

# 1 EVALUATION PRINCIPLES

---

The evaluation framework is based on a set of evaluation principles, which amount to an evaluation philosophy. The principles reflect the hopes, standards, and reasonable expectations of the evaluator, given real-world constraints. They declare the evaluator's position on such matters as why evaluation is conducted, its intended effects on stakeholders, data quality expectations, and what data are deemed important.

## Determine Evaluation Stakeholders

Stakeholders are those with an interest in evaluation. They vary with circumstances, but may include program managers, developers, training evaluators, military decision-makers, and others. Stakeholders must cooperate to make an evaluation succeed. The first step in any evaluation is to determine who these stakeholders are. It is essential to determine what information, obtained during evaluation, will satisfy each stakeholder.

## Define Objectives

An evaluation requires clearly-defined objectives at the outset. An evaluation may be conducted with more than one objective in mind; for example, to satisfy a milestone requirement while simultaneously demonstrating training effectiveness. Further, there must be consensus among stakeholders on evaluation objectives.

## Treat Evaluation As a Process, not an Isolated Event

Evaluations are often thought of as one-shot events that answer a question at a particular point in time. This makes little sense when evaluating complex and expensive LSTS that undergo years of development before becoming operational. LSTS evaluation may occur as a series of evaluation events, culminating periodically in larger milestone events. Given that evaluation cannot be done in a single stroke, the question becomes one of developing a logical progression of events that will support the development and fielding of an LSTS with the greatest possible training effectiveness.

## Attempt To Influence Design and Development

Training experts should play an important role in the design and development of training systems. Historically, this has not always been the case. Training evaluators should be brought into the system design process to influence system design from a learning perspective; that is, to assure that the design provides an adequate learning environment.

## Evaluate Multi-Dimensionally

An evaluation must link together four dimensions:

- (1) evaluation objectives (see Chapter 4)
- (2) time
- (3) evaluation criteria (dependent variables)(see Chapter 6)
- (4) evaluation methods (see Chapter 5)

Evaluation is a process that occurs across time.

Evaluation objectives at one point in time may differ from those at another.

The third dimension, evaluation criteria, may be added to this pair by considering that different sets of dependent variables may be used depending upon the evaluation objective.

The fourth dimension, evaluation methods, may be added to this triad by considering the logical types of evaluation methods needed to collect the dependent variable data.

## Obtain the Best Data Possible

The worth of an evaluation depends upon the quality of its data in terms of relevance, validity, and reliability (see chapter 8). Beware the fallacy that one evaluation method is inherently superior to another. The quality of data obtainable with a particular method may outweigh other considerations.

## Develop Learning Curves

If a training activity can be repeated several times during an evaluation, it may be possible to develop learning curves. The curves show not only that learning occurred or did not occur, but the rate of learning across time. Learning curves are more informative than point measures in determining the course of learning.

## Measure Transfer of Training

Transfer experiments measure the effects of learning in one situation to performance in another. Obviously, the greater the amount of transfer that can be demonstrated to the combat environment, the more convincing the evidence.

## Further Reading

See the following chapter in Volume I for further discussion:

- Chapter 8 (Evaluation Framework) discusses evaluation principles (pp. 113-115).



## 2 PERSONNEL CONSIDERATIONS

---

### Evaluation Stakeholders

Stakeholders--those with an interest in evaluation-- were discussed in the previous chapter. The first evaluation principle is to determine who these stakeholders are.

### The Evaluation Team

LSTS evaluation remains more art than science. Assure that the evaluation team includes people with the necessary experience and technical skills to analyze the problem, apply their expertise, and conduct the evaluation wisely. In general, an evaluation team requires personnel with some combination of the following backgrounds and technical skills:

- Program management-- expertise in the design and conduct of large-scale field evaluations
- Military decision-maker-- expertise in coordination with military end users of LSTS
- Academic or Service laboratory-- expertise in experimental design, judgment-based evaluations, surveys, and analysis
- Operational Test & Evaluation--expertise in conduct of field tests
- Operations research-- expertise in operations research, systems analysis, and mathematical modeling
- Military trainer-- expertise in schoolhouse and field military training evaluation

The exact combination of backgrounds and technical skills depends upon what evaluation methods are used. Regardless of what these turn out to be, all evaluation teams require competent program management and military decision-making skills. Ideally, program managers will recruit personnel with other backgrounds and skills to work on the evaluation as needed.

### The Community of Training Evaluation Experts

In planning and conducting a new evaluation, it always makes sense to consult the experts first. Consult with them and get their advice. If resources are available, recruit experts to work on the project.

Where are the experts? There is a small community of them. It consists mainly of personnel in the Service R&D (research and development) laboratories, Federally Funded R&D Centers, Service operational testing laboratories, and civilian contractors who have conducted training evaluations for the Services and DoD. There are perhaps a few hundred such individuals. Within this community are individuals with experience evaluating LSTS. Acknowledgments in Volume I lists the names of many of the most experienced personnel. In addition, authors of key references cited in Volume I belong on this list.

## Further Reading

See the following in Volume I for further discussion:

- Acknowledgments lists several evaluation SMEs. See Author Index for authors cited in this manual (pp. 183-185).
- Chapter 2 (Building an Evaluation Framework) describes evaluation perspectives of different evaluation participants (pp. 16-17). See also Chapter 7 (Evaluation Criteria) (pp. 104-105).
- Chapter 5 (Evaluation Problem Areas) discusses lessons learned in conducting field evaluations; many of these relate to program management (pp. 77-80).
- Chapter 8 (Evaluation Framework) discusses evaluation stakeholders (p. 113).

## 3 DEVELOPMENTAL PHASES AND MILESTONE DECISION POINTS

---

### Overview

DoD acquisition directives and regulations promote an orderly succession of developmental phases.

- Phase 0 (Concept Exploration) typically consists of short-term studies to evaluate the feasibility of alternative concepts
- Phase I (Program Definition and Risk Reduction) evaluates promising concepts more closely via prototyping, demonstrations, and early operational assessments
- Phase II (Engineering and Manufacturing Development) refines the design, demonstrates system capabilities through testing, and works out the manufacturing process
- Phase III (Production, Fielding/Deployment, and Operational Support) achieves an operational capability that satisfies mission needs

Milestone decision points prior to each developmental phase determine whether to continue or stop development. The decision is based on the evidence presented to the Milestone Decision Authority during program reviews, and typically consists of a combination of studies of training and cost-effectiveness.

These phases correspond approximately to three simple categories: pre-development (Phase 0), developmental (Phases I & II), and post-development (Phase III).

An evaluation *event* is a single evaluation of some aspect of an LSTS. A complete evaluation may include several different evaluation events, staged across time, throughout the LSTS life cycle.

### Further Reading

See the following chapters in Volume I for further discussion:

- Chapter 2 (Building an Evaluation Framework) discusses DoD directives and regulations, developmental phases, and milestone decision points (pp. 22-24).
- Chapter 7 (Evaluation Criteria) discusses DoD guidance on developmental phases, evaluation criteria, and verification, validation, and accreditation (pp. 104-105).



## 4 EVALUATION OBJECTIVES

---

### Overview

The logical beginning of an evaluation is to define its objectives. Note that an evaluation may be conducted with more than one objective in mind. To simplify discussion in what follows, define the time window prior to and including Phase 0 as pre-development, Phases I and II as developmental, and Phase III as post-development. Table 1 summarizes some of the most common objectives for conducting evaluations.

**Table 1. Common Objectives for Conducting Training Effectiveness Evaluations**

CODE	PRE-DEVELOPMENT (A)	DEVELOPMENTAL (B)	POST-DEVELOPMENT (C)
1	Estimate need for new training system	N/A	
2	Predict training effectiveness	Measure training effectiveness	Determine training effectiveness
3	Predict transfer of training	Measure transfer of training	Determine transfer of training
4	Predict user acceptance	Measure user acceptance	Determine user acceptance
5	Support training design		Determine training status
6	Support system design		Evaluate system design

Before system development begins, a decision is made to start development. This decision may be based on a study to estimate the need for a new training system. This objective is represented by the row labeled Code 1 in Table 1.

By far the most common evaluation objective is to predict, measure, or determine training effectiveness (Code 2). Most milestone evaluations are conducted to satisfy this objective. This objective actually consists of three sub-objectives:

- 2A. Predict training effectiveness (pre-development): estimate effectiveness before the training system is operational.
- 2B. Measure training effectiveness (developmental): estimate effectiveness during system development.
- 2C. Determine training effectiveness (post-development): integrate data post-development to reach definitive conclusions about training effectiveness.

Evaluations may be used to predict, measure, or determine transfer of training (Code 3). Of particular interest is transfer of training from the simulator to settings that reflect, in varying degrees, performance in wartime; for example, field training or live simulation training. Also of interest here is the effect of training on unit readiness. This objective consists of three sub-objectives, analogous in timing to those for Objective 2.

Evaluations may be conducted to predict, measure, or determine user acceptance (Code 4). This objective consists of three sub-objectives, analogous in timing to those for Objectives 2 and 3.

Evaluations may be conducted to support training design (Code 5); for example, to select among alternative training strategies. Studies may be conducted to identify and correct training problems. Post-development, evaluations may be conducted to determine training status; for example, how well individuals in a particular MOS (Military Occupational Specialty) are able to perform their jobs.

Evaluations may be conducted to support system design (Code 6); for example, to assure that the design provides an effective learning environment. After development is complete, the design may be further evaluated.

## Further Reading

See the following chapters in Volume I for further discussion:

- Chapter 2 (Building an Evaluation Framework) introduces evaluation objectives (pp. 10-11).
- Chapter 8 (Evaluation Framework) gives an overview of evaluation objectives and shows how they change during system development (pp. 111-113), and discusses each objective with examples of evaluations (pp. 116-130).

## 5 EVALUATION METHODS

---

### Overview

Military training evaluations tend to use one of four main methods: experiment, judgment, analysis, or survey. In general terms, here is how the methods are applied:

- Experiments determine effectiveness based on *observational* data.
- Judgment-based evaluations determine effectiveness based on *human judgments*.
- Analytical evaluations determine effectiveness based on common *analytical techniques* and using common analytical *strategies*.
- Surveys gather data from a sample of a knowledgeable target population and determine effectiveness based on analysis of the collected data.

Each of the methods can be performed in several different ways, comprising a set of submethods.

The frequency of use of the methods varies. The most commonly used method is Experiment. Analysis, Judgment, and Survey are used in far fewer cases. In practice, different methods are sometimes used in combination. Each method and submethod is described below.

### Experiment

The submethods of the Experiment method are based mainly on particular details of how the experiments are designed. Some features of the submethods are:

- True experiment— Experiments that maximize validity but are often difficult to perform under real-world conditions.
- Transfer experiment— Experiments that attempt to measure the effects of training in one situation (e.g., using a flight simulator) to performance in another (flying an aircraft).
- Pre- experiment— Compromised experiments that cannot promise the validity of true experiments.
- Test— Experiments that measure performance against a pre-determined standard.
- Quasi experiment— Compromised experiments that can nonetheless provide useful data if the evaluator can find suitable ways to compensate for their limitations.
- Ex post facto experiment— Studies that use historical data to mimic experiments.

## Judgment

The submethods of the Judgment method are based on respondent category; that is., the group whose judgments are gathered and considered:

- Analysts— Members of the evaluation community who are technically knowledgeable but not SMEs (subject-matter experts), such as civilian analysts and test managers.
- Subject-matter experts— Typically, very senior and knowledgeable members of the user community, such as master gunners and instructor pilots.
- Users— Typically, the class of individuals who's training is being evaluated, such as students, equipment operators, and crew members.

## Analysis

The submethods of the Analysis method are based on differences in the objectives of analysis:

- Evaluate— Evaluate a single way to train.
- Compare— Compare the relative effectiveness of two or more ways to train.
- Optimize— Refine the attributes of a training design to maximize its effectiveness.

Analyses use different strategies to assess training effectiveness. Some of these strategies are modeling, use of analogy, extrapolation, task list analysis, use of historical data, and mathematical models.

## Survey

Surveys vary in scale from small to large and in how data are collected (e.g., using questionnaire, interview, or observation.) They commonly use judgment data. Hence, Judgment and Survey methods overlap. Surveys are usually larger in scale than judgment-based evaluations.

## Selecting Methods

The degree of system development influences what type of evaluation method is most appropriate. For example, to use *experiment*, there must be something to conduct an experiment with. Ideally, this would be a complete system, but could be a mockup or simulation.

*Judgment* can be used in a limited way before a system exists (e.g., to estimate training potential of a hypothetical design or the perceived need for a system), but usually requires an existing, functional system.

*Analysis* can be performed without an existing, functional training system.

*Surveys* are usually conducted after a system has been developed, but they may be used during development or even before (e.g., survey potential users to determine the need for a new training program).

Despite the foregoing generalizations, there are numerous examples of all four of the evaluation methods being used throughout the various phases of system development.

## Further Reading

See the following chapters in Volume I for further discussion:

- Chapter 3 (Evaluation Methods) describes evaluation methods and provides examples of their application: Experiments (pp. 25-36), Judgment (pp. 36-40), Analytical (pp. 40-46), Survey (p. 47).
- Chapter 6 (Procedural Guidance) identifies and summarizes published evaluation guidance for the methods: Experiments (pp. 85-88), Judgment and Survey (pp. 91-93), Analytical (pp. 89-90).



## 6 EVALUATION CRITERIA

---

Evaluation criteria are the measures collected during an evaluation whose values are used to decide the outcome of the evaluation. Dependent variables in experimental research are one type of evaluation criteria.

### Reactions

The simplest variable to measure is reactions of participants to a training experience. This is done with a post-training questionnaire, interview, or videotaped group discussion such as an AAR (After-Action Review).

### Combat Performance

The operational testing community emphasizes the use of measures of *combat* performance such as engagement or battle outcomes. There are analogous variables for training systems. First, the evaluator could measure trainee performance during the simulation in relation to combat objectives. For example, did the simulated tank company defeat the simulated enemy or did the senior commanders participating in a war game win the war? Second, the evaluator might want to measure transfer of training from the simulation to the real world. One way would be to measure the impact of simulator training on performance in live simulation; for example, performance of Army units at the National Training Center (NTC).

### Student Learning

It is common to evaluate student performance in the schoolhouse based on test scores. Standardized tests can be used to evaluate training effectiveness. There are no collective tests in an LSTS. There *is*, however, collective performance. Improvement in collective performance demonstrates learning. Hence, LSTS could be evaluated based on collective learning.

## Collective Task Performance

Training is built upon tasks. Large-scale training simulations provide training on collective tasks. The Universal Joint Task List (UJTL) describes the tasks that are to be performed by a joint military force, the conditions under which the tasks are performed, and standards of performance. Comparable Service-specific task lists define the relevant collective tasks at the Service level. These task lists define what tasks the Services and Joint forces are expected to be able to perform. They are the logical tasks to use when building scenarios to evaluate LSTS. Performance on these tasks can be used as evaluation criteria.

Volume I systematically derives the set of evaluation criteria shown in Table 2. This scheme uses five different classes of variables.

Note that:

- The first three (Reaction, Collective Performance, Results) are obtained in the training system.
- The last two (Collective Performance, Results) are obtained post-training.

Think of Table 2 as a shopping list of criteria or measures to consider during evaluation.

The measures are not all of equal significance. Reaction data are useful, but less important than Collective Performance, which itself is less important than Results in the simulator. None of these is as important as performance in the real world, which means that the post-training measures are the most important of all.

**Table 2. Consolidated List of Recommended Dependent Measures with Descriptions**

<b>WHEN</b>	<b>DEPENDENT MEASURE</b>	<b>DESCRIPTION</b>
During training	1. Reaction	What were user and O/C reactions to simulator?
	2. Collective Performance	How well did teams and other collective echelons <i>perform</i> in the simulator?
	3. Results	What were the <i>tangible results</i> during training? (exchange ratio, percent losses by force, shots/kill, etc..)
Post-training	4. Collective Performance	How did team and other collective echelons perform after training?
	5. Results	What were the <i>tangible results</i> after training? (readiness, field exercise performance, combat outcomes)?

## Further Reading

See the following chapters in Volume I for further discussion:

- Chapter 2 (Building an Evaluation Framework) introduces evaluation criteria (pp. 21-22).
- Chapter 4 (Case Studies) describes the evaluation criteria used in the MDT2 (Multi-Service Distributed Training Testbed) evaluation (pp. 61-64).
- Chapter 6 (Procedural Guidance) identifies and summarizes published guidance on collective performance assessment (pp. 95-96).
- Chapter 7 (Evaluation Criteria) discusses evaluation criteria in detail (pp. 101-109).
- Chapter 8 (Evaluation Framework) summarizes several different evaluations and describes the evaluation criteria used in each (pp. 116-130).



## 7 C A S E S T U D I E S

---

### Overview

Case studies are concrete examples of how evaluations have been conducted in the past. They are a valuable resource in training evaluation. Cases provide insight into evaluators' decision-making, problem-solving strategies, evaluation methods, reporting, lessons learned, and general practices. They may show what was done well and poorly, what mistakes were made, and where the risks lie in the future. Cases provide vicarious experience that theory cannot. Good cases illustrate good evaluation practice. However, even flawed cases are useful if they help evaluators avoid future errors.

### Further Reading

See the following chapters in Volume I for further discussion:

- Chapter 3 (Evaluation Methods) describes evaluation methods at a procedural level, but illustrates discussion by summarizing studies using each method and telling where to locate study reports. Much of the chapter is based on actual evaluations contained in the TCEF (Training and Cost-Effectiveness File) data base.
- Chapter 4 (Case Studies) describes two well-documented evaluations of LSTS: SIMNET/CCTT (Simulation Networking/Close Combat Tactical Trainer) and MDT2.
- Chapter 5 (Evaluation Problem Areas) includes a section titled *Lessons Learned* that documents lessons learned in several past evaluations. It provides case-based recommendations for conducting future evaluations (pp. 77-80).
- Chapter 8 (Evaluation Framework) describes several cases of evaluations conducted to meet different types of evaluation objectives.



## 8 EVALUATION PROBLEM AREAS

---

### Overview

Most TEAs are field evaluations as opposed to laboratory evaluations. In laboratory evaluations, extraneous factors that may influence the outcome can usually be tightly controlled. Field evaluations are usually conducted with actual equipment, personnel, and under operational conditions— with all the uncertainties and messiness that implies. The typical field evaluation is a struggle to make the best of a less than ideal situation. A number of studies have critiqued field evaluation practice, with a particular focus on experiments whose outcomes were made questionable because of various compromises in research design. One critique found that most of the studies it examined contained one or more of the following limitations:

- Small sample size— Small samples result in low statistical power that makes it more difficult to detect true differences between groups. The differences may in fact be real, but statistical tests will not detect them.
- Unreliable performance measures— Unreliable performance measures do not provide consistent indications of performance and cannot be used to make comparisons between groups.
- Groups treated differently— If groups participating in an experiment are treated differently (other than for experimental/control treatments), the differential treatment may influence their performance, confounding with the experimental/control treatments.
- Device system errors— These errors may have a negative effect on subject performance.
- Subjects not random or matched— Subjects should be randomly assigned or matched prior to an experiment to assure that any differences found between them later can be attributed to the treatment and not to pre-existing differences.
- Ceiling effect— This generally occurs when the experimental task is too easy. If subjects perform at very high levels proficiency on a task, their scores may show little or no difference.
- Insufficient amounts of practice— Subjects who are not given sufficient time to practice with an unfamiliar device will still be learning when the experiment takes place and their performance will not reflect the true potential of the device.
- Floor effect— This generally occurs when the experimental task is too difficult. The inverse of the ceiling effect, if subjects perform at very low levels, differences may be undetectable.

When an evaluation suffers from these kinds of problems, its results become untrustworthy. Beware that all data are not equal. Interpreting such data has the same value as reading tea leaves. Making decisions based on such data is irresponsible.

### Further Reading

See the following chapter in Volume I for discussion:

- Chapter 5 (Evaluation Problem Areas) contrasts laboratory and field evaluations, discusses lessons learned from past evaluations, and critiques field evaluation practice (pp.73-84).

## 9 EVALUATION FRAMEWORK

---

The following discussion describes the framework and tells how it may be used.

### Pieces of the Puzzle

An evaluation framework is a set of evaluation *principles* and a description of evaluation *events, objectives, timing, and criteria*. It is intended to help the evaluator select the most suitable evaluation *methods* based on the circumstances, provide procedural descriptions of the methods, and identify case studies that may apply as models to emulate.

Evaluation principles, objectives, methods, criteria, and the use of case studies were discussed earlier.

An evaluation *event* may be thought of as a single evaluation of some aspect of an LSTS. A complete evaluation may include several different evaluation events, staged across time, based on *stage of system development*. (*Developmental phases* and *milestone decision points* were discussed earlier.)

### Putting Together the Pieces

#### 1. Determine Stage of System Development

A complete evaluation may include several different evaluation events, staged across time, based on *stage of system development*. Determine stage of system development in accordance with the phases and categories described earlier in this manual. These stages and phases are:

- pre-development (Phase 0)
- developmental (Phases I & II)
- post-development (Phase III)

Stage of system development will influence evaluation objectives.

## 2. Define Evaluation Objectives

Define evaluation objectives based on the categories in Table 1:

- Estimate need for new training system (1)
- Predict, Measure, or Determine training effectiveness (2A, 2B, 2C)
- Predict, Measure, or Determine transfer of training (3A, 3B, 3C)
- Predict, Measure, or Determine user acceptance (4A, 4B, 4C)
- Support training design (5A,B) or Determine training status (5C)
- Support system design (6A,B) or Evaluate system design (6C)

## 3. Identify Possible Case Studies

Chapter 8 in Volume I provides a separate table for each objective that (1) identifies relevant case studies, (2) lists evaluation criteria, (3) lists evaluation methods, and (4) summarizes the studies. These tables allow the reader to map from evaluation objectives to these four factors.

Relevant tables for each evaluation objective are as follows:

- Objective 1: Estimate need for new training system— see examples on page 116
- Objective 2A: Predict Training Effectiveness— Table 8-2
- Objective 2B: Measure Training Effectiveness— Tables 8-3 and 8-4
- Objective 2C: Determine Training Effectiveness— Table 8-5
- Objective 3B: Measure Transfer of Training— Tables 8-6 and 8-7
- Objective 4B (Measure User Acceptance)— Table 8-8
- Objective 4C (Determine User Acceptance)— Table 8-8
- Objective 5(AB): Support Training Design— Table 8-9
- Objective 5C: Determine Training Status— Table 8-10

Select the appropriate table for the evaluation objective and use it to identify possible case studies.

#### 4. Review Case Studies

Review the possible case studies to see if any could be used as a model for the evaluation being conducted. Review them in terms of these properties:

- What was evaluated
- Evaluation criteria
- Evaluation methods used
- How the evaluation was designed and conducted

Select case studies based on the best match of these four factors with the new evaluation.

If no adequate case studies are found, proceed without one.

#### 5. Determine Evaluation Criteria

If case studies have been identified, consider the evaluation criteria they used. Determine whether the same, similar, or analogous criteria may be suitable for the new evaluation.

Consult with evaluation stakeholders to determine what evaluation criteria will satisfy their evaluation requirements.

Independently of case studies and inputs from stakeholders, compile a set of evaluation criteria that will convincingly establish the training effectiveness of the LSTS being evaluated. Table 2 presents a shopping list of criteria or measures to consider during evaluation.

Consider the evaluation criteria from these three sources and select the criteria to use.

#### 6. Determine Evaluation Methods

Consider the evaluation methods used in the case studies identified. Determine whether the same methods may be suitable for the new evaluation.

Consult with evaluation stakeholders to determine what evaluation methods will satisfy their evaluation requirements.

Independently of case studies and inputs from stakeholders, identify methods that will provide valid and reliable training effectiveness data.

Making this selection is a matter of judgment. Often the choice of methods is constrained by available resources.

## 7. Design and Conduct Evaluation Event

There is no cookbook recipe to design and conduct an evaluation event. On the personnel side, a few general principles apply:

- Consult with training evaluation experts before the evaluation
- Let evaluation experts design and conduct the event
- Let stakeholders oversee the event

On the technical side, these principles apply:

- If possible, model the evaluation on a similar case study
- Review published *lessons learned* in similar evaluations
- Assure that common problem areas do not compromise the evaluation

## Further Reading

See the following chapter in Volume I for further discussion:

- Chapter 8 (Evaluation Framework) (pp. 111-130) presents the evaluation framework in terms of evaluation objectives and principles. It also describes evaluation events and links them to relevant examples and procedural guidance.