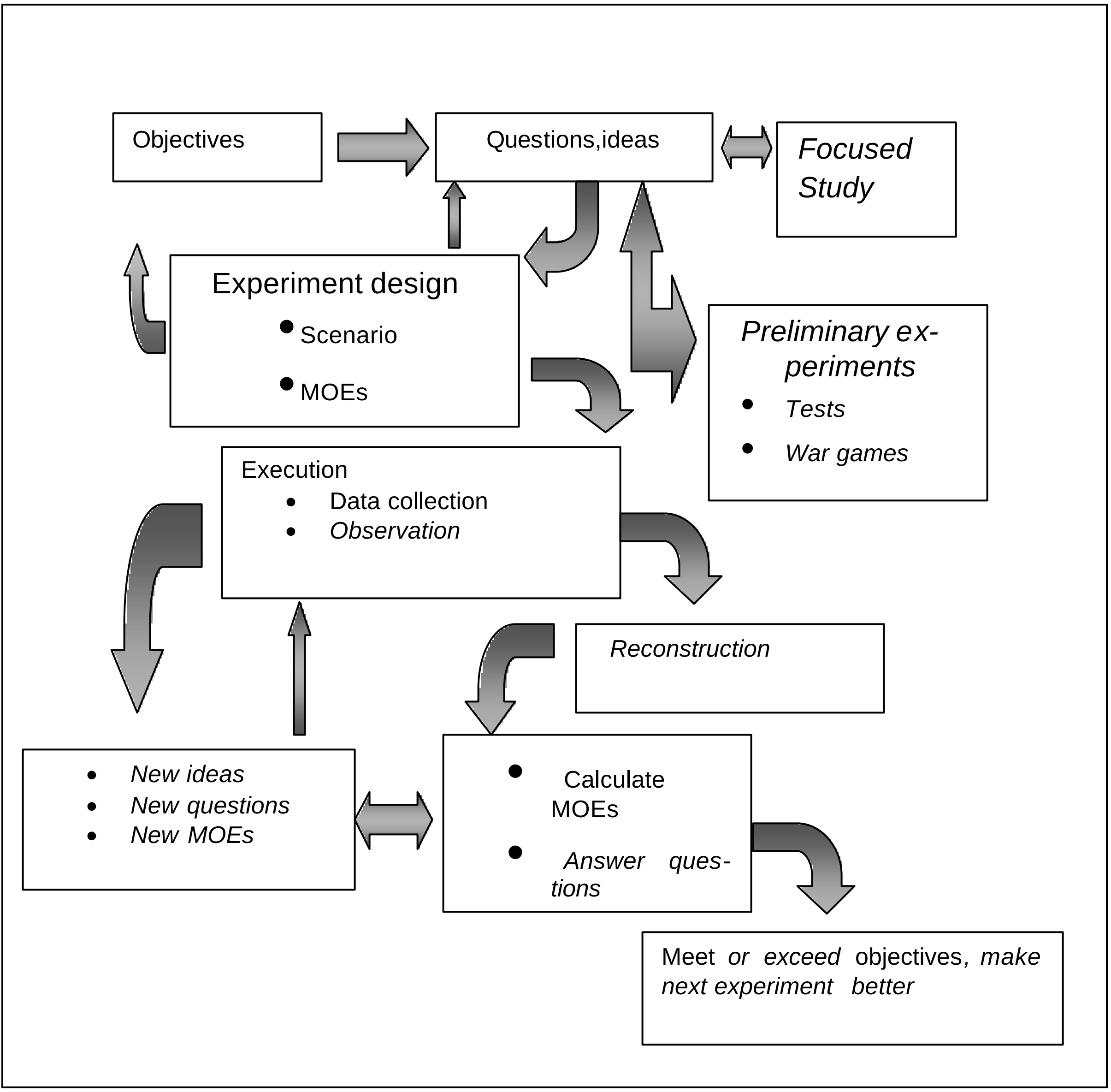
But the "basic view of the experimentation process" is really rather naïve, especially in its view of the analyst as somebody who comes in at the end and does the analysis step, defined as computing the MOE(s).

When the analyst was found and brought into the process—somewhat before the experiment was scheduled to take place—she had a number of concerns. These are detailed below: they address ground truth, MOEs and MOPs, sample size, data collection, and how the data were to be turned into results. More generally, she thought that it would have been better if she had been brought into the process earlier. Whereas the organizers of the experiment subscribed to the "basic view" of experimentation shown in figure 11 and saw "analysis" as adding up the numbers at the end to calculate the MOE, the analyst saw experimentation as shown in figure 12, with the added elements (shown in italics), being the analysis and ideally starting almost as soon as planning itself.

Questions and ideas

"Questions and ideas," for example, include the analyst's realization that she would not have any means of knowing the *ground truth*, i.e., the Marines' movements. She encountered resistance on this point: the sensor itself would do an adequate job of locating the Marines. But the analyst persisted in her desire for separate knowledge of their whereabouts, and argued that if the sensor could be guaranteed to track the Marines, then there was no point in doing the experiment. This argument eventually prevailed, and it was decided that each Marine, or at least one out of each team of four Marines, would carry a GPS device that could record his track. Commercial, off-the-shelf GPS devices from the sporting goods store would suffice for this application.

Figure 12. More sophisticated view of experimentation



One reason to want to know the Marines' whereabouts was to know whether any teams got lost and took some alternate trail other than the one leading past the sensor, which would eliminate a detection opportunity, but another was to be able to detect any *false alarms* that the sensor might produce. The analyst knew from previous experience with sensors that the manufacturers and prospective users focus on the target detection probability, and not on the false alarm rate. But, having studied detection theory and having observed in her field work that false alarms are the bane of real-world sensor operators, the analyst insisted on having the GPS units, and on the use of false alarm rate as a second MOE.

Focused study

Some "focused study" was in order when those in charge of the experiment asked the analyst whether nine teams would provide a large enough sample size. Her reaction to this question was mixed. On the one hand, she knew that a sample size of nine was large by the standards of military experimentation, in which sample sizes of one are not unusual. On the other hand, her training in statistics [59] had left her with the impression that when estimating a proportion (in this case, the probability that a sensor would detect a passing team of Marines) the smallest usable sample size was thirty.

However, in talking with various people about the upcoming experiment, she had noticed a strong divergence of opinion. The manufacturer of the sensor predicted that eight of the nine teams would be detected, but a crusty old veteran of Viet Nam predicted that one would be detected, "if that." This disparity impelled the analyst to raise the question with others. The Chief of Staff said seven, the sergeant whose Marines would be on the trail said three, and the lieutenant said two. Her own middle-of-the-road guess was five. Looking at these answers, the analyst felt reassured about the experiment, because she could do a thought-experiment: *whatever* the number of detections turned out to be, it would differ markedly from somebody's expectation, because the predictions disagreed so much.

Having done hardware LTAs before, the analyst insisted that the sensor be brought to her organization and tested in the region adjoining the parking lot. The manufacturer objected that the device would not be ready in time, leading the analyst to realize that he planned to finish the prototype just in time to take it to the LTA—he had not been planning on testing it himself, figuring that the LTA would be a test in itself. Therefore the "preliminary experiment" in the parking lot had the beneficial effect of forcing the manufacturer to finish the sensor and test it before the LTA.

Observation

After the GPS issue was settled, the data were to consist of the operator's notes regarding detection and the GPS devices' tracks. To the surprise of the organizers, the analyst insisted on a third set of data: *observation* of the experiment herself, sitting with the sensor operator and watching as he monitored the sensor's output. When she did so on the day of the experiment, she noticed that he made only a tally mark whenever the sensor registered Marines on the trail. Without criticizing, the analyst noted the *times* of the detections herself.

At the end of the day, the Marines had all arrived at the collection point at the far end of the trail, and nine detections had been registered. The organizers demanded a "quicklook" report, which the analyst duly produced. It said that the sensor had made nine detections, given nine opportunities, and was therefore 100 percent successful. The analyst included a caveat that this was only a quicklook, and that more analysis would be needed before the results could be considered final. The organizers could not imagine what work could remain to be done in such a clear-cut case, and felt confirmed in their view of analysts as people who revel in needless detail.

Reconstruction

Upon plotting the track data taken from the GPS devices, the analyst discovered that one team had gotten lost and taken an alternate trail, not passing the sensor at all. So eight teams had created nine detections, and the analyst compared her log of detection times to the GPS track data. Two detections turned out to have happened when no Marines were near, and were thus clearly false alarms. Another two occurred when only one team of Marines was nearby. The analyst was tempted to decide that one was a false alarm and one was a true detection, but just to be sure, she viewed the data in such a way that she could see the teams' motions. This view made it clear that one team had, for whatever reason, doubled back past the sensor and then turned around a second time and continued on its way, and was detected two out of the three times that it passed the sensor.

Analysis

Only after having completed this *reconstruction* was the analyst ready to try calculating the MOEs, doing what would be envisioned in the narrowest possible view of “analysis,” rather than the expansive view advocated above. The seven true detections were the result of ten opportunities, taking into account the team that got lost as well as the one that doubled back, and two false alarms had occurred in the six hours in which the sensor was in operation. So the probability-of-detection MOE turned out to be 70 percent, and the false-alarm-rate MOE was one per three hours.

The analyst wondered whether people would be surprised at the results, and remembered that her course in “Bayesian Statistics and Decision Theory” had addressed almost exactly this situation. From the textbook, [60] she adapted a quality-control example to be her analysis of the detector’s miss rate. The “prior probabilities” were the predictions made by the Marines, the manufacturer, and the analyst herself. She thought that the person from the company that built the sensor might have been biased, and that she herself might not have been as knowledgeable about the sensor as the Marines, so she weighted her guess and that of the manufacturer as being half as likely as those of the Marines: she and the manufacturer each had a 10 percent “prior probability” of being right, and the Marines each had 20 percent chances. These percentages added up to 100 percent, though in fact only their relative weights mattered.

This set-up is shown in the first three columns of table 4.

Table 4: Prior and posterior distributions, linked by likelihood of observed results

Person	Predicted Detection Rate	Prior, i.e., p(right)	Likelihood of 7 detections in 10 opportunities	Product of prior probability and likelihood	Posterior
Veteran	1/9	20%	0.002%	0.0004%	0.005%
Sergeant	1/3	20%	2%	0.3%	4%
Lieutenant	2/9	20%	0.2%	0.030%	0.412%
Chief	7/9	20%	23%	4.5%	62%
Builder	8/9	10%	7%	0.7%	10%
Analyst	5/9	10%	17%	1.7%	23%
TOTAL	--	100%	--	7%	100%

With the experiment done and the result (seven detections in ten opportunities) in hand, the analyst could compute, for each prediction, the probability of the observed result if the predicted rate were the true rate and the observed result's departure from the true rate occurred by chance alone. Some of the predictions were far enough from the observed result that, even with such a small sample size, the observed result would be very unlikely if the prediction had been true. The likelihoods not being mutually exclusive or collectively exhaustive, there was no reason to expect them to add up to 100 percent, and they didn't.

But the probability that a particular prediction was right must depend not only on how it squared with the results of a small-sample size test, but also on how likely it was to be correct a priori. The table's fifth column shows the product of each person's a priori probability of being right and the probability that, if he or she were right, the observed results would occur. (Because the second probability is stated as being conditioned on the first, there is no need for a statement of independence.) These probabilities don't have to sum to unity either, but it can be useful to re-weight them so that they do, as in the sixth column. This column shows that, even after only a relatively modest number of trials, the pessimists are likely to be wrong, and the Chief of Staff is most likely to be right—his accurate estimate reinforced his status as a Marine, on the basis of which his estimate was judged a priori to be twice as good as those of the civilian optimists, to give him the most credibility of all.

The analyst was glad she had gone to the trouble of using this Bayesian method. It had provided her with a line of reasoning that let her use not only her small sample, but also the expertise of the Marines while taking into account the fact that their stances, however confidently stated, could not all be right at once. It had also underscored the point that for some purposes, especially the ruling-out of far-off estimates, the sample was not really so small after all.

She wrote up the report in a way that did not make clear who had maintained far-off views of the sensor's probability of detection.

When she sent out the report for review, most of it was well received, but some of the readers wanted to know why she had made such a big deal out of the two false alarms. “That’s only two false alarms out of 11 total alarms, not a very big proportion,” they argued. She had to make clear, not only to them but also in the revised write-up, that the MOE for false alarms isn’t the ratio of false alarms to true alarms, but the rate at which the false alarms occur. The reason is that the number of true alarms is an artificiality of the test: in real life, there might not be any true alarms, but if the sensor and operator behaved in real life as they had in the test, there would be eight false alarms every day.

New ideas, questions, and MOEs

This report led, of course, to a *new question*, regarding the causes of false alarms, and to the *new idea* (or at least a new occurrence of a standard idea) that there must be a trade-off between the false alarm rate and the probability of detection: false alarms could be eliminated completely if all detections could be ignored, and failures of detection could, conversely, be eliminated if one were willing to tolerate a near-infinite false alarm rate. More *new questions* came up, especially after the analyst formalized the group’s thinking by introducing the idea of a “receiver operating characteristic curve,” formalizing the tradeoff between missed detections and false alarms: Where on this curve would it be optimal to operate? How, as a practical matter, would an operator be sure of being at this optimum? The analyst’s study of *decision theory*, as well as detection theory, began to look as though it was going to pay off.

In a more elaborate experiment, sometimes there is time to act on new ideas, questions, and MOEs. This was only an LTA, so there wasn’t enough time to do so. Still, after the analyst completed her report, it was valued not only for its answers regarding the sensor’s performance, but also for its suggestions regarding the next experiment: it said that the next experiment should see whether the false alarm rate could be lowered, and the detection probability increased, by operating a sensor string as one large sensor, filtering out false alarms (and increasing confidence in true detections) by paying attention to the time that passed between the activation of one sensor and the activation of the next one along the trail. The experiment after that should be an LOE in which the Blue side would have such sensors, and would be responsible for setting up a reasonable tradeoff between the effort wasted in sending out patrols to investigate detections that turn out to be false alarms, and the cost of having an OpFor team leak through the sensor network undetected.

Why military experimentation is so hard

> I tested all this with wisdom, and I said, "I will be wise," but it was far from me.
>
> —Ecclesiastes

Military experimentation makes laboratory experimentation look easy. Some of the reasons for this are treated in *The Practice of Military Experimentation*, in the chapter entitled “Obstacles to Effective Military Experimentation.” The difficulties discussed in the present chapter are more integral to military experimentation.

Because of these difficulties, one cannot proceed in a pure fashion, deciding on a suitable question, identifying possible answers, and then thinking of an experiment to match. To a considerable degree, the experiment may be shaped by outside forces and constraints, and the question will have to be modified to match.

Free will

Even a small military experiment involves dozens of people; a large one can involve hundreds. Though subject to military discipline and obedient to orders, each of these persons is a free-willed agent. Indeed, one of the reasons to do experiments is to involve people on this basis: a criticism of computer models is that they do not replicate human behavior, and an experiment (versus a run of a computer model) averts this criticism by introducing the use of human beings. Dictating particular courses of action to the participants can quickly undercut the experiment. In fact, the participants’ previous experience in exercises often renders them hyper-sensitive to seeming hints as to what they are supposed to do—especially since their performance in exercises is sometimes reflected in their personnel evaluations. In conjunction with the fact that much of the experiment represents departures from present practice, the desire of the participants to do what they think is expected will sometimes lead them to draw the wrong draw conclusions in this regard, do something that seems utterly strange, and then explain—to the frustration of one and all—"Sir, we thought that was what we were supposed to do."

This problem can be especially extreme in the case of the Opposing Force, whose exercise experience (on either side) leads them to believe that they are supposed to do all manner of strange things, and eventually to lose.

Yet the participants' ability to do something other than what is expected can wreak havoc with the experiment. The force commanders assert their "commander's prerogative" to do what they want to, not what the experiment plan says they should. The prerogative makes sense, though, only in a real-world case in which the commander will bear the consequences if he fails and therefore deserves to have authority commensurate with his responsibility. In an experiment, with no lives at stake and tactics imposed from above, for the purpose of advancing the goals of the experiment, commanders should feel neither the need nor the prerogative to act entirely as they see fit.

We have seen a number of experiments obviated, or nearly so, by participants' prerogative to use free will:

- When we wanted to observe the results of giving intra-squad radios (ISRs) to all hands, the experimental force, though supplied with enough radios to do this, deemed that they only wanted to operate on a single channel, and that there would be too much traffic on that channel if ISRs were handed out below the squad-leader level. Accordingly, only squad leaders and above had radios, obviating experimentation with an intra-squad radio capability.

- It was difficult to get the experimental force to use the experimental Penetration tactics—they thought Penetration was a bad idea, and they didn't want to do it.

- The experimental force refused to try the "squad leader call for fire," as recounted in some detail below.

To get desired things to happen without impinging upon the participants' free will, MCWL designed experiments with many (e.g.,

50) analytic objectives, in the awareness that perhaps only a third or a half of these would be supported by the experiment as it unfolded.

The need for cooperation and "buy-in"

A military experiment depends on the cooperation of many people. The troops must come from some particular unit; the experiment must take place in a particular locale; perhaps there is experimental gear may need to be borrowed for the occasion; and so on. Each step involves the enlistment of the aid of one or more individuals. Some of them will explicitly ask for something in return for their cooperation—"You can use my Marines if you promise me that your experiment will give them at least one day of training in doing assaults," or "You can borrow these radios if you promise that your Marines will try to use them in dense foliage at a range of at least 1.5 kilometers," or "You can use my building for training if you promise not to break any of the glass." Some may not ask for anything in return, but they will perform better if they buy into the premises of the experiment and don't think that it is pointless.

To get the cooperation or buy-in of these people, the experimental event may have to be altered. The notional quotes in the preceding paragraph suggest typical alterations that might be required.

During the execution of the experiment, the continued buy-in of the participants is essential. Of course with or without buy-in, they will continue to follow orders, but more than that is required of them, and it will not be forthcoming without their buy-in. This situation is perhaps most acute for members of the Opposing Force: if they conclude that the deck is stacked against them and they will lose no matter what, they will simply lose in the easiest possible way and stand by to be dismissed. Such a course of action would ruin the value of the experiment; it has very nearly happened in more than one MCWL experiment.

Participant rebellions

One MCWL experiment, in the Urban Warrior series, had as an objective the exploration of something called “squad leader call for fire.” The idea was that a variety of proposed technologies held out the promise of making calls for fire (e.g., artillery, mortars, naval surface fire support, and perhaps even close air support) so easy to do that little or no training would be needed. As a result, squad leaders—with whom awareness of a need for supporting fire might well begin—would be able to call for fire themselves instead of needing a qualified Forward Observer or Forward Air Controller. Inasmuch as the supporting fire in the force-on-force experiments was done purely by adjudication anyway, the idea was to let squad leaders call for fire themselves and see how much difference this made.

In a meeting on the very eve of the experiment, personnel from the Laboratory’s attached Special Purpose Marine Air-Ground Task Force (Experimental) professed not to understand what “squad leader call for fire” was supposed to be. The head of the analysis branch explained to them that it meant that the squad leaders could call for fire.

“They can’t do that,” objected the SPMAGTF(X)’s artillery officer, “that’s a whole separate MOE.”

“We’re asking, ‘What if it wasn’t?’” said the analysis branch head.

“OK,” said the SPMAGTF(X)’s commander, “Suppose they do call for fire. Does the company commander have to clear it?”

This was a new and good question, and actually the first fruits of the experiment—the involvement of the company commander, or lack thereof, had not been considered.

“You do it your way,” said the analysis branch head. “Set up a policy that you consider reasonable, tell us what it is, do it, and we’ll keep track of the results.”

“OK, I’ve got my policy,” said the SPMAGTF(X)’s commander. “Squad leaders’ calls for fire have to be cleared with the company commanders. And the company commanders are always to deny them."

Thus ended MCWL experimentation with "squad leader call for fire." Maybe "squad leader call for fire" was a good idea and maybe it was not, but the analysts had a hard time believing that it was so self-evidently bad that it ought not even to be part of a one-day experiment.

If new ideas could be screened by asking somebody's opinion as to whether they're bad, the rejection of the ISR or Penetration out-of-hand would be dispositive. But plenty of new ideas have been greeted with initial rejection by the military (for example, gunpowder, steam-powered warships, and tanks), [61] and the whole point of experimentation is to test new ideas open-mindedly.

Inability to try again

The analyst will find it hard enough to schedule the amount of experimentation needed for statistical significance; he or she will find it impossible to persuade the powers-that-be that something went wrong and that the whole experiment needs to be performed again.

In such a case, the analyst has no choice but to write an honest explanation of what happened and try to salvage as much value from the experiment as is possible. For example, in a test of a mortar, the radar that was supposed to track the rounds' travel downrange did not produce reliable data, but another radar measured the rounds' muzzle velocity, and it did so successfully for each round. One purpose of the test was to measure the precision of the mortar; the analyst made the most of the rounds for which complete data were available, but then also used the measured variations in muzzle velocity—and the mortar's ballistics table—to calculate the dispersion due to variations in muzzle velocity, and reported this dispersion as a lower bound on total dispersion.

In another such case, an AWE-sized experiment was to examine a number of ideas and issues, but some fell through because surrogates failed to work. The analysts had no choice but to state what had happened with those surrogates, and to devote analysis effort to those parts of the experiment in which the surrogates worked.

In a third case, the OpFor in an AWE-sized urban warfare experiment was supposed to hide in the built-up "Mainside" region of Camp Lejeune. This region contained over a hundred buildings, but most had been placed off-limits to experimentation because their regular inhabitants did not want Marines running around in them; a dozen or so buildings remained and were designated as usable by the OpFor. The list of forbidden buildings had to be made known to Blue, so the list of possible hiding places was not difficult to deduce. Of these, about half turned out to be locked up anyway, so the OpFor had only a handful of buildings in which to hide. Again, the analyst had no choice but to describe what happened and the light that it shed on Blue's apparent efficiency in locating the OpFor.

In each case, readers objected. A real laboratory experiment in which the supporting equipment fails would not even be written up. A science-oriented reader accordingly castigated the report of the mortar experiment, and the report on the experiment in which so many surrogates failed, as bad reports because they were reports of bad experiments. To a military reader, the accounts of the experiments smacked of excuse-making, if not whining.

Yet there was really nothing else that could have been done. A real laboratory experiment could have been done again, but military experiments (to include even relatively modest LTAs, such as that with the mortar, and certainly the AWE-sized events) cannot be repeated, and the analyst must work with whatever data have been obtained. Caveats regarding the data are not excuses; they are explanations.

The military person who understands this reasoning is likely to take it a step further and suggest that the objectives of the experiment be re-aligned, to better match what happened. To do so is to take the reasoning too far.

Intelligence

Intelligence play has always been difficult to include in exercises, and it is likewise difficult to include in experimentation. The difficulty is that realistic intelligence play has to include a major force in the shaping of intelligence activities—that is, intelligence "gold" must be buried in a large amount of "dross." (In a more modern metaphor based on information theory Roberta Wohlstetter described the intelligence "signal" as buried in a large amount of "noise." [62]) Truly tactical use of intelligence being the rare exception—operational and strategic uses being the norm—intelligence activities must take place on at least an operational[32] timescale. Probably the only source that can produce enough intelligence "dross" to allow for realistic intelligence play is the real world, though it is intriguing to contemplate the use of computer-inhabiting "artificial worlds" to address this problem.

Most units cannot afford a multi-week intelligence exercise, but a dedicated experimental intelligence activity could afford to do a multi-week experiment.

The other alternative, familiar from exercises, is to abandon the goal of realistic intelligence play and use the intelligence assets as "trusted agents" to feed information into the rest of the experimental force.

Emphasis on winning

Some maintain that nothing is learned except from failure:

> As we develop Joint concepts and conduct experiments, we must take intellectual risks informed by military judgment. We must … look for failure as the metric of intellectual honesty and the hallmark of a vibrant entrepreneurial culture. [63]

[32] *This* use of the word "operational" *is* meant in the military sense, and not in the sense in which the term is generally used in this document.

This view may be a bit extreme, but it is preferable to the all-too-often-expressed view that nothing is learned except from simulated victory, and that an exercise or experiment is therefore a failure if the battle is won by the side representing the enemy.

Everybody wants to be on a winning team, and experimentation benefits from this fact because it impels the participants to make great efforts even though they are not in danger of death or imprisonment if they are defeated, as would be the case in a real war.

However, the desire to see the experimental side win or, after the experiment, to see it depicted in the analysis as having won, can readily overcome the desire to learn something from the experiment. One MCWL analyst had ongoing concerns about this perceived importance of winning, and was overjoyed when the *Commandant of the Marine Corps*, speaking at a retirement attended by nearly the entire MCWL staff, recounted a pre-WW II USMC experiment in which the "American" side's opposed landing was defeated, but the experiment was a success because it led to an understanding of how opposed landings could be done successfully. The analyst thought that even if everybody had ignored an analyst, surely the Commandant's words would carry some weight. But this was not the case, and in fact some of the most egregious instances of seeking victory rather than knowledge (such as a company-sized defense in which the attacking force was only a platoon) still lay ahead.

This problem is neither new nor confined to the Marines:

> General [George C. Marshall] talked with one senator who objected to the money that was being spent on maneuvers. The senator was particularly upset because the troops had made numerous mistakes, and he asked why maneuvers were held with so many errors. [General Marshall] replied, "My God, Senator, that's the reason I do it. I want the mistake [made] down in Louisiana, not over in Europe, and the only way to do this thing is to try it out, and if it doesn't work, find out what we need to make it work." [64]

After Millennium Challenge 02, a major joint experiment, numerous press accounts reported that the officials running the exercise had taken steps to counteract the clever ideas of the commander of the "enemy" (a retired Marine Corps lieutenant general), lest the

Opposing Forces win. Some of those who objected to doing the experiment this way saw themselves as vindicated some time later, during Operation Enduring Freedom (the second Gulf War), when an Army lieutenant general was quoted as saying that the Iraqi enemy did not fight the way enemies in American war games did.

A standard for military experimentation

A military experiment typically entails considerable trouble and expense, so there needs to be a way to tell beforehand whether it is worth doing.

One good test is, “Will this experiment yield a greater increase in knowledge regarding the desired topic than we could get simply by sending the same analysts out to observe an exercise?” The exercise, of course, will take place anyway, so if it would be of more use to the analysts than an experiment, then there should be no experiment.

A possible reaction to this test, once heard in a related context, is “But the purpose of the experiment isn’t to collect data for the analysts.” Then what is it? If it’s to train for the participants, or to demonstrate a system to VIPs, then it’s an exercise or a demonstration, not an experiment. If it’s to gain insight, then those who want the insights should attend and see what they can see, but they would be well advised to wait for a report from the analysts before they reach their conclusions. All too often, people come away with “insights” that were merely artifacts of where they were sitting, which person they happened to speak with, or what they expected to see.

This page intentionally left blank.

The writing of reports

Never express yourself more clearly than you think.

—Niels Bohr

One big difference between a conventional experiment and a military experiment is that if the latter goes badly, there is not usually an option to try it again. Nonetheless, a report must be written, putting the analyst in the position of writing a report regarding a badly done experiment, a task for which academic training provides little preparation, and which is exceedingly difficult to do well. The best approach to writing a good report is to do a good experiment.

The Data Collection and Analysis Plan

Ideally, the Data Collection and Analysis Plan (DCAP) is written into the experiment plan.

The DCAP consists primarily of tables that connect experimental objectives and /or goals to indicators, MOEs, or MOPs via events in the experiment, and then show what data must be collected in order to calculate the indicators, MOEs, or MOPs, and who is responsible for doing so and getting the data to the analysts.

Structured in this way, the DCAP is a useful tool all the way through to the end of experiment execution, because the analyst will often be asked to assess the damage that would be incurred if a particular event were to be removed from the schedule. The DCAP allows the analyst to give a quick answer to this question, phrased in terms of which objective(s) would go unmet.

Of course, some objectives may go unmet even if all events take place. For example, the course of free play may unfold in an unforeseen manner.

It is also important for all to recognize that the analysis process ought not to be limited by the DCAP, either: the analysts should be free to perform analysis of unforeseen topics, if events unfold in such a way as to provide insight into them.

The "quicklook" report

The "quicklook" report is written within 24 to 48 hours after the experiment is completed. It is the bane of analysts, because:

- It has to be written before any serious analysis can be done;
- Those who have read it may feel released from the obligation to read any later analysis product; and
- There is great pressure from the experimenting organization to depict the experiment as successful, with "success" defined as "all experimental objectives were met" *and* "the simulated American side attained simulated victory."

These concerns can be at least partly mitigated by writing a quicklook that emphasizes what was done rather than what it means. Also, the analyst should try to have in mind at least a question or two that can be answered based on data available immediately after the experiment: these questions can be addressed in the quicklook, especially if accompanied by a caveat to the effect that they are preliminary.

The full report

A full experiment report is likely to contain sections answering to most of the following descriptions. To a considerable degree, the structure of the report ought to follow the "inverted pyramid" arrangement of newspaper articles, made famous by *The New York Times*: major items are addressed first, with both supporting detail and background generalities given later. The reader can start at the beginning and stop at any point, secure in the knowledge that everything up to that point is more important than anything thereafter.

Ideally, the intended audience of the full report will include a group of officers serving on an “assessment board,” whose task is to decide what to make of the experiment. This Board and its function are described in *The Practice of Military Experimentation*. [65]

Some sections of a full report may not be appropriate for some types of experiment, or for some particular experiments. For example, a report on an LTA might not have Background or Reconstruction sections, and a report on a larger event might not have a Data annex.

Summary

The Summary comes as early as possible in the report, i.e., before the table of contents and, if possible, even the letter of transmission. It explains the results of the experiment *to the reader who understands the questions that the experiment was meant to address, and how it was supposed to do so.* It ought to be short—perhaps one page long for anything less than an AWE, and two pages long for an AWE.

Introduction

The Introduction sets forth the question(s) that the experiment was meant to address, and explains how it did so. It is for the reader who, without it, would not know enough about the experiment to understand the Summary.

In an LTA, this section might explain how the equipment was employed and why. It may also explain what data were taken and why, without presenting the data.

Some LOEs and AWEs have a great number of objectives. In terms of figure 1, each of these has its own question, answers, and set of outcomes in the experimental event. What unifies them is that they all share a single event. These objectives, and how the LOE was expected to satisfy them, must be described. In a simple LOE, this section might be subsumed into the Introduction; in a more complex experiment, a separate Objectives section might be in order.

Background

The Background section explains the origin and importance of the question(s) addressed by the experiment. It is for the reader who, without it, would not know enough about the experiment to understand the Introduction, much less the Summary. The order of these sections is dictated by the inverted pyramid concept's goal of best serving the most knowledgeable reader, and of placing the most major points earliest.

Reconstruction

This section recounts, probably in great detail, what happened in the experiment's events. In the case of a force-on-force engagement, it will read like a historical account *except* that it is not prefiltered regarding what is and is not important: it describes all known events at equal levels of detail, and points out instances of events or details that are unknown. In the case of a weapon or sensor LTA, each shot or detection opportunity should be recounted here and/or in the Data Annex.

This section explains the experiment's artificialities unless they are so important and numerous as to merit an Artificialities section of their own, as is often the case. Artificialities are treated at length in *The Practice of Military Experimentation*.

Analysis

This section contains the conversion of the data, via the reconstruction, into the indicators, MOEs, and MOPs, as planned in the DCAP.

It also contains parallel development and use of indicators, MOEs, or MOPs not developed in the DCAP.

The usual argument against including analytic content is that most people don't really want to read it. But without the analytic content, the other conclusions and observations will appear to be mere opinion. If the readers see the analytic content, they are far more willing to accept the conclusions.

To use a naval analogy, the analytic part of a report is like the ballast of a ship: even though most people don't see it or care about it, and even though it is heavy, it actually holds the rest of the ship *up*.

Other topics

During the experiment, there will be unforeseen occurrences worthy of analysis. These are analyzed in a section entitled “Other topics.” The principal difference between this section and the preceding Analysis section is that the contents of the Analysis section were foreseen in the DCAP.

For an LOE, this section traditionally includes a casualty analysis. Even if such an analysis has no direct bearing on any of the experiment’s objectives, it serves the useful purpose of indicating, to the knowledgeable reader, the flavor of the combats and/or whether the experiment’s method of conducting the combats was at all realistic. Casualty analyses, when carefully done, have also been of benefit to later experiments, e.g., when Project Metropolis availed itself of Urban Warrior casualty data in order to see whether the Basic Urban Skills Training developed by Project Metropolis was an improvement over the standard USMC MOUT training through which the Urban Warrior participants had passed.

In force-on-force experimentation, a subsection entitled “What Worked for the OpFor” belongs in this section, and will be one of the very few parts of the report, other than the Summary, to be read by every uniformed reader.

Conclusions

This section uses the DCAP’s indicators, MOEs, and MOPs described and computed in the Analysis section to answer the questions set forth in the experiment’s goals and/or objectives.

Observations

This section uses non-DCAP indicators, MOEs, and MOPs described and computed in the Other Topics section to answer questions that were not in the experiment’s goals and/or objectives, but—in the analysts' estimation—deserve to be addressed. The principal difference between this section and the preceding Conclusions section is that the contents of the Conclusions section were foreseen in the DCAP.

Recommendations

An LTA is often a test to see whether a particular technology is ready for use in an upcoming LOE, so a recommendation as to that point may be required. Other recommendations may also emerge, especially regarding follow-on LTAs; many a successful MCWL technology went through multiple LTAs.

An LOE or AWE can be expected to result in recommendations regarding budget, acquisition, training, and/or doctrine. It will also result in recommendations regarding further experimentation: an LOE may generate recommendations regarding future LOEs or the AWE, and an AWE may suggest entire future lines of experimentation.

Data annex

This section contains substantially all of the data taken in the experiment, with only the slightest of processing; in principle, at least, another analyst could take this section and, guided only by the DCAP, recreate the Analysis section.

Providing results in a timely manner

Most non-analysts involved in military experimentation would agree that the biggest problem with analysis results is that they take so long to arrive.[33]

Probably analysis will always take uncomfortably long, but certain palliatives can help.

- One is to insist that the data be made available to the analyst as soon as the experiment has ended.
- Another—related to the first—is to create instrumentation that collects as much of the data as possible in an automatic way. Especially because of GPS, network software, and other recent technological advances, the data-collection step can often be completely automatic, or nearly so.
- Much of the report—certainly the Introduction and the Background, and probably the parts of Analysis that explain how the analysis is done—can be written, or at least drafted, before the experiment. Very possibly they can be "lifted" straight from the DCAP.
- Finally, report-writing can be done better, faster, and with more satisfaction all around if the experiment is easy to understand and runs smoothly. To the degree that the experiment entails surrogates whose performance does not adequately mimic that of the real thing; events that one is supposed to assume did not, in fact, happen; missing data; poorly prepared participants; inconsistently applied manual

[33] Old-hand analysts at CNA are strangely schizoid on this point. Based on their considerable exposure to high-ranking Flag and General officers, they emphasize the need for timely results. But in the same conversation, they will say that in the Good Old Days, analysis took much *longer* because there was no automatic data collection, and that this was Good because the analysts became more immersed in the data while doing the reconstruction manually, and thus produced better analysis.

adjudication methods; and the like—in other words, to the degree that the experiment is not a good experiment—it will be all the harder and slower to write up.

One way to write reports *much* faster is to leave out a lot of the analysis, and then periodically write analytic reports, organized around topics rather than individual experiments. This radical approach was suggested at MCWL, but was not tried, so there is no basis for estimating its effectiveness. However, MCWL analysts have occasionally write ten short synoptic reports on particular topics, summarizing the findings of multiple experiments. These have been quite well received.

Perhaps other solutions are possible, and the analyst attached to an organization devoted to military experimentation is encouraged to seek imaginative new solutions to the problem of providing timely results. Interest in an experiment will peak at the moment the experiment ends, and decay steadily thereafter—the sooner the report is written, the greater the interest with which it will be received.

Organization for experimentation

> Hell, there are no rules here—we're trying to accomplish something.
>
> —Thomas A. Edison

Organizational difficulties have plagued the Marine Corps Warfighting Laboratory. Major re-organizations have occurred approximately once a year. In part, these have been attributable to the unique nature of the Lab; even when—*especially* when—it was functioning reasonably well, it did not act as other USMC entities do, and if a new group of leaders arrived, they accordingly felt impelled to alter the organization.

The Lab functioned most smoothly and was most productive when it first began operating. Some claimed that there was no real organization at all (other than one on paper, to which no attention was paid), but I believe that there actually was an organization, albeit un-written and un-recognized. It had a circular structure and pattern of operation, somewhat resembling that presented as a “design for a Warfighting Laboratory” later in this chapter.

Conversely, when the Lab has functioned poorly, the proposed cure invariably has been a re-organization. The various resulting organizational structures have followed traditional military patterns—either the hierarchical pattern of a combat unit’s organization, or the linear pattern of a paper-processing military staff. Neither is suited to the art of military experimentation.

This chapter will describe the organizational difficulties faced by MCWL, and present a suggested design for a Warfighting Laboratory. This design includes more than just organization: it also includes patterns of operation, which are probably more important than the organization.

MCWL designs and problems

As mentioned above, MCWL organized itself according to different designs and different times. Typically, several branches (e.g., Technology, Plans, Operations, Analysis) and several efforts were devoted to different aspects of warfighting (e.g., Aviation, Ground Combat, C^3), and these were arranged side-by-side or contained one another in some pattern. The particular patterns did not make very much difference because the same problems emerged, with only slight variation, in all configurations:

Loss in the translation

Whenever an experiment plan made the transition from one office to the next (e.g., from Future Plans to Plans, or from Plans to Current Operations), information would be lost. Only two organizational schemes avoided this: that of the small group that had end-to-end responsibility for everything (e.g., MCWL in its earliest days, and Project Metropolis), and a successful scheme in which an Experiment Planning Cell migrated through the organization along with its experiment.

Conflation of in-experiment and day-to-day roles

One common problem was the conflation of elements' in-experiment and day-to-day roles. In some re-organizations, for example, one branch was responsible for day-to-day administration of the Lab's telephones and computers, procurement of experimental C^3I gear, running C^3 equipment in experiments and acting as S-2 in experiments. This made superficial sense in that the necessary knowledge and skills for these functions resided in largely overlapping sets of people; however, the different roles were really not the same and the difficult task of separating the make-believe world of the experiment from the real world was not made easier by, e.g., having the same group of people doing in-experiment counter-intelligence and at the same time giving out visitor badges to VIPs. One particular organizational scheme elevated this fault to an organizing principle, by arranging the whole lab along “G-x” lines, with no distinction at all between in-experiment roles and day-to-day roles.

The role of the Special Purpose Marine Air-Ground Task Force (Experimental)

Experiments of LOE size or larger would require forces, and an organization into which to fit them. Since they were needed only intermittently, the forces could be obtained in an opportunistic manner, but the organization into which they would fit during the experiment had to consist of people who understood MCWL work. Thus there arose the idea of the SPMAGTF(X), whose day-to-day establishment consisted of a battalion's worth of field-grade officers, captains, and senior enlisted, and a handful of first lieutenants. One or more companies' worth of Marines would be chopped to the SPMAGTF(X) when needed for an experiment, and returned to their parent organization when the experiment was over. *The difficulty with this idea was that it left the day-to-day role of the SPMAGTF(X) cadre undefined.* Being energetic, dutiful, inquisitive, knowledgeable, and possessed of great initiative—in short, being United States Marines—they embarked upon their own lines of experimentation, which would then either collide with those of the rest of MCWL, or eclipse them.

The problem of the Analysis Branch

MCWL initially had an Analysis Branch, co-equal with the various other branches such as Technology and Plans.

Most of those in the Analysis Branch were CNA employees, and CNA's field program has repeatedly demonstrated that a huge benefit inheres in physically sitting near the uniformed people with whom one is supposed to be working. On this basis, it was suggested that the Analysis Branch should retain its organizational identity, but that the analysts should cease to sit in the Analysis Branch room and should instead be farmed out to the various other branches. Some argued that while this arrangement would increase the analysts' awareness of what each piece of MCWL was doing, it would decrease their influence because there would be a perception that the Analysis Branch had been abolished, and because individual analysts would lack the total picture.

The change was made. Some time later, the analysts still did not agree as to whether the change had been of overall benefit or not, but they did agree that both sets of expected effects—the positive effects that would confer greater awareness, and the negative effects that would reduce the branch's capability and influence—were stronger than they had expected.

MCWL as a whole repeatedly showed signs of dissatisfaction with its Analysis Branch. One perceptive aspect of dissatisfaction was the assertion that the Analysis Branch provided only analysis, and that a Synthesis Branch was also needed. For some reason there was no attempt to see whether the analysts could also do synthesis, and a separate Synthesis Branch was created, but its final form was not what had originally been intended. Meanwhile, Project Metropolis had been created as a microcosm within MCWL, to continue the urban work after Urban Warrior was ended, and it, too, supposedly had a synthesis capability. At the periphery of the Lab, one or two other synthesis-oriented entities came into being.

"Push" versus "pull"

Another unresolved MCWL tension was that which pitted "technology push" against "user pull."

In "technology push," new inventions or technologies are discovered, a military utility is perceived or at least alleged, a weapon or other device is created, and its use is urged upon the Services. In the case of the Marines, much of the urging occurred at MCWL. At each of several stages (e.g., the inventor trying to get the attention of somebody in MCWL's Technology Branch, the Technology Branch trying to get the SPMAGTF(X) to be willing to do an experiment, MCWL trying to report the device out to Systems Command), there tended to be a strong negative reaction, usually explicitly invoking the term "push," as in: "If this was a good idea, somebody would have asked for it, but here you are trying to push it on us, and as far as we are concerned, it is a solution looking for a problem."

In "idea pull," by contrast, Marines or other Service people would conceive of a need and communicate it to technologists, who would set about creating a device to meet the need. This method is much preferred, at least in principle, and the system of “requirements” and “unfilled mission needs” documents is meant to support a process of innovation via “idea pull.”

The difficulty with the devotion to “idea pull” and the corresponding aversion to “technology push” is that any number of technologies with which the Marine Corps is now very happy came about as a result of technology push. Airplanes are a good example: the Services didn’t ask the Wright brothers to invent the airplane, but were sufficiently impressed to adopt it when given the opportunity. Conversely, very few military inventions have come about in any other way.[34]

Both paradigms—“push” and “pull”—subscribe to the erroneous belief that all innovation is technological innovation, and therefore seek to manage the innovation process by managing the process of device-invention. Yet innovations of tactics, or even of operations (e.g., wolfpack, and Blitzkrieg—see *Wotan’s Workshop*), can be important too, and are certainly fit for experimentation.

Finally, the descriptions of “push” and “pull” lead to the impression that experimentation is a one-time-through process, when in fact all successful experimentation efforts have pointed to the benefits of iteration.

Transition

MCWL had considerable difficulty in “transitioning” its results to such entities as Marine Corps Systems Command or others.

The principal obstacle was, fundamentally, that nobody had told these other organizations that they should expect to receive input

[34] Tanks are a partial exception, in that although Winston Churchill conceived of the tank (as we know it—armored land vehicles, albeit impractical ones, had been on the drawing boards since the time of Leonardo da Vinci, if not before) as one of several means of breaking the First World War’s stalemate in the trenches, he did so from a position in the Royal Navy, and then had to persuade the British Army to take it. (Brodie and Brodie, page 196.)

from MCWL. Thus they perceived MCWL as “horning in” on their work, and were accordingly unreceptive.

This difficulty was most apparent with MCWL’s Technologies Division, which sometimes had trouble finding its focus.

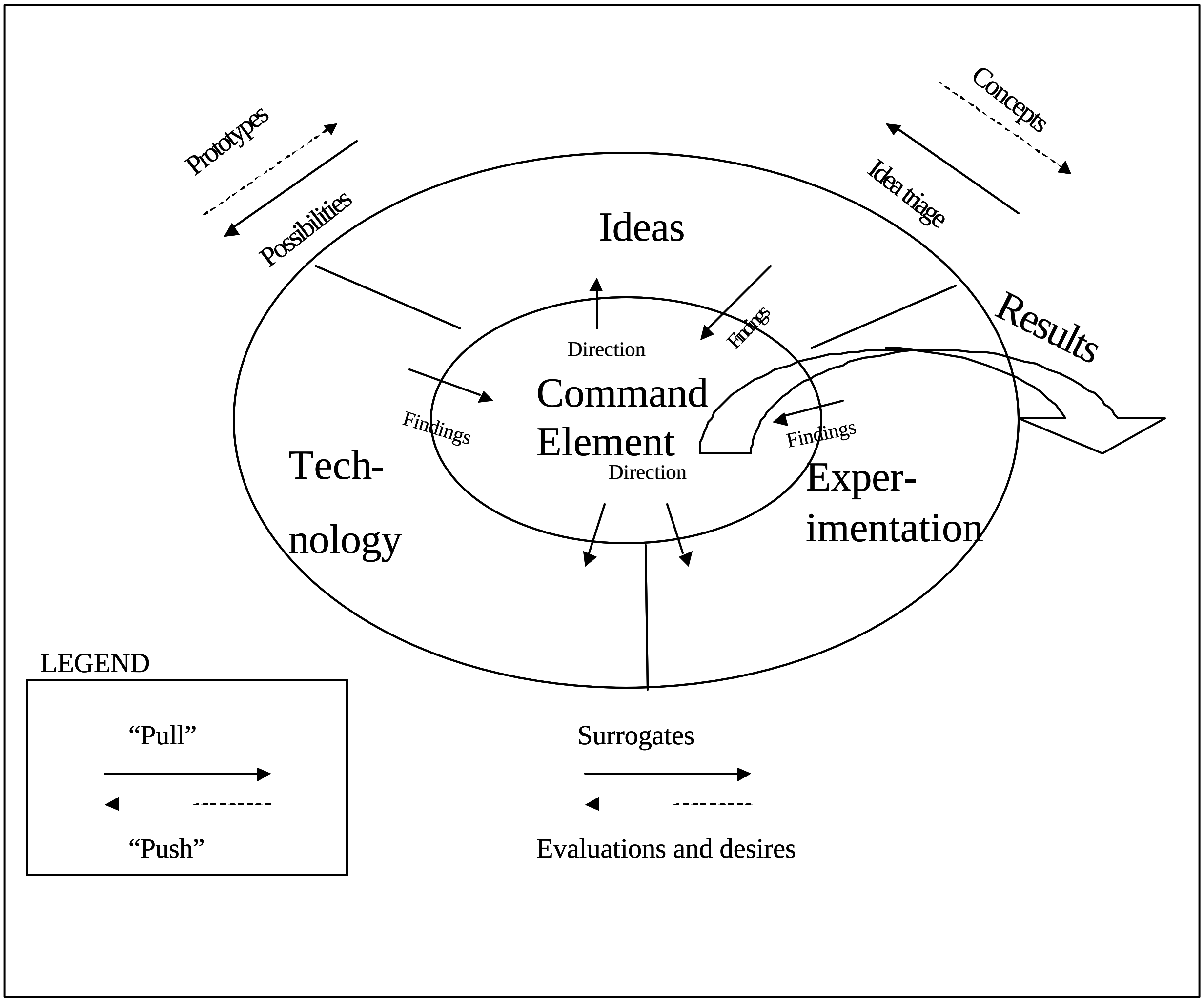
- Existing technologies seemed unworthy of effort because they already existed, along with the Tactics, Techniques, and Procedures (TTPs) to use them.
- Coming technologies had already streamed into the acquisitions process, and to experiment with them would be to interfere with that process.
- Far-future technologies were not the topics of signed MNSes, UNSes, or Requirements Letters, and thus experimentation with them risked the charge of wasted effort and of having an unruly “cowboy” group mentality.

The effort to find focus was not helped by those who claimed, perhaps with justification, that the Laboratory had been set up as a means of circumventing the acquisition process.

Design for a Warfighting Laboratory

An alternative design, shown in figure 13, has a central command core surrounded by three divisions, each in close touch with the other two and the command element. Three divisions are devoted to Ideas, Technologies, and Experimentation. The organization is circular, rather than end-to-end: people, projects, and notions move through the organization in each direction. Work circulates through the divisions, under the control of the Command Cell in the center, and a given piece of work might well complete more than one full circle before the Command Cell pronounces it complete—or it might complete less than one full circle before the Command Cell pronounces it dead.

Figure 13. Design for a warfighting laboratory



These two directions represent fundamentally different functions of the laboratory; linear designs failed because they supported only one function or the other (if any). In fact, the two directions are “push” and “pull.”

The counterclockwise flow is “pull.” The Ideas Division conceives of possibilities, and passes them to the Technology Division, which develops any necessary surrogates and supplies them to the Experimentation Division, which experiments and reports to the Ideas Division. However, this loop need not start with the Ideas Division—it could start with an insight gained by the Experimentation Division while experimenting with something else, or even with the Technology Division if its personnel were to discover a technology (a surrogate, simulation, or method of instrumentation) that would facilitate a long-wanted line of experimentation.

The clockwise flow is “push.” The Technology Division discovers or develops a technology, and passes it (possibly in prototype form) to the Ideas Division, which develops a means of using it. The Ideas Division then passes it on to Experimentation, which does the experiment and creates results on which the Technology Division can act. However, this loop does not need to start with the Technology Division—it could start in Ideas, with a concept for a new technology, that would be passed to Experimentation for war gaming to see if it is a good idea, and then if it is, it goes to Technology so that they can try to develop such a thing. Or it could start in Experimentation, when an experiment causes them to (in a textbook instance of “pull”) realize a need.

People, including analysts, would tend to move around the circle along with their work.

All divisions would produce written reports regarding their discoveries and conclusions.

Ideas Division

This division would have the most challenging task: to think of new ideas. However, it is important to note that the Ideas Division would include people following their projects, in one direction or the other, through the system, not just “idea-people” who remain stationary in the Ideas Division.

The Ideas Division would pass concepts (and the Technology Division’s prototypes that had inspired them) clockwise to Experimentation. It would pass possibilities counterclockwise to the Technology division, which would see whether they were realistic and, if they were, would devise surrogates for the Experimentation Division to use in testing them.

The Ideas Division would be responsible for a considerable amount of writing, since ideas are best communicated as briefings and documents.

While the Technology and Experimentation Divisions would certainly develop their own ideas, the Ideas Division would bear primary responsibility for this important task. Previous experimentation efforts have derived ideas from other branches of the armed forces, the militaries of other countries, law enforcement, science fiction, the behavior of animals, and the behavior of entities in computer programs (e.g., the ISAAC program of Andrew Ilachinski).

Technology Division

The Technology Division would pass prototypes clockwise to the Ideas Division, which would try to figure out how these might best be used (and then pass the results to Experimentation). It would pass surrogates counterclockwise to Experimentation, where they would be used to test ideas, with the results reported to the Ideas Divison.

In today's environment, a warfighting laboratory's Technology Division would do well to avoid technologies that are anywhere in the formal Pentagon procurement process: nothing that a laboratory can do will affect this process. Nor can laboratories readily become part of the procurement process by starting and running "programs" themselves. The laboratory's technology should focus on technologies that have yet to enter the Pentagon's "pipeline," those that have just emerged or are about to do so, and COTS technologies in which some new promise has been perceived. Technologies already in the pipeline are probably best seen as "givens," and experiments regarding them should focus on the development of TTPs rather than on the development of the technology itself.

But in a future environment, it might be otherwise: one can imagine a laboratory that has tight ties to the procurement process, and can do experimentation to address questions that arise during procurement, even as various studies-related organizations (including CNA) address such questions through analysis.

Totally new technologies can be found by searching for them, e.g., among the fruits of the Office of Naval Research and other basic research funded under the "6.1" and "6.2" rubrics, or simply by opening the door when the purveyors of such technologies arrive. These technologies will be far from ready for use, and experimentation with them will probably consist largely of using surrogates to try to assess whether the envisioned fruits of the research would have utility.

Once in the procurement pipeline, a technology is embedded in a particular product, and therefore in an acquisition "program" managed by somebody outside of the laboratory. These programs defend themselves zealously against meddling by outsiders. When a technological product is about to emerge, however, a warfighting laboratory could beneficially obtain a prototype, pre-production item, or surrogate, and experiment with TTPs.

COTS technologies can also present themselves for fruitful experimentation, either in themselves or as surrogates for a militarized version.

When experiments involve innovations that are not technologies, the Technology Division's involvement would certainly be reduced, but it would probably not be reduced to zero. For example, it would still have cognizance over any surrogates that were to be used, and over instrumentation intended to collect data for analysis.

Experimentation Division

Considerable difficulty can arise from the failure to draw a clear distinction between surrogates and prototypes. In the proposed design for a warfighting laboratory, the Experimentation Division personnel would find it easy to maintain the distinction, because surrogates would come to them counterclockwise, directly from the Technology Division, whereas prototypes would arrive clockwise, via the Ideas Division.

The Experimentation Division would be responsible not only for executing the experiments, but also for designing them. Like the other divisions, the Experimentation Division would include at least one analyst, who would help to design the experiments. Experimentation would also benefit from the presence of those involved with particular ideas or technologies; as these ideas and technologies arrived at the experimentation stage, these people would arrive with them, and their understanding would help in the creation of a useful experiment.

It is important to keep in mind that "experimentation" includes war games and LTAs as well as LOEs and AWEs, and the Experimentation Division would have cognizance over all of these activities; however, those who held the topic of the war games or LTAs close to their heart would follow their projects into the Experimentation Division.

Command Element

The Command Element would have the task of remaining cognizant of the activities of the three divisions. The Command Element would be responsible for deciding that a particular project was as finished as the Laboratory would be able to get it, to pass judgment on the result as presentable and meritorious or not, and—if the former is the case—to transition the result to the appropriate part of the outside world. More generally, it would receive findings from the three divisions and exert direction over them as necessary.

Also, the Command Element would be responsible for communicating results to the outside world. In many cases, this responsibility would be discharged simply by signing out a report written by one of the divisions, and coordinated (through the workings of the Command Element) with the other divisions. In other cases, a report would be formulated within the Command Element, and then coordinated with all three divisions before publication.

Like all divisions, the Command Element would include at least one analyst; assuming that the analysts have some organization of their own, the chief analyst would probably be the one to be in the Command Element.

When MCWL was started (as the Commandant's Warfighting Laboratory), the commanding officer foresaw outreach to three types of outside body: industry, foreign countries' marine (or army) branches, and academia. Cooperation with the first two groups (especially the Royal Marines, the Norwegian Marines, and the Australian Army) proved fruitful, but the involvement of academia was never really attempted. This probably constitutes a missed opportunity, and a re-designed laboratory would do well to make a serious attempt to involve academia. Specific parts of academia that might be interested, each in its own way, could include those involved with modeling and simulation, "virtual reality," the psychology of learning, and defense policy.

Appendix: References from The Practice of Military Experimentation

This appendix lists the references to this document that were made in *The Practice of Military Experimentation*.

Page of *The Practice*	Reference	Page(s) of *The Art*
37	"…the section entitled 'Methodology' in *The Art*…"	Page 85 and following
41	"…how and why a small number of trials may be made to suffice…"	Pages 93-96, and 108
49	"…despite all the inaccuracies and artificialities, the truth can be found."	Page 53 and following
76	"Analysis is treated at greater length in *The Art*…"	Pages 101-110
88	"Report-writing on the part of analysts…"	Pages 121-128

This page intentionally left blank.

Glossary

ASWORG	Anti-Submarine Warfare Operations Research Group
AWE	Advanced Warfighting Experiment
CAS	Close Air Support
C^3I	Command, Control, Communications, and Intelligence
CEP	Circular Error Probable
CNA	Center for Naval Analyses
CNAC	The CNA Corporation
COTS	Commercial Off-The-Shelf
CTP	Common Tactical Picture
DCAP	Data Collection and Analysis Plan
ExCon	Experiment Control
HC	Hydro-Chloric (smoke grenade)
GIGO	Garbage In, Garbage Out
GITO	Garbage In, Truth Out
GPS	Global Positioning System
ISAAC	Irreducible Semi-Autonomous Adaptive Combat
ISR	Intra-Squad Radio
JCATS	Joint Conflict and Tactical Simulation
LD	Line of Departure
LOE	Limited Objective Experiment
LTA	Limited Technical Assessment
MANPADS	Man-Portable Air Defense System
MCCDC	Marine Corps Combat Development Command
MCCRES	Marine Corps Combat Readiness Evaluation System
MCWL	Marine Corps Warfighting Laboratory
MEU	Marine Expeditionary Unit
MNS	Mission Needs Statement
MOE	Measure of Effectiveness
MOP	Measure of Performance
MOUT	Military Operations in Urbanized Terrain

OEG	Operations Evaluation Group
OpFor	Opposing Force
ORMO	Officer, Retired, Military, Omniscient
PDA	Personal Data Assistant
PGM	Precision-Guided Munition
PT&S	Penetration, Thrust, and Swarm
RBI	Runs Batted In
ROC	Receiver Operating Characteristic
SAW	Squad Automatic Weapon
SMAW	Shoulder-Launched Multipurpose Assault Weapon
SPMAGTF(X)	Special Purpose Marine Air-Ground Task Force (Experimental)
STOM	Ship-To-Objective Maneuver
TITO	Truth In, Truth Out
TTP	Tactics, Techniques, and Procedures
UAV	Unmanned Air Vehicle
U-boat	Undersea Boat (*Unterseeboot*)
UNS	Universal Need Statement
USMC	United States Marine Corps
VIP	Very Important Person
WW II	World War II

Notes

1. Sorensen, page 186.
2. Sorensen, pages 126-7.
3. Sorensen describes this idea of Mach's on pages 61-63, and attacks it as overly limiting on pages 74-75.
4. http://galieoandeinstein.physics.virginia.edu/lectures/michelson.html
5. See Bondi, pages 54-58.
6. Quoted in Rhodes, pages 49-50.
7. Ilachinski; see also Levy.
8. Epstein and Axtell; see also Levy.
9. Schelling, Micromotives and Macrobehavior.
10. Zuckerman, page 172.
11. This section drawn from Isely and Crowl.
12. Isely and Crowl, page 36.
13. McCue, The Practice of Military Experimentation, pages 16-17.
14. Dönitz, page 33.
15. Dönitz, page 34.
16. See also McCue, *Wotan's Workshop*.
17. Rhodes, page 50.
18. Plato, *The Republic*, Book VII.
19. *Webster's II New Riverside University Dictionary*, page 1200.
20. See Bondi, page 58.
21. Morse and Kimball (1951), page 130.

22. These *ideas* are discussed by Kuenne, Gardner, and McCue (forthcoming). Their attribution to Dönitz is more problematic; several sources (e.g., Frank, Keegan, and many others) describe Dönitz as thinking these thoughts, often in the context of his loss, in the First World War, of the submarine U-68, but do not provide solid references. Keegan (page 224) provides a reference, but it dates from 1939. On page 223, Keegan mentions and even quotes pack-oriented doctrine statements applied to destroyers, but Dönitz seems to have been largely a recipient of this doctrine rather than a source.

23. Dönitz, page 20.

24. Padfield says there is "even the possibility that some of the exercises were actually designed to study the problem of U-boat surfaced attack. No direct evidence to support this has appeared, but..." and then goes on to list a number of pieces of circumstantial evidence (page 101).

25. Dönitz, page 19.

26. This paragraph drawn from Frank, page 23, and Dönitz, page 21.

27. This point is forcefully made by Davis and Blumenthal in their RAND report, *The Base of Sand Problem*.

28. B.O. Koopman, *Search and Screening*.

29. *The New Hacker's Dictionary*.

30. Blaug, pages 103 and following.

31. See Ilachinski, or Epstein and Axtell.

32. Gause, page 7.

33. Leitch, Champion, and Navein, citing Craig Lewellyn and Eran Dolev, "Health Service Support for Military Operations in Urbanized Terrain," *Medical Bulletin*, volume 2, number 6, June 1985.

34. Marine Corps Warfighting Laboratory, Project Metroplis, pages 35, 36; Coates, pages 324 and following.

35. Gannon, *Black May*. This book is also of interest to analysts on account of its description of the arguments regarding analytic “strategy by slide rule” v. choices based on military intuition.

36. This paragraph drawn from John R. Pierce.

37. Examples are Pierce, page 4, and ff; Raisbeck, page 1; Weaver, p. 8 (in Shannon and Weaver).

38. See Klingbeil and Sullivan.

39. See Mark Sakitt.

40. See Chernoff and Moses.

41. Lanchester, *Aircraft in Warfare*.

42. For example, Weiss, Engel.

43. Marine Corps Warfighting Laboratory, Project Metropolis, pages 32 and 33.

44. Morse and Kimball (1946), pages 77 and following. McCue (1990) provides added explanation, and a Poisson-based line of reasoning justifying the use of the exponential expression for L.

45. Blackett, page 180 and following; Waddington’s book is based on an account written by the British operations researchers immediately after the war, edited by Waddington and including contributions by Blackett and many others. (Waddington, page x). I have taken the liberty of reformatting the equation to conform to modern notation.

46. Blackett, page 189.

47. Blackett, page 191, quoting E.J. Williams. It is not 100 percent clear which of Blackett’s examples are supposed to correspond to each of the methods he presents, but this solution appears to me to be a verbalization of the variational method.

48. Waddington, page 165.

49. Blackett, pages 179 and 180.

50. Blackett, page 231 and following.

51. See Morse and Kimball, and McCue (2001).

52. Page 63, in the section entitled, “Reliance on participants’ opinions.”

53. Fuchida and Okumiya, pages 123-125. Allen, page 122.

54. Fuchida and Okumiya, page 125.

55. Crow, Davis, and Maxfield, page 52.

56. The line of reasoning presented in this paragraph is drawn from Jaynes, pages 258 and 336-338.

57. Crow, Davis, and Maxfield, page 14.

58. See Davis, 1992, or Hodges and Dewar.

59. As embodied in e.g., Appendix Charts II, III, and IV of Crow, Davis, and Maxfield.

60. Winkler and Hays, pages 475 and following.

61. Brodie and Brodie: gunpowder, pages 41 and ff; steamships, page 155; tanks, page 196.

62. Roberta Wohlstetter, *Pearl Harbor: Warning and Decision*.

63. Secretary of Defense Donald Rumsfeld, speaking at U.S. Joint Forces Command, 1 Oct 02, as quoted in CO U.S. FFC Edward Hanlon, and CG MCCDC Robert J. Natter, joint letter to CNO and Commandant USMC, entitled “Naval Operating Concept,” 7 Nov 02.

64. Pogue, page 89.

65. *The Practices of Military Experimentation*, pages 77 and following.

Bibliography

Thomas B. Allen. *War Games.* New York: McGraw-Hill, 1987. Berkley paperback reprint 1989.

Robert Axelrod. *The Evolution of Cooperation*. New York: Basic Books, 1984.

Patrick Blackett M.S., *Studies of War.* New York: Hill and Wang, 1962.

Mark Blaug. *The Methodology of Economics*. Cambridge: Cambridge University Press, 1980.

Hermann Bondi. *Relativity and Common Sense: A New Approach to Einstein.* New York: Anchor (Doubleday Books), 1964.

Percy Bridgeman. *The Logic of Modern Physics.* (New York: Macmillan), 1927.

Bernard Brodie and Fawn M. Brodie. *From Crossbow to H-Bomb* (revised and enlarged edition). Bloomington: Indiana University Press, 1973.

Herman Chernoff and Lincoln Moses. *Elementary Decision Theory.* New York: Dover, 1987. (Reprint of 1958 original, published by John Wiley and Sons, New York).

James B. Coates, et al., of the Medical Department, United States Army, *Wound Ballistics in World War II, Supplemented by Experiences in the Korean War.* Washington, DC: Office of the Surgeon General, Department of the Army, 1962.

James S. Corum. *The Roots of Blitzkrieg*. Lawrence: The University Press of Kansas, 1992.

Edwin L. Crow, Frances A. Davis, and Margaret W. Maxfield. *Statistics Manual.* Originally published by the U.S. Naval Ordnance Test Station, China Lake; republished by Dover Publications, 1960 and later years.

Paul K. Davis and Donald Blumenthal. *The Base of Sand Problem.* Santa Monica: RAND, 1991.

Paul K. Davis. *Generalizing Concepts and Methods of Verification, Validation, and Accreditation (VV&A) for Military Simulations*, Santa Monica, RAND, 1992.

Karl Dönitz. *Memoirs: Ten Years and Twenty Days.* R.H. Stevens, translator. London: Weidenfeld and Nicholson, 1959. (Originally published 1958.)

Arthur Conan Doyle. "Silver Blaze," short story, widely available in many collections of stories about Sherlock Holmes, first published in *The Strand Magazine*, December 1892.

Melvin Dresher. *The Mathematics of Games of Strategy.* New York: Dover, 1981.

J.H. Engel. "A verification of Lanchester's law." *Operations Research*, volume 2 (1954), pages 163-171.

Joshua Epstein. *Nonlinear Dynamics, Mathematical Biology, and Social Science*. Reading: Addison-Wesley, 1997.

Joshua M. Epstein and Robert Axtell. *Growing Artificial Societies: Social Science. from the Bottom Up*. Washington, DC: Brookings Institution Press, and Cambridge: MIT Press, 1996.

David Hackett Fischer. *Historians' Fallacies: Toward a Logic of Historical Thought*. New York: Harper Torchbooks, 1970.

Wolfgang Frank. *The Sea Wolves*. New York: Ballantine, 1955 and many re-issues.

Mitsuo Fuchida, and Masatake Okumiya,. *Midway: The Battle That Doomed Japan, the Japanese Navy's Story*. Annapolis: United States Naval Institute Press, 1992. (Reprint; original date of publication was 1955.)

Christopher R. Gabel. *The U.S. Army GHQ Maneuvers of 1941*. Washington, DC: The Center for Military History, United States Army, 1991 (U.S. Government Printing Office).

Michael Gannon. *Black May*. New York: Harper Collins, 1998.

Gause, G(eorge) F(rancis). *Vérifications expérimentales de la théorie mathématique de la lutte pour la vie*. Actualités scientifiques et industrielles', 277. Paris: Hermann, 1935. Translated as *The Struggle for Existence*, Dover reprint, 1971.

Theodore E. Harris. *The Theory of Branching Processes*. New York: Dover, 1994. (Reprint of original edition jointly published by Springer-Verlag, Berlin, and Prentice-Hall, Englewood Cliffs, 1963.)

David C. Hoaglin et al. *Data for Decisions*. Cambridge, 1982.

James S. Hodges and James A. Dewar. *Is It You or Your Model Talking? A Framework for Model Validation*. Santa Monica: RAND, 1992.

William Honan. *Visions of Infamy*, New York: St. Martin's Press, 1991.

H. E. Huntley. *Dimensional Analysis*. New York: Dover, 1967.

Andrew Ilachinski. *EINSTein: An Artificial-Life "Laboratory" for Exploring Self-Organized, Emergent Behavior in Land Combat*. Alexandria, VA: Center for Naval Analyses, 2000 (CNA Research Memorandum D0002239.A1.)

Rufus Isaacs. *Differential Games: A Mathematical Theory with Applications to Warfare and Pursuit, Control and Optimization*. New York, Wiley, 1965. ("Complete, unabridged republication" by Dover Books, 1999).

Jeter A. Isely and Philip A Crowl. *The U.S. Marines and the Art of Amphibious War*. Princeton: Princeton University Press, 1951.

E. T. Jaynes. *Probability Theory: The Logic of Science*, Cambridge: Cambridge University Press, 2003.

Reginald V. Jones. *Most Secret War*. London: Hamish Hamilton, 1978.
Released in the United States as *The Wizard War*, by the Putnam Publishing Group, New York, 1978.

Herman Kahn. *On Thermonuclear War*. Princeton: Princeton University Press, 1960.

Abraham Kaplan. *The Conduct of Inquiry*. San Francisco: Chandler Publishing Company, 1964, and various later reprintings.

Richard Kass. *Understanding Joint Warfighting Experiments*. Joint Forces Command.

John Keegan. *The Price of Admiralty*. New York: Viking, 1989.

Ralph Klingbeil and Keith Sullivan. *A Proposed Framework for Network-Centric Maritime Warfare Analysis, Naval Undersea Warfare Command*. Newport, NUWC-NPT Technical Report 11,447, 15 Jul 2003.

Bernard Osgood Koopman. *Search and Screening*. New York: Pergamon, 1980. (Originally published in 1946 as OEG Report 56 from the Operations Evaluation Group, a forerunner of The CNA Corporation.)

T. W. Körner. *The Pleasures of Counting*, Cambridge: Cambridge University Press, 1997.

Thomas Kuhn. *The Structure of Scientific Revolutions*. Third edition.Chicago: University of Chicago Press, 1996.

Frederick W. Lanchester. *Aircraft in Warfare: The Dawn of the Fourth Arm*. New York: D. Appleton & Co., 1916. Nearly the whole of the small portion that relates to attrition modeling may be found in volume 4, pages 2138-2157 in James R. Newman, editor. *The World of Mathematics*. New York: Simon and Schuster, 1956, with a commentary on pages 2136-2137.

Steven Levy. *Artificial Life*. New York: Pantheon, 1992. (Reprinted in New York by Vintage, 1993).

Marine Corps Warfighting Laboratory, Project Metropolis, *Military Operations on Urbanized Terrain (MOUT) Battalion Level Experiments, Experiment After Action Report*, Feb 2001.

Brian McCue. “A Chessboard Model of the Battle of the Atlantic,” *Naval Research Logistics*, forthcoming.

Brian McCue. with the assistance of Christine Hughes and Kathleen Ward. *Analysis Planning for a Domestic Weapon-of-Mass-Destruction Exercise*. Alexandria, VA: The CNA Corporation, 2003.

Brian McCue. *Wotan’s Workshop*. Alexandria, VA: The CNA Corporation, 2002.

Brian McCue. *The Practice of Military Experimentation*, Alexandria, VA: The Center for Naval Analyses, 2003 (CNA Research Memorandum D0007581.A1.)

Brian McCue. *Estimation of the Circular Error Probable (CEP)*, Alexandria, VA: The Center for Naval Analyses, 2002 (CNA Information Memorandum D0005820.A1.)

Brian McCue. “Operational Search Rate—the Quintessential MOE.” *Phalanx*, volume 34, number 1 (Mar 2001), pages 6-8.

McCue, Brian. “U-Boats in the Bay of Biscay.” Washington, DC: National Defense University Press, 1990.

Stanley Milgram. “The Small World Problem.” *Psychology Today* 1, 1967, pages 61-67.

Philip Mirowsky. *Machine Dreams: Economics Becomes a Cyborg Science*. Cambridge: Cambridge University Press, 2002.

Oskar Morgenstern and John von Neumann. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press, 1980. (Originally published 1944 and 1947 as by von Neumann and Morgenstern.)

Philip Morse and George Kimball, *Methods of Operations Research*, Washington, DC: Operations Evaluation Group, 1946. OEG Report 54. (Reprint available from the Center for Naval Analyses.)

Philip Morse. *In at the Beginnings: A Physicist's Life.* Cambridge: MIT Press, 1977. (Autobiography of an analyst.)

Peter Padfield. Dönitz: The Last Führer. New York: Harper and Row, 1984.

Thornton Page. "A Tank Battle Game." *Journal of the Operations Research Society of America*, Volume I, (1952) pages 85-86.

Peter P. Perla. *The Art of War Gaming.* Annapolis: United States Naval Institute, 1990.

Peter P. Perla et al. *Gaming and Shared Situation Awareness*. Alexandria, VA: Center for Naval Analyses, Nov 2000 (CNA Research Memorandum D0002722.A2 / Final.)

John R. Pierce. *An Introduction to Information Theory: Symbols, Signals, and Noise.* Second Revised Edtion. New York: Dover, 1980. (Original title, *Symbols, Signals, and Noise: An Introduction to Information Theory.* New York: Harper and Brothers, 1961).

Forrest C. Pogue. *George C. Marshall: Ordeal and Hope 1939-1942.* New York: Viking, 1965.

Fletcher Pratt. *Fletcher Pratt's Naval Wargame*. New York: Harrison-Hilton Books, 1940. Republished Milwaukee, Z&M Publishing Enterprises, 1973.

Gordon Raïsbeck. *Information Theory: An Introduction for Scientists and Engineers.* Cambridge: MIT Press, 1964.

Anatol Rapoport. *N-Person Game Theory: Concepts and Applications*, New York: Dover, 2001. (Original edition 1970.)

Anatol Rapoport. *Two-Person Game Theory*, New York, Dover, 1999. (Original edition 1966.)

Richard Rhodes. *The Making of the Atomic Bomb.* New York: Simon and Shuster, 1986.

Thomas L. Saaty. *Elements of Queueing Theory with Applications*. New York: Dover, 1983. (Reprint of McGraw-Hill edition, 1961.)

Mark Sakitt. *Submarine Warfare in the Arctic: Option or Illusion?* Stanford: Stanford University Press, 1988.

Michael Schrage. *Serious Play*. Boston: Harvard Business School Press, 2000.

Ivan Selin. *Detection Theory*. Princeton: Princeton University Press, 1965.

Thomas C. Schelling. *Micromotives and Macrobehavior.* Norton, 1978.

Thomas C. Schelling. *Arms and Influence.* New London: Yale University Press, 1967.

Claude E. Shannon and Warren Weaver. *Mathematical Theory of Communication*. Urbana. University of Illinois Press. 1949 and many reprintings, sometimes as *The Mathematical Theory of Communication*.

Ray Solomonoff and Anatol Rapoport. “Connectivity of Random Nets.” *Bulletin of Mathematical Biophysics* 13, 1951, pages 107-117.

Roy A Sorensen. *Thought Experiments*. Oxford, Oxford University Press, 1992, paperback edition 1998.

Charles M. Sternhell and Alan M. Thorndike. *Antisubmarine Warfare in World War II.* Washington, DC: Operations Evaluation Group, 1946. OEG Report 51. (CNA reprint, 1977.)

Robert H. Thompson. “Lessons Learned from ARMVAL.” *Marine Corps Gazette,* Jul 1983.

Jeffrey Travers and Stanley Milgram. “An experimental study of the small world problem.” *Sociometry* 32, 1970, pages 425-443.

Myron Tribus. *Rational Descriptions, Decisions, and Designs*. New York: Pergamon, 1969.

John von Neumann and Oskar Morgenstern. (See Oscar Morgenstern and John von Neumann.)

C.H. Waddington. OR *in World War 2: Operational Research against the U-Boat*. London: Elek Science, 1973.

Alan R. Washburn. *Search and Detection*. Second edition. Arlington: Operations Research Society of America, 1989.

Herbert George Wells. *Little Wars*. First published by Frank Palmer, London, 1913. Facsimile edition published by Arms and Armour Press, London, 1970.

Weiss, Herbert K. “Lanchester. Type Models of Warfare.” *Proceedings of the First International Conference on Operational Research, Oxford 1957*. Baltimore, MD: Operations Research Society of America, Dec 1957, pages 82-99.

J.D. Williams. *The Compleat Strategyst: Being a Primer on the Theory of Games of Strategy*. New York: Dover, 1986. (Reprint of 1954 McGraw Hill original.)

Andrew Wilson. *The Bomb and the Computer: War gaming From Ancient Chinese Mapboard To Atomic Computer.* New York: Delacorte, 1968.

Robert L. Winkler and William L Hays. *Statistics: Probability, Inference, and Decision*. Second edition. New York: Holt, Rinehart, and Winston, 1975.

Roberta Wohlstetter. *Pearl Harbor: Warning and Decision*. Stanford: Stanford University Press, 1962.

Gene T. Zimmerman. “More Fiction than Fact—The Sinking of the Ostfriesland.” Warship International, 12, 2 (1975), pages 142-154.

Solly Zuckerman. *From Apes to Warlords*. New York: Harper and Row, 1978.

This page intentionally left blank.

CRM D0010079.A3/1Rev

CNA