IDA

INSTITUTE FOR DEFENSE ANALYSES

On Measuring the Effectiveness of Large-Scale Training Simulations

John E. Morrison
Colin Hammon

October 2000

IDA Paper P-3570

Log: H 00-002743

This work was conducted under contract DASW01 98 C 0067, Task BE-2-1709, for ODUSD(R) R&T, PP. The publication of this IDA document does not indicate endorsement by the Department of Defense, nor should the contents be construed as reflecting the official position of that Agency.

INSTITUTE FOR DEFENSE ANALYSES

IDA Paper P-3570

On Measuring the Effectiveness of Large-Scale Training Simulations

John E. Morrison
Colin Hammon

PREFACE

This document was prepared for the Deputy Under Secretary of Defense for Readiness (Readiness and Training Policy and Programs) under an Institute for Defense Analyses (IDA) task order entitled "Concepts of Training Effectiveness for Large-Scale Simulations." Technical cognizance of this task is assigned to Mr. Daniel Gardner.

Dr. Henry K. Simpson of the Defense Manpower Data Center (DMDC), Dr. Jozsef A. Toth of IDA, and Dr. J. Dexter Fletcher of IDA reviewed this paper. We acknowledge and appreciate their assistance in revising the report.

We also acknowledge the contribution of the late Dr. Jesse Orlansky of IDA. His pioneering work in the measurement of simulator training effectiveness provided the foundation and inspiration for this paper.

CONTENTS

SUMMARY S-1

I. INTRODUCTION I-1

A. Problem I-1

B. LSTSs and Training Effectiveness I-1

1. Limited Relevance of Transfer Paradigms I-2
2. Assessment of Functions Other Than Training I-2
3. Scope and Diversity of Training Objectives I-2

C. Training Effectiveness Measurement Issues I-4

1. Performance Measurement Issues I-4
2. Research Design Issues I-4

D. Organization I-5

II. MILITARY GUIDANCE II-1

A. Military Services II-1

1. Army II-1
2. Navy and Marines II-2
3. Air Force II-2

B. Joint Services II-3

C. Summary II-3

III. PERFORMANCE MEASUREMENT III-1

A. Fundamental Measurement Concepts III-1

1. Validity III-1
2. Reliability III-3
3. Sensitivity and Specificity III-5
4. Utility III-6

B. Types of Performance Measures III-8

1. Objective Performance-Based Measures III-8
2. Judgment-Based Measures III-11

IV. RESEARCH DESIGN IV-1

A. Training Surveys IV-3

1. Devising a Sampling Strategy IV-3
2. Choosing an Appropriate Survey Mode IV-4
3. Constructing the Instrument IV-4
4. Analyzing the Data IV-6

B. Empirical Performance-Based Research Designs IV-7

1. Randomized (True) Experiments IV-8
2. Quasi-Experimental Designs IV-17
3. Correlational Studies IV-18
4. Comparison of Design Approaches IV-21

C. Training Analyses IV-22

1. State of the Art IV-23
2. Examples of Formal Models IV-25

V. RECOMMENDATIONS V-1

A. Identify Specific Measurement Issues V-1

1. Determine the Types of Decisions That the Measure Is Intended to Inform V-1
2. Identify the Audience for Whom the Measure Is Intended V-2
3. Specify Whether Absolute or Relative Effectiveness Measures Are Required V-2

B. Devise a Measurement Plan That Specifies Performance Measurement and Research Design V-2

C. Use Valid and Reliable Performance Measures V-2

D. Obtain the Most Valid Measure of Proficiency Possible Within the Constraints of the Evaluation V-3

E. To the Extent Possible and Reasonable, Impose Experimental Control on the Research Situation V-4

F. Use Analytic Models in the Early Stages and Continue Using These Models Throughout System Development V-5

G. Include User Reactions as an Adjunct to Performance Data and Analysis V-6

References Ref-1

Glossary GL-1

FIGURES

IV-1. Example of a 5-Point Likert Scale IV-6

IV-2. Median Target Prosecution Times From a Joint Experiment on Attack Operations IV-12

IV-3. Isoperformance Curve for Total Training Costs IV-15

TABLES

III-1. Outcomes From a Diagnostic Test III-6

III-2. Comparison of Outcome and Process Measures III-9

IV-1. Outline of Pretest-Posttest Control Group Design IV-9

IV-2. Outline of Groups X Trials Transfer Design IV-16

IV-3. Comparison of Empirical Research Design Approaches IV-22

IV-4. Calculated Utilities for Two Simulation Models IV-27

IV-5. Example Results of a Comparison-Based Prediction of Training Effectiveness IV-31

SUMMARY

A. PURPOSE

This paper identifies and evaluates methods for measuring the benefit or utility of large-scale training simulations (LSTSs). Such methods are usually discussed under the rubric of training effectiveness analyses (TEAs). Although TEAs can be concerned with the cost and the effectiveness of training, this paper focuses primarily on the latter (i.e., the benefits of training). The purpose is to provide a discussion of relevant issues for those who are directly responsible for assessing the training effectiveness of LSTSs.

B. METHOD

We examine literature related to the measurement of LSTS training effectiveness, including guidance provided by the individual military Services and the Office of the Secretary of Defense (OSD), the body of research on performance measurement, and research practices related to experimental design and analysis. We then synthesize the findings from these literature sources to provide recommendations for performing TEAs of LSTSs.

C. FINDINGS

1. Military Guidance

The Army provides the most comprehensive guidance on conducting TEAs. However, although Army publications specify *when and where* TEAs should be performed, these publications are less precise about *how* TEAs should be performed. Existing joint publications are ambiguous in setting the minimum requirements for a TEA, but some of the new draft publications may mitigate this problem.

2. Performance Measurement

In the context of training effectiveness, performance measurement refers to an assessment of learning outcomes. This broad definition includes not only the measurement of task performance (i.e., skills), but also the evaluation of task-related knowledge and attitudes when these are valid outcomes of LSTS training. At the same time, the definition

of performance does not pertain to the engineering performance of the system (i.e., how often a training system crashes), because such metrics are generally not relevant to LSTS training objectives.

Performance measures can be categorized into two general types: those based on actual performance and those derived from expert judgment. All things being equal, measures based on actual performance are the preferred data source for TEAs. However, objective, performance-based measures may not be available or even appropriate for certain effectiveness issues, and researchers have often turned to measures based on the judgment of subject matter experts (SMEs).

Researchers use several fundamental measurement concepts to evaluate the adequacy of any measure or set of measures. These concepts apply equally well to judgment-based and performance-based measures and can be cast into four broad areas of concern:

1. **Validity.** A performance measure is valid to the extent that it measures what it is purported to measure. Accordingly, the index of validity is based on the relationship between the performance measure and the concept of interest. Different methods—each having somewhat different implications for training effectiveness—exist for assessing validity. The text discusses validation methods based on the content of measures, the relationship to other variables, and the relationship to criterion variables.
2. **Reliability.** Reliability is the consistency or stability of measurement. Reliability of measurement is a fundamental requirement for assessing training effectiveness because a measure cannot be valid if it is not reliable. As with measuring validity, researchers have used several different types of measurement operations (e.g., procedural, test-retest, inter-item, and inter-observer reliability) to calculate reliability indexes. This paper discusses these definitions of reliability.
3. **Sensitivity and specificity.** The concepts of sensitivity and specificity, which evolved from medical testing traditions, refer to the classification accuracy of tests that determine whether patients have a particular disease. In the context of performance measurement, a sensitive test correctly discriminates among levels of the measurement in question. Insensitive measures, on the other hand, can mask real differences in training effectiveness. A specific test correctly excludes individuals who do not have the characteristic in question. A nonspecific test, on the other hand, taps into individual attributes other than the one in question.
4. **Utility.** Performance measures might be reliable and valid but still be of limited utility for measuring training effectiveness. When considering the utility or

value of measuring a particular aspect of performance, the associated costs must be considered.

3. Research Designs

Training effectiveness designs are often described in terms of Kirkpatrick's (1976) four types or "levels" of training effectiveness evaluations. These levels are usually described as a sequence of four progressively difficult and time-consuming steps or procedures that yield increasingly comprehensive and convincing information about a training program or system. Although the higher-level designs (Levels III and IV) provide more information that is relevant to LSTSs, these levels are also the most difficult to implement and interpret.

Kirkpatrick's schema classifies training effectiveness evaluations by abstract evaluation functions or questions. Research designs, however, are more concrete because they lay out specific procedures used to collect performance measures.[1] Accordingly, we offer a taxonomy that classifies TEAs by the research operations used or the types of data collected. The resulting three categories can, nevertheless, be mapped back into Kirkpatrick's schema to discuss evaluation functions and goals. The three categories are:

- **Training surveys.** Training surveys are the standard approach for obtaining user reactions to training devices or simulations. "Users" include the training audience (i.e., the trainees) and the relevant unit commanders and trainers. These reactions may document the users' general impressions of the system, or they may provide specific feedback about the training system and its relation to training objectives. However, even though such data are easy to obtain, they provide the weakest evidence for training effectiveness.
- **Empirical performance-based research.** Performance-based research represents the most highly respected approach for measuring training effectiveness because it examines the effects of training on task execution in realistic contexts.
- **Training analyses.** In several situations, neither survey nor performance-based designs are feasible or applicable as methods for determining LSTS training effectiveness. In such situations, the only viable alternative is to use a training analysis model, which is a method or procedure for estimating or forecasting the effectiveness of training systems. This approach is different from

---

[1] The present use of the term "design" is broader than the standard academic approach. The academic usage implies that the data are collected in a controlled experiment or study. This implication does not apply to the use of the term in this paper.

the previous approaches (training surveys and empirical performance-based research) because of its reliance on nonempirical data (i.e., data other than that derived from actual performance or user input).

D. RECOMMENDATIONS

We recommend that analysts take the following actions when measuring the effectiveness of LSTSs:

- **Identify specific measurement issues.** The analyst must decide among a variety of approaches for measuring the training effectiveness of an LSTS. The first step is to determine and document the exact issues that the measure should address.
- **Devise a measurement plan that specifies performance measurement and research design.** The analyst should devise a plan for measuring training effectiveness. This plan must specify the performance measures and the design for collecting these measures.
- **Use valid and reliable performance measures.** To the extent possible, analysts should employ measures that have known validity and reliability. If this is not possible, analysts should provide the opportunity to measure reliability and validity within the TEA, particularly those measurement characteristics that they suspect are problematic.
- **Obtain the most valid measure of proficiency possible within the constraints of the evaluation.** In most situations, the most valid measure of job proficiency is on-the-job performance in which weapons systems and related equipment are used as these systems/equipment would actually be used in combat. If this is not possible, the next best alternative is a surrogate measure (e.g., a virtual simulation). If performance data are not available from primary or surrogate measures, analysts should seek the best input from an SME (using some systematic model for estimating training effectiveness).
- **To the extent possible and reasonable, impose experimental control on the research situation.** In general, analysts should always opt for the approach that allows them the greatest degree of control over the experimental situation. Thus, a randomized experiment is generally preferred to a quasi-experiment, and a quasi-experiment is generally preferred to a correlational study.
- **Use analytic models in the early stages and continue using these models throughout system development.** The cost and effort involved in initializing and executing the model should decrease progressively during system development. The model should also be updated and revised

periodically to reflect any performance-based results. The point of the updates is not to confirm what analysts already know. The idea is to perform excursions from the known cases and to test conditions and circumstances that are not known.

- **Include user reactions as an adjunct to performance data and analysis.** Although user reactions are easy to obtain, these data provide the weakest evidence for training effectiveness. Such data should not be used as sole or even primary data for the TEA of an LSTS. Rather, these judgment-based measures should be used to supplement performance- or analytic-based data.

I. INTRODUCTION

A. PROBLEM

Modern simulation-based training methods encompass several technologies, including live (e.g., systems appended to actual equipment), virtual (e.g., flight simulators), and constructive (e.g., computer-based war games) simulation. Large-scale training simulation (LSTS) technology provides the capability to link a large number of entities generated by disparate systems and place these entities into a common, simulated battlespace. As such, LSTS technology promises to provide a unique and potentially powerful approach for training joint and combined forces. However, to realize this potential, the Department of Defense (DoD) and the military Services must invest substantial amounts of time and money. To ensure that LSTS systems are appropriate investments of those resources, the DoD Office of the Inspector General (OIG) recommended that DoD "… establish policy and procedures for evaluating the training effectiveness and cost-effectiveness of large scale training simulations" (Office of the Inspector General, DoD, 1997, p. 36).

This paper identifies and evaluates methods for measuring the benefit or utility of LSTSs. Such methods are usually discussed under the rubric of training effectiveness analyses (TEAs). Although TEAs can be concerned with the cost as well as the effectiveness of training, this paper focuses primarily on the latter (i.e., the benefits of training). The purpose is to provide a discussion of relevant issues for those who are directly responsible for assessing the training effectiveness of LSTSs.

B. LSTSs AND TRAINING EFFECTIVENESS

In the context of simulation-based training, training effectiveness can be defined as the extent to which simulations prepare individuals or collections of individuals to conduct military operations. The traditional concepts of training effectiveness were derived from the analyses of older small-scale systems. Although many of these concepts apply equally well to the evaluations of LSTSs, the evaluator must take into consideration some of the distinguishing features of LSTS. Some of the more relevant differences are discussed below.

1. Limited Relevance of Transfer Paradigms

The traditional benchmark for the effectiveness of small-scale simulation systems has been performance on the actual equipment (i.e., the extent to which skills trained on the simulator transfer to performance on the military equipment). LSTS systems, because of their scale and modular nature, may not be associated with specific sets of equipment. In recognition of this fact, the DoD Inspector General (IG) suggested using collective field training exercises as the appropriate benchmark for LSTS effectiveness. In response to this suggestion, the Under Secretary of Defense for Acquisition and Technology (USD(A&T)) correctly pointed out that LSTS training should be viewed as augmenting rather than as substituting for field training. This official also pointed out that field training may not be the appropriate venue for training certain collective tasks for several reasons, including cost, security, safety, environmental restrictions, and political constraints (Office of the Inspector General, DoD, 1997). Moreover, researchers (e.g., Fletcher and Chatelier, 2000; Hiller, 1987) have also argued that field training exercises are also simulations and, as such, do not replicate many crucial the conditions and contingencies that exist in actual combat. Thus, appropriate alternative benchmarks for LSTS systems are needed to conduct valid cost-effectiveness analyses.

2. Assessment of Functions Other Than Training

Traditionally, the effectiveness of a small-scale simulation system has been defined narrowly by its function to train specific tasks. In comparison, an LSTS system provides a wider array of functions related to training and readiness assessment, including mission rehearsal and individual and collective performance assessment. Each of these functions has different implications for assessing effectiveness. Starting with the basic training functions, as an example, an effective LSTS would have a large library of canned scenarios for exercising certain core skills. Mission-rehearsal functions would require the capability to develop scenarios tailored to particular missions. Assessing the individual and collective performances that underlie such complex combat scenarios requires the capability to collect and archive performance data automatically. Thus, to address the total value of an LSTS system, the concept of effectiveness needs to be broadened to assess such capabilities.

3. Scope and Diversity of Training Objectives

The size and complexity of the LSTS distinguish it from other simulation-based training systems. The implication for the measurement of training effectiveness is that the

objectives of LSTSs are correspondingly large in scope and diverse in content. The following two examples illustrate the impact of this increased scope and diversity.

a. Simulation Networking (SIMNET)

SIMNET, which was initiated as a joint effort of the Defense Advanced Projects Research Agency (DARPA) and the Army in 1983 to train battalion-size forces in tactical warfare, exemplifies the effect of LSTS complexity on evaluation objectives. SIMNET represents the prototypical LSTS in many ways. It has tank and aircraft simulators, communications networks, command posts, and extensive data processing capabilities. Initial production of the Close Combat Tactical Trainer (CCTT) was begun in 1991 to provide a follow-on to SIMNET (Alluisi, 1991).

The evaluation of such a complex system requires several approaches. Simpson (1999) cited 26 different evaluations of SIMNET/CCTT, of which 11 were based on task performance data, 8 were founded on analyses of system capabilities, 6 were based on expert judgment, and 1 was derived from user surveys.

b. Multi-Service Distributed Training Testbed (MDT2)

MDT2 linked eight different types of simulators from different Services to provide training in the Close Air Support (CAS) mission. Army, Marine, and Air Force personnel participated at three locations. The MDT2 evaluation (Orlansky, Taylor, Levine, and Honig, 1997), conducted during two 5-day periods in 1994 and 1995, points out the need to collect a diverse set of process and outcome performance measures.

Among the many process measures collected were the observer/controller (O/C) ratings of a trainee's performance in tactics, techniques, and procedures. In addition, at the end of each day, the Senior Trainer judged the relative combat proficiency of each trainee in integrating CAS into the planning process, controlling aircraft control in the area, executing the plan, and reacting unforeseen problems on the battlefield. Likewise, several outcome measures were collected to show performance trends related to the amount of training.

Among the many outcome measures collected were the number, timing, and frequency of bombs released; the number of enemy vehicles hit, damaged, or destroyed; the number and percent of bomb releases that resulted in vehicle impact or proximal impact; and timing a volume of CAS fires in relation to artillery fires.

C. TRAINING EFFECTIVENESS MEASUREMENT ISSUES

We assume that two conceptually separate, but nevertheless related, issues underlie the measurement of LSTS effectiveness (and any other training device, simulation, or simulator):

1. Performance measurement
2. Research design.

At the outset, we acknowledge that these two topics are related. For instance, the negative effects of variability in performance measures can be mitigated partially by using repeated measures designs. Nevertheless, we consider them separately because they have evolved from two separate and distinct research traditions. Performance measurement issues have evolved from research and development (R&D) on individual differences, whereas research design issues have developed from the fields of learning and transfer of training. Consequently, the concepts and constructs for these two topic areas are quite different, and separating them facilitates their exposition.

1. Performance Measurement Issues

Performance measurement issues relate to a fundamental question: What aspect of effectiveness should be measured? The usual metric for determining the effectiveness of a training device is to assess task performance, including measures of behavioral processes and the outcomes of those tasks. In some cases, task performance cannot be assessed directly, and analysts must rely on more subjective estimates of the training devices' capabilities to train specific tasks. Also, training devices can be assessed for more general attributes, including the impressions of trainers and trainees who use the devices. Clearly, the validity of training effectiveness for a large-scale system ultimately depends on the validity and reliability of the TEA performance measures.

2. Research Design Issues

As with performance measurement issues, we can reduce research design issues to a single fundamental question: How should training effectiveness measures be collected and summarized? Design issues are important because they determine the extent to which data from performance measures can be attributed to the use of training devices. For instance, experimental designs are intended to establish direct cause-and-effect relationships between the device and performance. Design issues also relate to the mathematical formulae and statistical techniques used to summarize the results. Since research design and analysis

issues are often intimately intertwined, analysts must consider these issues together when measuring training effectiveness.

D. ORGANIZATION

In this paper, we examine the problem of measuring the training effectiveness of LSTSs from several points of view. In Section II, we review and evaluate the guidance provided by the individual military Services and the Office of the Secretary of Defense (OSD). In Sections III and IV, we discuss the considerable body of research that relates to planning and executing TEAs—the first relating to performance measurement and the second relating to research design. In Section V, we provide some recommendations that we synthesized from this information.

II. MILITARY GUIDANCE

Simpson (1995) reviewed military guidance on conducting TEAs. This guidance covered the period from 1980 to 1995. The following summarizes and updates Simpson's report. It includes a discussion of military organizations designated to perform TEAs and identifies and describes published guidance that the Services provide.

A. MILITARY SERVICES

1. Army

The TRADOC Analysis Center at White Sands Missile Range (TRAC–WSMR) is the Army's lead agency for providing technical assistance and conducting TEAs. Training and Doctrine Command (TRADOC) Regulation 350-32, *The TRADOC Training Effectiveness Analysis (TEA) System*, provides official guidance for designing and conducting TEAs (U.S. Army Training and Doctrine Command, 1994). This document recognizes three types of TEAs:

1. **TEAs related to system acquisition.** These TEAs investigate the training effectiveness of new systems and are designed to coincide with system-acquisition decisions and milestones.
2. **TEAs for evaluating current training programs.** These TEAs study the effectiveness of existing training programs and investigate alternative training approaches and technologies.
3. **TEAs for improving training study methods.** These TEAs are methodological studies designed to improve the overall TEA program.

This paper focuses on the first two types of TEAs and is itself is an example of the third type.

Although TRADOC Regulation 350-32 sets guidelines for conducting TEAs, it also recognizes that no single "best" method exists for measuring training effectiveness. The types of questions that the TEA designers intend to ask should determine the method used. In addition to standard quantitative methods for evaluations, Army TEAs encompass other diverse methods, such as qualitative analyses, field observation, task analyses, survey research, and questionnaire design and analyses.

2. Navy and Marines

No single organization within the Department of the Navy is responsible for TEAs. Navy systems commands (SYSCOMs) develop system hardware and software and may identify the need to perform a TEA as part of the acquisition process. The Navy's research laboratories, such as the Naval Air Warfare Center Training Systems Division (NAWCTSD), may be asked to provide analytic support, or these organizations may initiate validation or cost-avoidance studies. The Department of the Navy Modeling and Simulation Management Office (NAVMSO) has oversight responsibility of these analyses but does not initiate such studies.

Type Commanders are responsible for decisions regarding the tradeoffs between the use of simulation and actual forces. They are, therefore, responsible for evaluating simulations as training devices. Fleet Commanders have ultimate responsibility for TEAs of simulations, including large-scale exercises. The Chief of Naval Education and Training (CNET) may also perceive a need for a TEA and direct that it be conducted.

Although this system may appear to be rather ad hoc, in effect, LSTS evaluations are conducted as part of the overall acquisition and training system. The Secretary of the Navy has directed NAVMSO to write a common verification, validation, and accreditation (VV&A) instruction. We do not know whether this instruction will include a single responsible agent for LSTS TEAs.

3. Air Force

The Air Force has no published guidelines for conducting TEAs, and no single Air Force agency or office is tasked to perform training effectiveness evaluations (Simpson, 1995). A variety of agencies perform evaluations on an "as-needed" basis. These agencies are tasked primarily to evaluate the new training technologies and the methods that result from the development of these devices. Prominent among these agencies is Human Effectiveness Directorate (HE), Warfighter Training Research Division (HEA) of the Air Force Research Laboratory (AFRL). Other evaluators include the Flight Training System Program Office at the Aeronautical Systems Center at Wright-Patterson Air Force Base (AFB), Ohio, and the Air Force Major Commands.[2] In addition, the Air Force's 29th Training

---

[2] Air Education and Training Command (AETC), Air Mobility Command (AMC), Air Combat Command (ACC), Air Force Materiel Command (AFMC), Pacific Air Forces (PACAF), Air Force Space Command (AFSPC), U.S. Air Forces in Europe (USAFE), Air Force Special Operations Command (AFSOC), and the Air Force Reserve Command (AFRC).

Systems Squadron, located at Eglin AFB, Florida, is charged with performing the certifications of all Combat Air Forces (CAF) training systems. However, these certifications are not, strictly speaking, equivalent to formal TEAs (D. H. Andrews, personal communication, 29 February 2000).

B. JOINT SERVICES

Draft DoD Directive (DoDD) 1430.13, *Training Simulators and Devices* (Department of Defense, 1996), establishes DoD policy for cost and training effectiveness analyses (CTEAs). This directive applies to any simulator or training device that meets the criteria for a major automated information system (MAIS) acquisition program or to any special-interest program so designated by the Secretary of Defense. A simulator or training device is classified as an MAIS if either of the following applies:

- It is estimated to require program costs in any single year in excess of $30 million in fiscal year (FY) 1996 constant dollars, total program costs in excess of $120 million in FY 1996 constant dollars, or total life-cycle costs in excess of $360 million in FY 1996 constant dollars
- It is designated as an MAIS by the Assistant Secretary of Defense for Command, Control, Communications, and Intelligence (ASD(C3I)).

For the simulators and devices within the directive's purview, the directive stipulates that TEAs should be conducted at all major milestones. Furthermore, costs should be considered (i.e., CTEAs should be conducted) at the model's earliest phases (0 and I) so that the impacts of these costs can support the decision-making process when system milestones are developed.

The present version of DoDD 1430.13 (dated 22 August 1986) provides little information on how CTEAs or TEAs should be performed. However, a recently revised draft of DoD Instruction (DoDI) 1430.1 (October 1999) seeks to rectify this shortcoming by providing methodological guidance in an annex. This annex is based on Simpson's (1999) extensive review of the state-of-the-art in conducting CTEAs and TEAs and stipulates the methods that should be used to conduct TEAs and the measures used therein. However, it has not been adopted officially.

C. SUMMARY

In 1995, Simpson concluded that the Army provides the most comprehensive guidance on conducting TEAs. However, the Army's approach has not been updated since the

September 1994 publication of TRADOC Regulation 350-32. Furthermore, whereas the Army publication is clear about *when and where* TEAs should be performed, it is less precise about *how* those analyses should be performed.

For Joint Service guidance, Simpson (1995) also concluded that DoDD 1430.13 (dated 22 August 1986) was ambiguous in setting the minimum requirements for a TEA. However, a recently revised draft of DoD Instruction (DoDI) 1430.1 (October 1999) goes a long way toward mitigating this long-standing problem.

III. PERFORMANCE MEASUREMENT

In contrast to the relative dearth of concrete guidance from military publications, a considerable number of research and analytic practices related to performance measurement have evolved over the last 50 to 60 years. This section discusses some of the more important research practices that pertain to performance measurement as it relates to training effectiveness.

In the context of training effectiveness, performance measurement refers to an assessment of learning outcomes. These outcomes are defined broadly to include not only the measurement of task performance (i.e., skills), but also the evaluation of task-related knowledge and attitudes when these are valid outcomes of LSTS training. At the same time, the definition of performance does not pertain to the engineering performance of the system (i.e., how often a training system crashes), because such metrics are generally not relevant to LSTS training objectives.

A. FUNDAMENTAL MEASUREMENT CONCEPTS

Several fundamental measurement concepts are used to evaluate the adequacy of any measure or set of measures. These concepts apply equally well to performance-based and judgment-based measures and can be cast into four broad areas of concern: validity, reliability, sensitivity and specificity, and utility.

1. Validity

A performance measure is valid to the extent that it measures what it is purported to measure. Accordingly, the index of validity is based on the relationship between the performance measure and the concept of interest. For example, the performance measure in question may be the score on a test that gauges a person's knowledge of the functions and components of operations orders, whereas the central concept is proficiency at producing an actual order. As described below, different methods exist for estimating the relationship between the measure and the corresponding concept—each of which has somewhat different implications for training effectiveness.

a. Methods Based on the Content of Measures

This approach seeks to determine the degree to which the measure provides a representative sample of elements in the performance domain. The formal assessment of content validity is usually assigned to subject matter experts (SMEs). One important aspect of content validity is called "face" validity, which refers to the surface features or characteristics of measures. This form of validity is used primarily to determine the initial acceptability of measures. One common approach to content validation is to have SMEs review the list of tasks that a simulator is intended to train. For example, SMEs reviewed the face validity of a list of training objectives for MDT2 (Orlansky et al., 1997). Based on SME comments, this list was modified and then used to support development of training scenarios and performance measures.

b. Methods Based on Relationships to Other Variables

This method concerns the extent to which the measure reflects the underlying theoretical construct. The correspondence is tested by the relationships between the measure in question and measures of other known constructs. Although no concrete example of this method emerges from the short history of LSTS research, one can readily be imagined. Suppose, for example, that an LSTS trains a large number of performers at distant sites and that sending evaluators to all sites to observe each performer would be impractical. An alternative might be to have performers evaluate each other using a systematic peer-rating scheme. The validity of the peer measure could then be validated by observing a small sample of performers and deriving traditional and demonstrably valid performance measures. The peer ratings for the entire group could then be validated by demonstrating that these ratings are correlated with traditional performance measures for the sample of performers.

c. Methods Based on Relationships to Criterion Variables

This approach quantifies the relationship between the measure in question and some criterion measure that provides the standard for the concept in question. As an example, Waag, Raspotnik, and Leeds (1992) examined the validity of performance measures in the Simulator for Air-to-Air Combat (SAAC), which networks two F-16 and two F-15 cockpits in a virtual simulation of air combat engagements. They examined two types of performance measures: aircraft state measures (e.g., altitude and airspeed) and positional advantage measures that calculate the vulnerability (probability of kill) of Blue and Red aircraft from their relative positions. They measured the relationships between these two

measures and an unambiguous criterion measure (the outcome of each air combat engagement categorized as either a win, lose, or draw). Their results, based on the performance data of pilots executing standardized engagements in the SAAC, indicated that the positional advantage measure was more strongly related to the outcome of each air combat engagement than the aircraft state measure was. However, the analyses also suggested that a composite of both types of measures was more accurate at predicting the outcome of each air combat engagement than positional advantage was by itself. Thus, developing a composite measure of performance for air-to-air combat that can be used to supplement, or be used in lieu of, outcome measures in TEAs is feasible.

The approach used by Waag, Raspotnik, and Leeds (1992) was to measure the *concurrent* validity of the performance measures. That is, the two performance measures and the criterion were obtained from the same sample at approximately the same time. This approach is often used to determine the ability to substitute one measure for another. An alternative approach is to measure the *predictive* validity of a measure by determining the degree to which it predicts performance on the criterion at some future point. For instance, an analyst would use the latter method to determine whether simulation-based performance predicts potential problems in a field environment.

2. Reliability

Reliability refers to the consistency or stability of measurement. It is a basic requirement for any performance measure because a measure cannot be valid if it is not reliable. Thus, reliability of measurement is a fundamental requirement for assessing training effectiveness (Boldovici, 1987). As with measuring validity, evaluators have used several different types of measurement operations to calculate reliability indexes.

a. Procedural Reliability

Procedural reliability refers to the consistency with which the test instrument is administered. To be useful, a measure must be collected in a procedurally reliable fashion, and the training evaluator must ensure that measurement operations are performed in a consistent manner. An analyst who administers any test instrument in an inconsistent manner injects an additional and unwarranted source of unreliability in the measures. In contrast to the other sources of reliability, procedural reliability is not usually measured directly; rather, the concept of procedural reliability is used to emphasize the need for standardized administration of performance measures. In other words, the results of any TEA cannot be interpreted properly if the performance measurement procedures are not conducted reliably.

This admonition applies to the evaluation of any training device or program, but it is particularly relevant to LSTS evaluations that may require assessment on a variety of performance measures by multiple observers.

b. Test-Retest Reliability

Test-retest reliability relates to the stability of a performance over time. Test-retest reliability is measured by the empirical correlation between multiple performance administrations within the same group of performers. A general finding in regard to this measure is that human performance is typically not stable in the early stages of learning but increases as a function of practice. Thus, repeated testing has been shown to be an effective approach for stabilizing human test performance and thereby increasing reliability of measurement (e.g., Jones, Kennedy, and Bittner, 1981).

In the context of TEAs, Boldovici (1987) argued that repeated testing partially mitigates problems inherent in evaluations based on small sample sizes. At the same time, he pointed out that repeated testing does not eliminate the problem of the reliability of measurement. However, the process of obtaining multiple measures of response from performers permits the evaluator to calculate the index of reliability. The advantages of repeated measurement are particularly relevant to the evaluations of LSTSs for two reasons:

1. Large samples of LSTS trainees are often impossible or impractical to obtain.
2. The reliability of relevant performance measures for LSTSs, particularly those under development, is unknown.

c. Inter-Item Reliability

Inter-item reliability pertains to the internal consistency of a multiple-item test. Several different methods have been developed to measure this concept. In the split-half approach, for example, test instruments are randomly divided into two parts, and the parts are correlated with each other. Inter-item reliability is a desirable characteristic only when the inventory is designed to measure a single characteristic or performance dimension (e.g., response speed or dexterity). In contrast, many performance inventories are multidimensional in character, and items should not necessarily intercorrelate. For instance, Smith and Hagman (1998) examined the inter-item correlation among tasks in a live-fire, tank gunnery exercise and the part-whole correlation between these tasks and the overall score on exercise. Their results revealed relatively small inter-item correlations but relatively large part-whole correlations, indicating that the performance on each task is associated with a unique source of variance in the overall score. This is the sort of finding that an evaluator can

expect from many types of performance tests, including those related to the evaluation of an LSTS.

d. Inter-Rater Reliability

Inter-rater reliability refers to the degree to which observers agree on their performance ratings. The index of inter-rater reliability is the correlation among two or more judges' independent performance ratings of the same individuals or groups under the same conditions. Some aspects of LSTS performance measurement can be automated and are not subject to this consideration. However, it is impossible or impractical to automate the measurement of many complex team behaviors, which must, therefore, be evaluated via some sort of human observers.

Evaluators can ensure high inter-rater reliability in several ways. Before the fact, they can design measurement procedures that are unambiguous and refer to aspects of performance that are easily observed. If the ratings demonstrate reasonably high inter-rater reliability, evaluators could be justified in aggregating the ratings. If observers do not agree, however, evaluators may take a number of steps to mitigate the problem. They could elect to use only those items about which the observers agree. If this results in the rejection of too much data, evaluators may elect to have the observers review each of the ratings in question, and negotiate, either directly or through a mediator, to resolve their differences.

3. Sensitivity and Specificity

The concepts of sensitivity and specificity of measures have evolved from the medical testing tradition (Yerushalmy, 1947). Sensitivity and specificity of measures refer to the classification accuracy of a diagnostic test that specifies that a patient either has or does not have a particular disease. Table III-1 (a 2 × 2 model) shows the possible outcomes of such a test.

Given this model, the sensitivity of a test is the probability that the diagnostic test correctly classifies diseased subjects, or $p(\text{TP})/[p(\text{TP}) + p(\text{FN})]$. Similarly, the specificity of the screening test measures the extent to which the diagnostic test correctly classifies subjects free of the disease, or $p(\text{TN})/[p(\text{TN}) + p(\text{FP})]$. Although this example is stated in terms of a dichotomous measure, the concepts of specificity and relevancy can be generalized from the 2 × 2 model to more complex continuous measures of performance, as discussed below.

Table III-1. Outcomes From a Diagnostic Test

		True State	
		Has Disease	*Does Not Have Disease*
Test Diagnosis	*Has Disease*	True Positive (TP)	False Positive (FP)
	Does Not Have Disease	False Negative (FN)	True Negative (TN)

a. Sensitivity

A sensitive test is one that correctly discriminates among levels of the measurement in question. Insensitive measures, on the other hand, can mask real differences in training effectiveness and should be avoided for evaluations of LSTSs or any other training device or system (Boldovici, 1987). An example of an insensitive measure is one limited by a performance ceiling, such as tests for accuracy that are distributed close to 100 percent or error measures close to zero. Other measures (e.g., activity counts) may show performance differences for extreme values of independent variables but may be relatively insensitive for moderate values. The sensitivity of the test can be estimated by examining the frequency distributions of measures themselves and the bivariate distributions (i.e., scatter diagrams) of the measures and the training variables.

b. Specificity

A specific test is one that correctly excludes individuals who do not have the characteristic in question. A nonspecific test, on the other hand, is one that taps into constructs other than the one in question. An example of a nonspecific measure is using a test of tactical knowledge as the sole outcome from training on an LSTS. Although performance on this test may be correlated with LSTS usage, it may also be correlated with other important variables, such as military education. The central problem with this nonspecific measure is that it underrepresents the target concept: learning outcomes that result from training on the LSTS, which includes relevant skills, attitudes, and knowledge.

4. Utility

In the context of practical performance measurement, Lane (1986) argued that performance measures might be reliable and valid but still be of limited utility for measuring

training effectiveness. What must be considered is the utility or value of measuring a particular aspect of performance in light of the associated costs of measurement. In that regard, Lane (1986) identified two dimensions or attributes that determine measurement utility: the effectiveness of measures relative to alternative measures and the practicality of implementing the measurement scheme. These two dimensions can be combined into a single index using multiattribute utility measurement.[3] It is more likely, however, that the analyst assesses these dimensions subjectively or considers them for minimum standards.

a. Effectiveness of Measures Relative to Alternative Measures

The effectiveness of a performance measure must be considered in relation to other performance measures or no measurement at all. The issue is whether the measure improves, or potentially improves, the analyst's understanding of the actual impact of the training device. Part of this decision depends on the extent to which trainee performance affects the measure in question. For instance, consider two outcome measures from attack operations of time-critical mobile targets: the number of opposing forces (OPFOR) missile launches and the number of OPFOR missile launchers destroyed. These two measures are equally reliable and valid; however, the first outcome is a product of OPFOR and friendly Blue forces (BLUEFOR) actions, whereas the second outcome is more directly caused by BLUEFOR targeting. Both measures may be useful in describing mission outcomes, but the launcher kills may be a more effective measure for BLUEFOR training outcomes.

b. Practicality of Implementing the Measurement Scheme

A measure that is impractical to implement has little utility. The practicality refers to the feasibility of implementation and to user acceptance. An example of an infeasible performance measure is one that requires individual observers for each participant in a large-scale tactical exercise. Furthermore, research on user acceptance has shown repeatedly that neither the instructor nor the trainee views raw (i.e., disaggregated) performance data favorably. Thus, whereas the collection of detailed performance in LSTS systems may be feasible, it may not be practical.

---

[3] The topic of multiattribute utility measurement is discussed in more detail in Section IV as a general method for determining the training effectiveness of training systems. Here, the suggestion is to use the method, in a more limited sense, to evaluate the utility of performance measures.

B. TYPES OF PERFORMANCE MEASURES

For measuring training effectiveness, performance measures can be grouped into two broad categories:

1. Objective measures based on actual performance
2. Subjective measures based on human judgment.

Although complex measures could have both objective and subjective characteristics, this dichotomy is useful in defining the domain of performance measures relevant to training effectiveness. These two broad categories of performance measures are described below. Section IV discusses the methods for using both types of performance measures to assess training effectiveness.

1. Objective Performance-Based Measures

Most analysts agree that, all things being equal, measures based on actual performance are the preferred data source for TEAs. The foremost reason, perhaps, is that these performance-based measures are relevant to the goal of TEAs, which is to determine the extent that training directly improves the performance of individuals and collectives. Another advantage of performance-based measures is that properly constructed measures provide objective and quantifiable indexes of training effectiveness that can be manipulated mathematically to predict effectiveness and to develop tradeoffs between performance and cost variables. Despite the agreement that performance-based measures are preferred, an issue remains: What specific aspect of performance should be measured—performance processes or performance outcomes?

a. Distinction Between Process vs. Outcome Performance Measures

Recent large-scale, simulation-based training programs have advocated the exclusive use of either process or outcome performance measurement approaches. *Process measures* indicate how well tasks are performed in terms of speed, accuracy, or completeness. Examples of process measures are the number or percent of task steps performed correctly or the time to complete a task. *Outcome measures* refer to task products or outcomes. Outcome measures can reflect the intermediate task products or outcomes (e.g., number of enemy targets destroyed during a mission segment) or the terminal outcome of a complex task or mission (e.g., total enemy losses/friendly losses). Table III-2 compares and contrasts these two approaches, which are then discussed in more detail.

Table III-2. Comparison of Outcome and Process Measures

Features	Process Measures	Outcome Measures
Focus	Required behavioral processes or actions	Intermediate and terminal products or outcomes of performance
Strengths	• Explains causes of performance • Provides more data points	• Directly relevant to task goals and values • Objective and comprehensive
Weaknesses	Reliance on subjective judgment	Often Inherently unreliable
Requirements	Knowledgeable human judges	Lasting trace or product of performance
Potential for automation	Can be partially automated	Can be fully automated
Example	Targeted Acceptable Responses to Generated Events or Tasks (TARGETs)	Unit Performance Assessment System (UPAS)

b. Process Measures

Advocates of the process-measures approach include the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) in its development of structured training for Army simulation-based exercises (e.g., Hoffman et al., 1995) and NAWCTSD in its development of a performance measurement system for air combat maneuvering (Lane, 1986) and for military teams in general (e.g., Fowlkes et al., 1994).

Supporters of this approach argue that outcome measures, although they may help identify exceptionally good or poor performance, do not provide information on the causes of performance. To diagnose performance problems and provide feedback, process measures are needed. Process measures are also favored for those tasks that have no permanent product or trace (e.g., certain staff actions). Process measures can also have a statistical advantage because they can provide more data points than outcome measures. A unit may provide only a single terminal outcome (e.g., the ratio of enemy to friendly losses) for each mission. In comparison, multiple process measures may exist, reflecting the fact that certain tasks are performed more than once by multiple units. Finally, recent advances in the measurement of simulation performance have improved the reliability of process measures. These include increased structure of mission scenarios to ensure that key actions occur and improved instructions for assessing performance. Further, aspects of process measurement

can be semi-automated. For instance, portable computers can be used to cue O/Cs when to score performance and to constrain responses appropriately.

c. Outcome Measures

Advocates of the outcome-measures approach include Dynamics Research Corporation (1996) in its development of performance measures for the Universal Joint Task List (UJTL) and BDM Federal, Inc. (1995) in its development of measures for the Army's Critical Combat Functions (CCFs).

Supporters of this approach argue that outcome measures, in comparison with process measures, are more directly relevant to task and mission goals and are, therefore, more valid measures of effectiveness. For instance, the combat value of a simulation is measured more directly by battle outcomes than by the proportion of correct responses. Advocates also feel that outcome measures are more objective and quantitative indexes of performance than process measures, which are often based on subjective judgments of behavior.[4] Moreover, since computer-based simulations can detect many important outcomes (e.g., target hits/misses, time to traverse between two points, precise locations), measurement can be automated and completely free of human bias. For these reasons, the outcome measures are often portrayed as being less labor intensive and more accurate and reliable than process measures.

d. Combination Approach

A third point of view advocates that process and outcome measures are complementary and should be used together. Combining process and outcome measures provides a more complete picture of performance. Lickteig (1996) has also pointed out that the relationship between process and outcome is itself an interesting research question. The Orlansky et al. (1997) evaluation of the MDT2 provides an example of this approach. The MDT2 study incorporated a method called TARGETs (Targeted Acceptable Response to Generated Events or Tasks) to measure team processes in a simulated close air support task. The collection of outcome measures was automated by the Unit Performance Assessment System (UPAS), which tapped into the distributed interactive simulation (DIS) data stream that controlled the multistation simulation. The use of process and outcome

---

[4] Outcome measures are generally objective but not necessarily. Certain outcomes must be subjectively measured (e.g., adequacy of operational plans) because of the lack of specific measurement criteria.

measures provided mutual support for the conclusion: Repeated exercises on MDT2 resulted in improvements to performance on close air support tasks.

This combination approach is hardly new. The Interservice Procedures for Instructional Systems Development (IPISD) (Branson et al., 1975) advised using both types of measures for tasks where the process leads to a product that is particularly influenced by process errors. The point of unearthing this approach is to counteract the recent tendency to advocate either the process or the outcome approach to the exclusion of the other. Both types of measures are useful for evaluating the training effectiveness of LSTSs.

2. Judgment-Based Measures

In many cases, TEAs cannot be based on objective, performance-based data. For instance, the training system may not have evolved to the point that it can be used to train (and test) personnel operationally. This is especially true for LSTS systems, which typically have long-lead development times. Also, performance alone cannot address some effectiveness issues, such as the case where performance only indirectly addresses the effectiveness of instructional features. Finally, performance-based measures may not be reliable, valid, or useful.

For these and other reasons, researchers often turn to judgment-based measures as substitutes for or supplements to performance-based measures. Two types of judgment-based measures can be distinguished:

1. Training capability measures, which are designed to assess the capability of training systems.
2. Training reaction measures, which are designed to assess user reaction to the systems or to the training that they have received.

a. Training Capability Measures

This category of judgment-based measures includes the detailed analyses of the capability of training devices to train skills that can be transferred to operational settings. Although this approach can be used for existing training simulations, the advantage of this method is that it can also be used for devices that are still in the conceptual stages of development. The analyses are usually performed by SMEs who are familiar with the tasks and the proposed device or simulation. In essence, these analyses attempt to measure the effectiveness of a training system through the systematic examination of its capabilities.

Section IV describes some of the specific designs that have been developed for obtaining SME judgments in assessing training effectiveness.

Regardless of the approach, one of the key concerns in obtaining training capability measures is the inter-rater reliability—or the agreement among the SMEs. To attribute validity to the SMEs' judgments, most of the methods assume that the raters must agree. If ratings are highly variable, analysts have often used consensus-building methods to induce agreement. An example of a consensus-building method is the Delphi Technique, a procedure that RAND developed in 1948 to achieve the consensus of experts while minimizing conflict. In this technique, group members work individually on estimates, which are then compiled by a group facilitator. The facilitator summarizes the results and identifies outliers (those group members whose responses fall outside an acceptable range). The outliers are asked to justify their estimates in writing. Then, the facilitator circulates the summaries and the justifications (without attribution to individuals), and the group members consider this feedback and revise their estimates accordingly. This process continues until the group achieves consensus.

Although the Delphi Technique is a well-known consensus-building technique, others are available. The interested reader should refer to Delbecq, Van de Ven, and Gustafson (1975) for more details on this and other methods for facilitating consensus decisions in small groups.

b. Training Reaction Measures

Training reaction measures are used to assess the users' reactions to the training system. "Users" include the training audience (i.e., the trainees) and relevant unit commanders and trainers. These reactions can be viewed as assessing two related aspects of training effectiveness:

- First, training reactions reflect the users' general impressions of the system. The implication of such impressions is that trainees, trainers, and commanders who react negatively to training devices will use them ineffectively or not use them at all. Thus, positive reactions to training devices are a necessary, but not sufficient, condition for training effectiveness.
- Second, training reactions provide specific feedback on those training objectives that the users think the system addresses and how well the users think the system trains those objectives. As such, training reactions can provide some of the basic information on the face validity of the training system.

Training reaction measures are usually obtained through structured questionnaires. The questionnaires can be formatted in several different ways, including open-ended items, multiple-choice items, scaled responses, and ranking tasks. Fletcher (1988) provides an example of a formal assessment of trainee reactions. He used a variety of questionnaire items to measure user perceptions of the SIMNET system—one of the earliest examples of a virtual, LSTS system. This analysis provided one of the earliest identifications of SIMNET's strengths and weaknesses.

As with training capability measures, training reaction measures are subjective because they are based on human judgments. However, in contrast to training capability measures, training reaction measures must be collected against existing devices, although users might conceivably be able to react to breadboard versions of the device. If, however, the evaluated version does not resemble the operational version, the user may not provide valid reactions to the system.

Another characteristic of training reaction measures that distinguishes them from training capability measures is that respondents are the users—as opposed to SMEs. Further, these users may differ in their reactions to the training system so that inter-rater reliability is not an overriding issue, as it is for training capability measures. For these reasons, the analyst must document the central tendency and the variability of user responses. If possible, the analyst should also try to determine the source of the differences in ratings. For instance, such differences may be caused by variability in the amount or type of experience that the user brings to the training context.

IV. RESEARCH DESIGN

In the context of statistical analysis, the term "design" refers to the arrangement of experimental treatments and subjects in an experiment. A good design is one that allows valid inferences about the effects of independent variables (e.g., variations in the quality or quantity of simulation training) on dependent variables (e.g., individual and collective measures of performance). In this paper, we define "designs" more generally as data collection procedures that allow us to make inferences about the effectiveness of LSTSs. Hoffman and Morrison (1992) and Simpson (1999) provide detailed discussions of research designs for evaluating the effectiveness of military training devices.

In a now-classic paper, Kirkpatrick (1976) distinguished among four types or "levels" of training effectiveness evaluations. Within each level, a variety of performance measures can be obtained. These levels are distinguished by the procedures for collecting the data (i.e., the research designs). The levels are ordered in a sequence of four progressively difficult and time-consuming steps or procedures that yield increasingly comprehensive and convincing information about a training program or system. Each of the four levels is described briefly, as follows:

- **Level I: Reaction.** This type of evaluation is based on the trainee's reaction to or opinions of the training program or simulator. The rationale for this analysis is that trainees must first like the system to initiate training and maintain interest in its contents. However, the fact that trainees like a training system does not ensure that the system trains the appropriate skills and knowledge effectively.
- **Level II: Learning.** The next type of evaluation focuses on determining whether the training program or device is effective in imparting the skills and knowledge in question. This effectiveness is established by measuring the change in performance that occurs because of practice on the device. Performance is measured either on the system itself or in an offline assessment of skills and knowledge. Evidence from a Level II evaluation is considered more convincing than trainee reactions (i.e., Level I evaluations).
- **Level III: Transfer.** The purpose of this type of evaluation is to determine whether the skills and knowledge learned during training transfer to improvements in job performance. As implied, Level III evaluations require methods

for assessing performance on the job and determining the relationship between training and job performance. Often, Level III evaluations also include a demonstration of learning on the training system (i.e., a Level II evaluation).

- **Level IV: Results.** The objective of this type of evaluation is to determine whether skills learned during training and transferred to job performance actually make a difference for the organizations. For commercial businesses, Level IV evaluations are usually cast in terms of return on investment (ROI). Few examples of Level IV evaluations exist because of the difficulty in attributing organizational outcomes to training. At the same time, Level IV evaluations are potentially valuable for determining the actual value of training to the organization.

Level III and Level IV evaluations are particularly problematic for military applications. Level III evaluations require measures of individual and small unit job performance, such as weapon marksmanship scores or aircrew proficiency assessments. The diligent analyst can obtain these measures, but not without considerable expenditure of time and resources. Level IV evaluations introduce a more difficult problem because they require measures of the military's "bottom line" (i.e., effectiveness in combat). As Fletcher and Chatelier (2000) point out, the opportunities to collect measures of actual combat effectiveness are mercifully rare and are not conducive to experimental control. In lieu of actual combat performance measures, surrogate measures can be obtained through live field simulations (e.g., the National Training Simulation or Red Flag exercises) or through analytic, computer-based simulations. The problem with these surrogate measures is that their relationship to actual measures of combat effectiveness is not known. From the point of view of face validity, it is certainly reasonable to expect that performance in simulated combat simulations should be positively related to performance in combat. Further, the finding that training on LSTSs increases performance at live field exercises is usually interpreted as evidence in favor of LSTS training effectiveness. However, the real-world implications of such a finding are limited because no data exist on whether or how much increases in simulated combat performance translate into increases in actual combat performance.

Kirkpatrick's schema classifies TEA designs by abstract evaluation functions or questions. We offer a similar taxonomy that classifies TEAs by the research operations used or the types of data collected. These three broad categories of TEAs are training surveys, empirical performance-based designs, and training analyses. Despite the apparent differences between the taxonomies, our categories can nevertheless be mapped back into Kirkpatrick's schema to discuss evaluation functions and goals.

A. TRAINING SURVEYS

The survey method is the usual approach for obtaining user reactions. The purpose of training surveys is to obtain users' opinions about the training system. This method corresponds roughly to Kirkpatrick's Level I evaluation, although Kirkpatrick's focus is explicitly on the reactions of trainees to the exclusion of other system users (e.g., training managers, unit commanders, and trainers). Even for TEAs that focus on actual performance (vs. user opinions), training surveys are often included in evaluations because user reactions are clearly relevant to training effectiveness and survey responses are relatively easy to collect and interpret.

The process of designing a survey is typically segmented into four phases:

1. Devising a sampling strategy
2. Choosing an appropriate survey mode
3. Constructing the instrument
4. Analyzing the data.

The following subsections discuss each of these concepts within the context of LSTS evaluations.

1. Devising a Sampling Strategy

The political or marketing surveys, with which many of us are familiar, are intended to capture the opinions of a large population of respondents, such as those likely to vote in the next presidential election or the number of U.S. men aged 18 to 65. To obtain a fair representation of these opinions, survey designers devise a strategy that systematically samples from that population. Sampling strategies are based on statistical theories and can sometimes be complex in execution. Fortunately, the sampling strategies for LSTS research are generally less complex than those used in these examples.

For LSTS systems currently under development, only a limited number of individuals have been exposed to these systems. In this case, the strategy usually adopted is to survey all—not just a sample of—the users as they finish their initial operational training on the device. The individuals participating in operational testing provide a captive audience, and surveying them soon after the test minimizes forgetting.

For systems that have been fielded for a substantial period of time, the analyst may devise a different sampling strategy. The most difficult aspect would be identifying which individuals or units have had substantial experience with the system in question. If the

numbers are sufficiently small, however, the analyst may still opt to administer the survey to all individuals who have been trained on the system, as opposed to a sample of individuals who have been trained on the system.

2. Choosing an Appropriate Survey Mode

The standard modes for administering surveys are by mail (standard or electronic), telephone, and in-person interviews. The most popular mode is the mail survey that incorporates a paper-and-pencil questionnaire; however, computer-automated versions of paper-and-pencil questionnaires are growing in popularity. The traditional paper-and-pencil questionnaires distributed by standard mail and computer versions transmitted by e-mail provide the means to obtain highly structured data from a large and diverse sample of respondents. On the other hand, in-person interviews and, to a limited extent, telephone surveys offer the advantage of having a live interviewer to provide immediate clarifications on survey items. Furthermore, the mere presence of an interviewer can increase cooperation rates.

The typical training survey combines the best of a paper-and-pencil questionnaire and in-person interviews. Someone who is familiar with the questionnaire (often the questionnaire designer) and can answer questions from individuals or from the group usually proctors the survey.

Another less-standard approach for obtaining training reactions uses a focus group. Users are part of a carefully planned discussion group designed to obtain perceptions on a defined area of interest in a permissive, nonthreatening environment. This group includes approximately 7 to 10 people and is facilitated by a skilled interviewer. The interviewer keeps the discussion comfortable and enjoyable for the participants and encourages them to share their ideas and perceptions (Krueger, 1994). Klein Associates recently used this approach to evaluate the effectiveness of technologies used in Prairie Warrior, a corps-level exercise that serves as the final event for students in the Army's Command and General Staff College (Space and Naval Warfare Systems Center, 2000). Although the focus-group approach is potentially effective for eliciting training reactions, it is not often used because it requires an experienced interviewer trained in focus-group techniques.

3. Constructing the Instrument

This subsection reviews some of the general principles of survey construction as they apply to LSTS evaluation. Babbitt and Nystrom (1989a) provide more detailed

guidance in constructing questionnaires for military applications, and Hagin et al. (1982) provide more specific information on survey construction as it relates to TEAs.

The content of a training reaction survey follows logically from its intent. For instance, users may be asked how they like specific features of the training device, to what extent the device trains particular tasks, the relative ease of learning, and the extent to which skills learned on the device transfer to the job. The survey developer is also able to address topics not obviously related to training effectiveness. This is where surveys can be tailored to the specific capabilities and problems inherent in LSTSs. For instance, survey items could be designed to assess the ease with which scenarios can be configured to support mission rehearsals.

In composing surveys, the analyst should ensure that each item addresses a single issue, is as brief as possible, and is easy to understand. Items should be carefully written to avoid bias and the use of leading questions. The analyst should also be careful to avoid questions that the respondent cannot or should not answer. For instance, the analyst should not ask the user to comment on the technical aspects of the simulation. Finally, items should be written so as to be easily computer coded to facilitate subsequent statistical analyses.

A more worrisome problem in constructing training reaction surveys is that these surveys often have the implicit demand that the user respond positively toward the system or device. Such surveys are obviously written and administered by individuals who have an interest in the success of the device, and, consequently, the respondent is often predisposed to "help" the developer by providing the outcome that the respondent perceives as being the desired one. Detractors of this approach often refer to training reaction surveys as "smiles" tests because the outcome is usually preordained to be positive.

Surveys can be constructed to mitigate (at least partially) the problem of demanding that the user respond positively toward the system or device. To illustrate, consider the following example, which is based on one of the most widely used survey scales: the Likert scale. According to this technique, the analyst constructs statements of opinion, and the respondents indicate the degree to which they agree or disagree by choosing one of five or more mutually exclusive and exhaustive response categories (Babbitt and Nystrom, 1989b). For instance, an evaluator of an LSTS may construct a statement about the ease of reconfiguring an LSTS as follows:

> **The simulation scenarios can be easily configured to support mission rehearsals.**

Along with this statement, the evaluator provides a scale from which respondents indicate the degree to which they agree or disagree (see Figure IV-1).

5	4	3	2	1
Strongly Agree	Agree	Neither Agree Nor Disagree	Disagree	Strongly Disagree

Figure IV-1. Example of a 5-point Likert Scale

To avoid the implication that the survey developer seeks only positive responses, it is advisable to word the items negatively on occasion. This has the added benefit of requiring respondents to read the items closely. For example, the previous item can be rewritten as follows:

> **The simulation scenarios *cannot* be easily configured to support mission rehearsals.**

Again, the respondent indicates the degree of agreement/disagreement by using a 5-point Likert scale (see Figure IV-1).

4. Analyzing the Data

The survey should be constructed to facilitate the analysis of data. The general goal of the analysis should be to describe the extent to which the respondents responded positively to the training device or simulation. If all items are similarly scaled (e.g., 5 = most favorable, 1 = least favorable), the analyst could report the average rating of all items for all respondents. For instance, Fletcher (1988) calculated the average rating for all items of a survey on SIMNET. Using a 5-point scale similar to the one in Figure IV-1, he reported an average rating of 3.63, which indicates an overall positive impression of the simulation system. The analyst must, however, note any items that do not conform to the overall trend in the data. Accordingly, Fletcher noted that only 4 of the 35 rated items on his survey were rated unfavorably (i.e., less than the 3.0 midpoint). This example points to the fact that numerical ratings can be used to determine the reactions to individual items or to characterize responses to the questionnaire as a whole.

If qualitative items are used, the analyst can choose to report the number or percent of respondents responding favorably. For example, Fletcher reported that 91 percent of the respondents (30 of 33) replied positively to the item asking whether they would use SIMNET if they had access to it. Qualitative items present a problem because they are not easily combined into an overall measure and, therefore, must be considered individually.

In designing the survey, retaining variables that can affect user reactions is important. Suppose, for instance, the analyst believes that the responses of the simulation users may vary as a function of experience. In that case, the analyst must design questions to determine the experience level of the respondent. Fletcher (1988) calculated one such revealing cross tabulation. He showed that armor commanders (tank commanders, platoon leaders, company commanders, and battalion staff) rated SIMNET more favorably than tank crewmembers (gunners, drivers, and loaders) rated it. Evidently, the senior respondents appreciated the abstract tactical value of SIMNET more than the tank crewmembers did.

B. EMPIRICAL PERFORMANCE-BASED RESEARCH DESIGNS

Performance-based research designs represent the most highly respected approach for measuring training effectiveness because they examine, in realistic contexts, the effects of training on the execution of job tasks. In a now-classic paper, Campbell and Stanley (1963) described the challenges and advantages of field experiments by extending the concept of the validity to the design of experiments. For research design, they distinguished between internal and external validity. Internal validity refers to the extent to which the design permits valid inferences about the effects of the experimental treatments on performance. External validity relates to the degree with which results from an experiment can be generalized to the population of interest.

Campbell and Stanley defined 12 different threats to internal and external validity and related each to 16 research designs. Included in the designs are the highly controlled "true" experiments that characterize laboratory research and the "quasi-experiments" that lack some of the controls of a true experiment but are more practicable for field situations. Their analysis revealed the strength and the weakness of quasi-experiments: whereas quasi-experiments may lack the controls required to mitigate threats to internal validity, they are generally less susceptible to threats to external validity than the "true" designs that require arbitrary and restrictive conditions, such as random assignment to treatments.

Although performance-based research designs represent a respected approach for measuring training effectiveness, they are also the most resource-intensive approaches for measuring training effectiveness. Hoffman and Morrison (1989) demonstrated that a positive correlation exists between the comprehensiveness of the design and the experimental resources that would be required to implement the design. In other words, the most

comprehensive designs (the ones that address the most questions and have the most controls for potential threats to validity) are also the most costly to implement.

To avoid incurring unnecessary costs, Hoffman and Morrison suggested that analysts must define clearly the most important objectives of the assessment for questions they want answered. Theoretically, this should dictate the type of design. However, experience indicates that the actual design choice represents a compromise between what researchers want to answer and what they can answer practically.

Hoffman and Morrison (1992) and Simpson (1999) detail several different types of performance-based designs that have been used to evaluate the effectiveness of training devices. The following summarizes these discussions by dividing designs in three broad categories:

1. Randomized (true) experiments
2. Quasi-experimental designs
3. Correlational studies.

Trochim (1999) and others have argued that real-world evaluations are actually amalgams of these "pure" types. However, considering them separately is useful in understanding the pros and cons of each. A comparison of these three approaches follows the individual discussion of each.

1. Randomized (True) Experiments

The randomized—or true—experiment represents the "gold standard" for TEAs. To illustrate, consider the simple experiment outlined in Table IV-1. Even though the experimental group is the only group exposed to the training intervention X, the experimental group and the control group are tested before and after training on knowledge and skills purportedly trained by X. The crucial requirement for a randomized experiment is that the participants (e.g., individual trainees, operational units) have to be randomly assigned to experimental and control groups. By requiring random assignment to experimental treatments, the researcher can make strong inferences about cause-effect relationships between independent variables (training conditions) and dependent variables (performance measures).

Field researchers rarely have full control over the situation, making it difficult to assign participants randomly to treatments. Although clearly a challenge, Cook and

Table IV-1. Outline of Pretest-Posttest Control Group Design

Treatment	Pretest	Train on X	Posttest
Experimental	Yes	Yes	Yes
Control	Yes	No	Yes

Campbell (1979) outlined several approaches for conducting randomized experiments in real-world settings. They argue that the value of randomized experiments is crucial in those situations where the costs of being wrong about a causal inference are high. In the context of LSTSs, such costs can be identified as the Service-wide or joint implementation of an ineffective system or the elimination or reduction in scope of an effective system. Thus, the most appropriate use of a randomized experiment for evaluating the effectiveness of an LSTS should be for crucial evaluations where the results are expected to affect implementation decisions.

For LSTS evaluation, two types of randomized training experiments can be distinguished: learning experiments and transfer experiments.

a. Learning Experiments

Learning experiments are evaluations to determine the amount of learning that occurs as a direct result of training on the simulation system. As an example, the experiment outlined in Table IV-1 could be considered a learning experiment if the pretests and posttests assess outcomes (i.e., knowledge, skills, attitudes) addressed by the training system. Clearly, learning experiments can be interpreted as a Level II evaluation in Kirkpatrick's schema.

The measures of learning that can be derived from such experiments depend on the exact design used. One of the simplest approaches is to measure performance before training (pretest) and after training (posttest) on the to-be-evaluated system. The pretest and posttest can be administered on the system itself or off line using a separate instrument for assessing learning outcomes (e.g., a paper-and-pencil test). In this case, a typical measure of learning would be percent improvement, defined as follows:

$$(\text{pretest} - \text{posttest})/\text{pretest} \times 100 \quad ,$$

where "pretest" is the mean performance on a test administered before training on the target system and "posttest" is the mean performance on a test administered after training on the target system.

If a two-group design similar to that depicted in Table IV-1 is employed, percent-improvement measures should be calculated for the experimental and the control groups. The rationale is that the improvement shown in the experimental group can be attributed to the effects of repeated testing and to the training program. The improvement in the control group, on the other hand, should reflect improvement caused by repeated testing only. Thus, the difference between the two values should reflect the effect of the training program controlled for the effect of testing.

Note that such simple pretest/posttest learning experiments provide results under one set of learning conditions—the conditions used in the experimental group. Changes to any of those conditions—most notably the length of training—would likely result in different values for the improvement measures.

The preferred approach is to determine the device learning curve, which is a mathematical function that describes the course of skill acquisition over repeated training and testing on the simulation. (This approach is facilitated when the to-be-evaluated training system has integrated performance-testing scenarios.) Given a series of successive performance tests, an analyst can use parameter-fitting techniques to derive a learning function for the data. Among the many different forms of the proposed learning functions, the so-called "power law of practice" is a relatively simple formulation that appears to apply to a wide variety of tasks and performance measures (Newell and Rosenbloom, 1981). According to this formulation, the effects of learning on speed of performance can be summarized by the following simple equation:

$$T = BN^{-k} \quad ,$$

where T is the time to perform task, k is the learning rate parameter, B is the initial performance on task before training, and N is the number of learning trials. Although this particular form of the power law is stated in terms of speed of performance, similar relationships hold for other aspects of performance, such as accuracy.

For example, a learning curve was fitted to data recently gathered in connection with J9901, the first human-in-the-loop (HITL) experiment conducted by the Joint Warfare Experimentation Battle Laboratory (Joint Advanced Warfighting Program, 2000). The J9901 experiment used an LSTS to conduct research on attack operations against time-critical mobile ground targets (e.g., mobile ballistic missiles). One of the dependent variables collected was the time that elapsed from the point that the targeting cell nominated a target to the point that a weapon was tasked against the target. This interval was intended to

measure the efficiency with which the targeting cell prosecuted a target. The median value of this measure was calculated for each of five successive experimental trials. The results, shown in Figure IV-2, indicated a 54-percent improvement in speed of performance from the first to the last trial. Furthermore, the data were well fit by the function $T = 7.57N^{-0.47}$. The function indicates the following: whereas the targeting cell substantially improved in its performance across trials, further improvement would be expected with additional trials. Specifically, the function predicts that performance on a sixth trial would be 17 seconds faster than that on the fifth trial—a substantial difference against time-critical targets.

b. Transfer Experiments

The conclusions that one can draw from learning experiments are limited to performance on the training device or simulation. To make inferences about performance outside of this context, the evaluator must consider a transfer experiment. The purpose of a transfer experiment is to assess the degree to which learning on the training system transfers to and improves performance on the operational equipment. Because transfer experiments potentially provide more convincing evidence of training effectiveness than learning experiments, they are considered higher type (Level III) evaluations in Kirkpatrick's schema.

The simplest two-group case (analogous to the simple pretest/posttest learning experiment) includes an experimental group that receives training on the system in question and a control group that does not.[5] Then, the evaluator compares the performance of both groups on the operational equipment. A typical measure of transfer that can be derived from this design is percent transfer, which is defined as follows:

$$[(\text{experimental} - \text{control})/\text{control}] \times 100 \quad ,$$

where "experimental" is the mean performance on operational equipment of the group receiving simulation training and "control" is the performance on operational equipment of the group receiving no simulation training.

If performance is measured by time or the number of trials required to reach some standard of performance on the operational equipment, this measure is sometimes

---

5 In fact, the design in Table IV-1 could be considered a type of transfer design if the performance tests were administered on the operational equipment. However, most transfer experiments employ a posttest-only design (i.e., no pretesting).

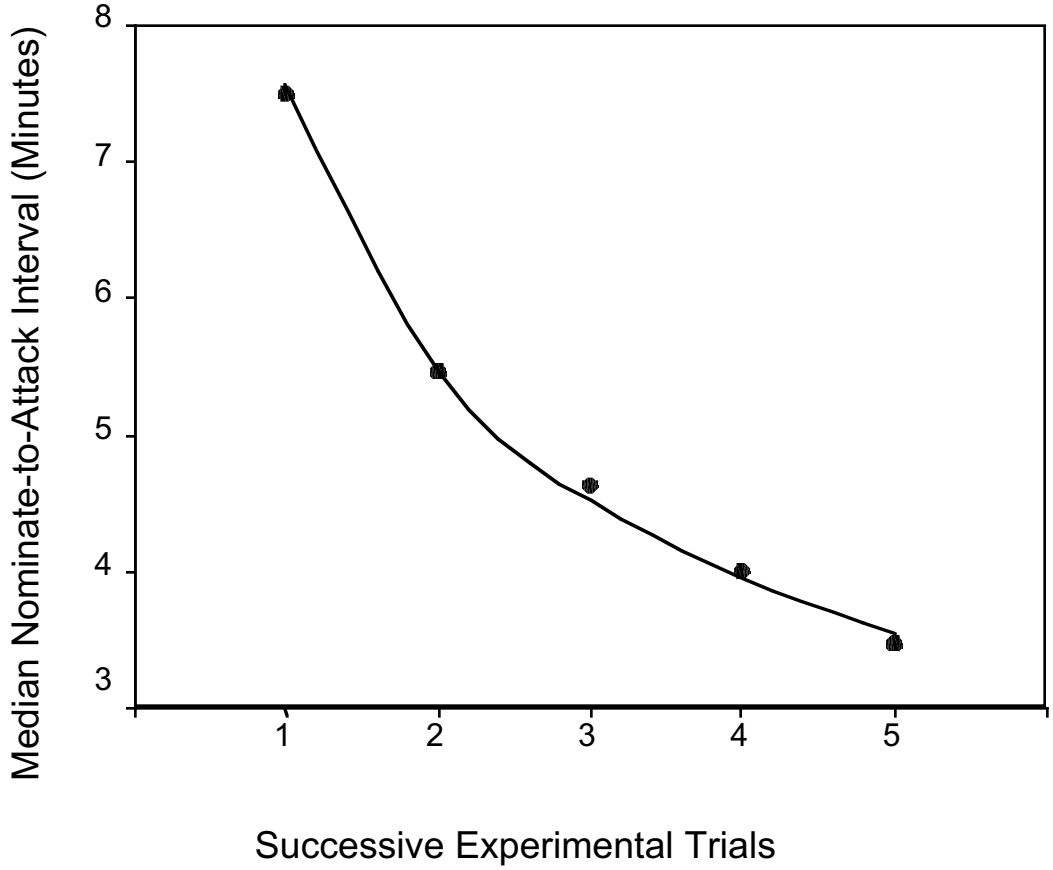

Figure IV-2. Median Target Prosecution Times From a Joint Experiment on Attack Operations

interpreted as "percent savings." The rationale is that this measure reflects the portion of training on the operational device that can be replaced (and therefore "saved") by training on the simulation system.

Roscoe (1971) argued that percent transfer measures do not reflect the efficiency with which the simulation trains the task(s) in question. To do so, he suggested an improved measure, the cumulative transfer effectiveness ratio (CTER), which he defined as follows:

$$[\text{control (OE)} - \text{experimental (OE)}]/[\text{experimental (SIM)}] \times 100 \quad ,$$

where "control (OE)" is the mean time or number of trials the control group required to reach standard on the operational equipment, "experimental (OE)" is the mean time or number of trials the experimental group required to reach standard on the operational

equipment, and "experimental (SIM)" is the mean time or number of trials the experimental group required to reach standard on the simulation.

A CTER of 1.0 indicates that the simulator is as efficient as the operational equipment, whereas a value more or less than 1.0 indicates that the simulator is more or less efficient than the operational equipment. However, as Boldovici (1987) pointed out, a result of 1.0 can be obtained from an infinite number of combinations of simulator and operational equipment training as long as training trials on the two media are interchangeable. Boldovici, in his example, assumed that using the operational equipment alone required 20 trials to reach the performance standard. In one case, 18 trials were required to reach the performance standard on the simulator, but only 2 additional trials were required to reach standard on the operational equipment. In the other case, 2 trials were required to reach standard on the simulator, but an additional 18 trials were required to reach standard on the operational equipment. In either case, the CTER was equal to 1.0, which seems counterintuitive given that the second case requires nine times more training on the operational equipment.

To determine how the amount of training affects transfer performance, Boldovici (1987) argued that researchers must employ more than two levels of training on the simulation (i.e., 0 vs. some fixed amount). Povenmire and Roscoe (1973) used such a design when they evaluated how training on a Link GAT-1 simulator affected flight performance in a Piper Cherokee Airplane. Student pilots were randomly divided into four groups that received either 0, 3, 7, or 11 hours of simulator training. All four groups were trained concurrently on the Piper Cherokee, and the number of training hours required to reach private pilot certification was recorded for each student. The results showed that the 11-hour group obtained the largest percent transfer value (16 percent); however, the 3-hour group obtained largest CTER value (1.53), which takes into account amount of simulation training.

As an alternative to Roscoe's CTER formulation, several researchers (e.g., Bickley, 1980; Sticha et al., 1988; Hoffman and Morrison, 1992; Holding, 1991; Jones and Kennedy, 1996) have advocated the use of "isoperformance" functions for determining the results from multigroup transfer experiments. These isoperformance functions are so named because they display different mixes of simulation and equipment training needed to maintain a fixed standard of performance. As with the CTER, the independent and dependent variables in isoperformance functions can be defined in terms of trials or training time. A particularly useful isoperformance function is one that defines those variables in terms of training costs. Although we have not, to this point in our paper, considered cost factors in

training evaluations, the following discussion illustrates how training costs and measures of training effectiveness can be integrated to conduct more comprehensive and realistic evaluations of training systems. For instance, Boldovici's (1987) previous example indicated that markedly different results from two experiments could lead to the same value in training effectiveness (i.e., CTER = 1.0). Had the measure taken costs into consideration, the two results would have diverged in a reasonable manner.

To illustrate how costs can be incorporated in an isoperformance function, performance results from Povenmire and Roscoe (1973) were integrated with the reported cost rates for training: $16/hour in the simulator vs. $22/hour in the aircraft. Figure IV-3 shows the isoperformance curve for total costs as a function of simulator time. These data indicate that 3 hours of simulation training paired with about 40 hours of training on the aircraft was the most cost-effective mix of training. Increasing the amount of simulation training to 7 hours decreased the aircraft training by 2 hours; however, the small decreases in aircraft training costs were offset by the increases in simulation training costs. Nevertheless, the 7-hour condition was still cost effective relative to the 0-hour control condition. Although the 11-hour condition yielded the greatest value for percent transfer (16 percent), the total cost data indicate that this amount of simulation training was actually not cost effective because the total costs in this condition ($996.60) were slightly greater than the total costs in the 0-hour condition ($978.78).

The research literature contains numerous examples of transfer designs similar to that used by Povenmire and Roscoe (1973), where groups differ on the amount of simulation training they receive and train to a common standard on the operational equipment. This design allows the analyst to determine the shape of the isoperformance function, and it restricts training on the operational equipment to just the level required to reach some performance standard. However, it does not allow the analyst to determine the precise effect that simulation training has on training with the operational equipment. To do so requires the construction of a transfer function, which is a type of learning curve where the effects of practice on the simulator are measured on the operational equipment, not on the device. Increases in simulation training could conceivably affect the transfer function by increasing the asymptote (the upper limit of learning), the learning rate (the speed at which the curve approaches the asymptote), or both. Each outcome has a unique implication for transfer effectiveness.

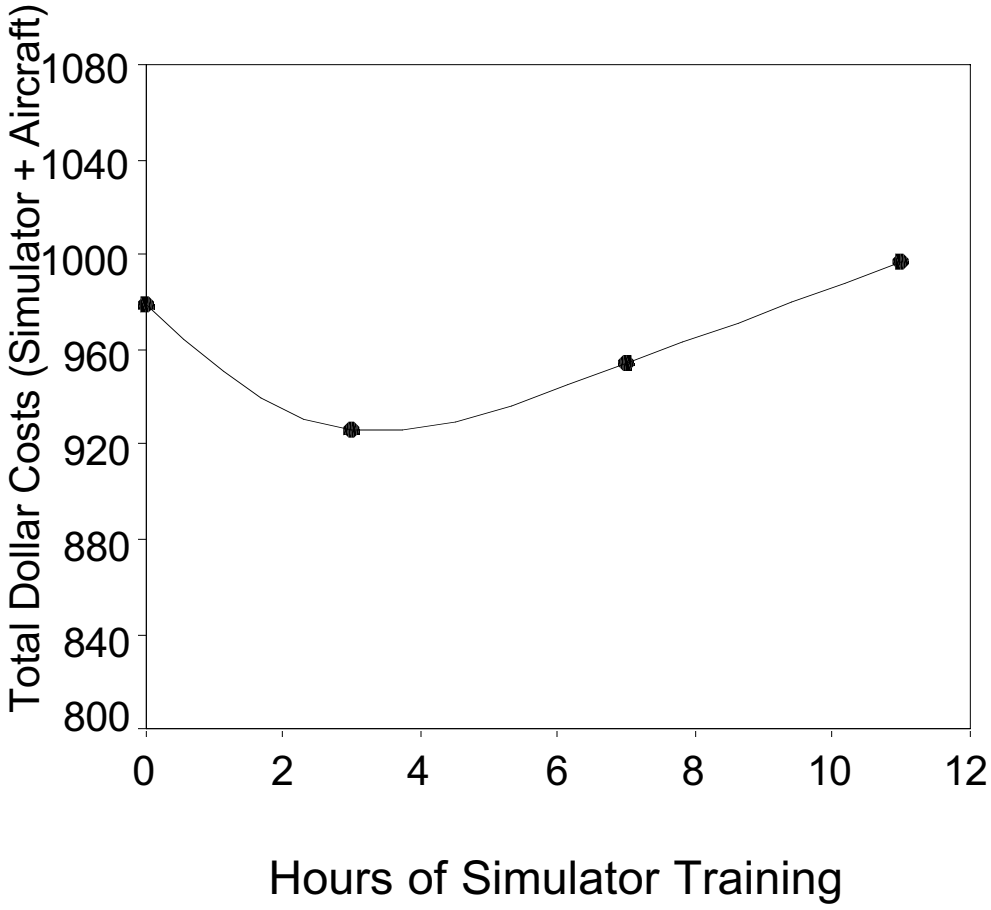

Figure IV-3. Isoperformance Curve for Total Training Costs (Source: Data reported in Povenmire and Roscoe, 1973)

To determine the shape of the transfer function, Hoffman and Morrison (1992) proposed a groups X trials design, which Table IV-2 depicts conceptually.[6] In this design, groups are differentiated by levels of training on the simulation, as in the Povenmire/ Roscoe design, but each group is then trained for a fixed number of trials on the operational equipment. If the simulation system has performance-measurement capabilities, the repeated measurements in the groups receiving high levels of simulation training can provide information about skill acquisition on the training device (Level II evaluation). Furthermore, if a reasonable performance standard can be set, the performance data from

---

6 The three trials depicted in Table IV-2 may not represent sufficient levels of training on the simulation. If so, the "trials" can be interpreted to mean groups of, say, 10 trials so that "1" represents Trials 1–10, "2" represents Trials 11–20, and "3" represents 21–30. Similarly, the number of trials on the operational equipment may be insufficient to show differences in the learning curve. If so, simply increase the number of trials for all groups.

Table IV-2. Outline of Groups X Trials Transfer Design

Treatment (Group)	Training on Simulation			Testing on Operational Equipment			
	Trial 1	Trial 2	Trial 3	Trial 1	Trial 2	Trial 3	Trial 4
G3	Yes	Yes	Yes	Yes	Yes	Yes	Yes
G2	Yes	Yes	No	Yes	Yes	Yes	Yes
G1	Yes	No	No	Yes	Yes	Yes	Yes
G0	No	No	No	Yes	Yes	Yes	Yes

the operational training phase of the evaluation can be rescored as trials or time to reach that standard. These data can be used to estimate isoperformance curves from which the user can derive useful parameters of transfer and cost effectiveness, which are Level III evaluation issues.

Although such complex transfer designs are potentially powerful ways to determine training effectiveness, Hoffman and Morrison (1992) also observed that they are not often used in research because they require high levels of research support. Given the scope of LSTSs, this issue becomes even more salient. As a result, transfer experiments are impractical approaches to evaluating LSTSs. Nevertheless, discussing this approach to evaluation is important for two reasons. First, discussing the pure transfer design is important so that evaluators fully appreciate the problems inherent in the approaches that exert less control over training. Second, in a few cases, the evaluator can exert extraordinary control over the training situation, and the transfer design can be used to good advantage.

For example, Bickley (1980) was able to control the amount of simulator training that students in the AH-1 transition-training course received before they received hands-on training in the actual helicopter. Bickley assigned students randomly assigned to one of four pre-specified levels of simulator training, including no training on the simulator. He then recorded the numbers of trials that the students required to reach proficiency, as defined by a passing score on a standard performance rating scale administered by instructor pilots. This particular study is regarded as an example of extraordinary cooperation between training providers (the U.S. Army Aviation Center) and the research organization performing the analysis (U.S. Army Research Institute Field Unit at Fort Rucker). Clearly, more cooperation would be needed to evaluate a large-scale simulation involving multiple Services.

2. Quasi-Experimental Designs

Although randomized design allows researchers to make inferences about the effects of training, the requirement for random assignment is often regarded as overly restrictive or impractical for evaluating LSTSs. Multiple observations and arbitrary scheduling become impossible because LSTSs require too many people and significant resources to implement. One alternative is to forego random assignment and instead seek intact groups that have appropriate levels of experience. By so doing, researchers can assess the effects of unusual experiences or high levels of training that could not be achieved in the context of a short-term experiment.

The systematic use of intact comparison groups in research provides an example of a "quasi-experimental" design, a term first used by Campbell and Stanley (1963). Although the term was new in 1963, the concept was not. In fact, the design depicted in Table IV-1 would be considered a "quasi-experiment" if the experimental and control conditions were formed from intact groups rather than from random assignment. The pretest/posttest design is particularly popular because the pretests allow the researcher to detect, and possibly control for, any differences that may exist between the groups before training.

The advantage of quasi-experimental designs is that they allow the researcher to assess the training effectiveness of LSTSs in a more realistic context than would otherwise be possible in the more constrained and highly controlled context demanded by randomized experiments. Because of the potential realism of quasi-experiments, they are recognized as possessing high levels of external validity, which is defined as the extent to which findings can be generalized to operational settings. At the same time, the lack of control in quasi-experiments implies that they are potentially more susceptible to threats from internal validity, which is defined as the extent to which changes in performance measures can be attributed to training variables. The possibility of preexisting differences in intact groups is a fundamental threat to internal validity. If differences exist before training, differences observed after training cannot be attributed to training. The pretest in the pretest/posttest control group design is the method by which this particular threat can be assessed and perhaps mitigated.

Several different quasi-experimental approaches have been described in the original classic exposition on the topic by Campbell and Stanley (1963) and the subsequent elaboration by Cook and Campbell (1979). An example of a study to assess the training effectiveness of SIMNET, one of the earliest LSTSs, illustrates how quasi-experimentation can be used to fit particular operational situations. Bessemer (1991) employed an interrupted

time series design to assess the transfer of SIMNET training to the field performance of students in the Armor Officer Basic (AOB) course at Fort Knox. He accomplished this by collecting detailed statistics on the performance of a field exercise administered at the end of AOB training. Performance data were obtained for 36 successive AOB classes, 24 of which occurred before the introduction to SIMNET and 12 of which occurred after the introduction to SIMNET. Linear trend analyses of the data indicated that introducing SIMNET caused small initial gains in field performance but that those gains increased steadily thereafter. The results were interpreted as being caused by a "learning-to-train" effect. In other words, instructors were initially ineffective in using the complex SIMNET technology to train tactical skills, but they improved during successive classes. These findings suggest that the "learning-to-train" effect may have resulted in serious underestimations of training devices from more conventional TEA designs.

3. Correlational Studies

Correlational studies of training effectiveness are the easiest approach to assessing training effectiveness because the researcher only has to observe passively or assess levels of training, rather than arbitrarily imposing levels of training on the research sample through random assignment or the scheduling of intact groups. In addition to being less intrusive, the results from correlational studies may be more generalizable because they reflect more "natural" variations in amounts and kinds of training.

At the same time, the results from correlational studies are difficult to interpret. The principal problem with correlational designs is the ambiguity in the relationship between independent variables (training) and dependent variables (performance measures). The essence of this problem is that the existence of covariation between independent and dependent variables in a correlational design does not imply that performance is somehow caused or controlled by the amount and kind of training. Suppose, for instance, that a researcher finds a positive correlation between the amount of LSTS training and performance. On the surface, this finding suggests that the LSTS is effective. However, the relationship between training and performance could be spurious. The empirical relationship could be caused by the fact that both variables covary with a third variable. This could happen if the LSTS were available to units that wanted to use it. In this case, amount of training might covary with the unit's motivation to seek out and capitalize on training experiences. It might be possible, then, that the observed positive relationship between the amount of training and proficiency might imply more about unit leadership and motivation than it does about LSTS effectiveness. On the other hand, if the researcher had randomly

assigned the amount of LSTS training to units, this variable would then have been uncorrelated with motivational differences and all other determinants of proficiency (measured or unmeasured). This contrived example illustrates the problems with correlational studies and the advantages of random assignment.

Researchers have tried to overcome the disadvantages of correlational designs by using advanced statistical analysis techniques. Of the proposed techniques, causal modeling by path analysis appears to be the most relevant to the analysis of training effectiveness. The concept is to estimate paths and path coefficients from correlations among variables. To do this, the researcher must first identify the major variables controlling and/or mediating the causal relationship in question (training → performance) and develop a structural diagram that indicates how the variables relate to one another. The resulting model would provide an estimate of the effect of LSTS training on performance, uncontaminated by the effects of other variables. Cook and Campbell (1979) have demonstrated, however, that omitting a relevant variable can distort the size of the path coefficient and even change the sign of the coefficient. Such an error would be disastrous to TEAs because it would mistake beneficial effects of a device for harmful ones and vice versa.

An example from Campshure and Drucker (1990) illustrates how the omission of a relevant variable changes the basic relationships between training and performance. They measured the progress of tank crew training on the Unit Conduct-of-Fire Trainer (U-COFT) and the performance on the live-fire, crew qualification exercise (Tank Table VIII). The initial data from two armor battalions (77 crews) indicated a complex but interpretable relationship between accomplishments on the training device and the live-fire performance. However, when they tried to replicate the findings on four additional battalions (136 crews), none of the relationships found in the initial data set were significant. An investigation of the two sets of data revealed an important difference. The battalions in the first set were under severe tank-mileage constraints that limited the amount of on-tank training they could conduct. In contrast, the battalions in the second set were under not under any tank-mileage constraints and therefore participated in virtually unlimited on-tank exercises. The researchers argued that the additional on-tank training received by the battalions in the second data set effectively reduced the relationship between U-COFT and Tank Table VIII. Unfortunately, these effects were not anticipated; therefore, measures of on-tank training were not obtained. These findings illustrate the need to identify all the determinants of performance before the start of data gathering and to devise operations for measuring each determinant.

Simpson (1999) identified several researchers who have suggested correlating LSTS usage data with historical or archival performance data from large-scale live exercises, such as those at the National Training Center (NTC), Twenty-Nine Palms, and Red Flag. The data can be arranged to mimic an experiment, as in an "ex post facto" design, but the approach remains essentially correlational because the independent variable (the quantity or quality of training) is measured, not controlled. This correlational approach to TEA has been suggested as an alternative to more conventional experimental and quasi-experimental approaches because the level of control and intervention required by the latter are practically impossible to implement at such venues. The correlational alternative approach would simply require that performance data on these units be passively and unobtrusively collected as units rotate through these training centers. To the extent that data from these exercises serve as surrogates for actual combat performance, the data provide valid measures of organizational outcomes in the military that are analogous to profit margins in for-profit organizations. Thus, this correlational approach appears to provide a method for performing a Level IV evaluation, the highest and most difficult analysis that assesses training programs in terms of their effects on organizational output.

Several objections have been raised to correlating training data and performance on large-scale live exercises. One objection relates to the assumption that performance on large-scale live exercises provides reliable and valid measures of combat proficiency. These training centers and participating units have long resisted—and for good reason—the standardization of training scenarios. Without such standardization, however, performance measures derived from such exercises cannot be reliable or valid. Also, whereas units use actual equipment to engage realistic OPFOR, aspects of the exercises must be simulated for obvious safety reasons. Whenever simulation is used, even in so-called "live" exercises, significant departures from fidelity can potentially reduce the validity of performance measures as indexes of combat performance.

In addition to concerns about the criterion measures of performance, this approach also has problems on the other side of the equation: the training predictors. Detailed information must be obtained on the kinds, amounts, and schedule of training conducted with the LSTS in question. In addition, information on other types of training that might possibly impact performance during the live exercise must also be obtained. Traditional measurement concerns also apply here because information on training may not necessarily be reliable and valid, especially if it is based on the memories of selected unit staff members.

Technical problems also complicate the analysis of correlational designs. The researcher must identify and obtain measures of all the important variables (not just measures of training) that potentially affect the relationship between LSTS training and field performance. Then, the analyst must specify, by path diagram or other convention, the likely relationships among the variables identified previously and the performance on the large-scale exercises. Identifying potential training correlations and determining a valid path diagram would comprise a substantial research effort in and of itself.

At first, the correlational approach would appear to provide an unobtrusive method for collecting LSTS system performance-based data that would otherwise be difficult, if not impossible, to obtain. After more serious consideration, however, what becomes clear is that this approach would likely have some serious conceptual limitations, consume greater-than-expected levels of research support, and require considerable technical expertise to execute and analyze. On the other hand, this approach is perhaps one of the few empirical approaches that allow us to address the question posed by the Level IV evaluation: What are the effects of training on battle outcomes?

4. Comparison of Design Approaches

We can summarize the previous discussion of empirical performance-based designs by comparing the three general approaches [randomized (true) experiments, quasi-experimental designs, and correlational studies] based on attributes that describe their inherent characteristics or requirements. In comparison to the original analysis of designs presented by Campbell and Stanley (1963), Table IV-3 contrasts broad categories of designs rather than contrasting individual designs within those categories. Although the resulting summary is not equivalent to Campbell and Stanley in form or detail, it is offered in the same spirit, which was to provide a method for summarizing the discussion and a general guide for the training evaluator.

Table IV-3 rates the three approaches on the degree to which they present problems for the evaluator. In general, the strengths of the experimental approach are the weaknesses of the correlational approach and vice versa. For each attribute, the quasi-experimental approach was only somewhat problematic, underlining the general applicability of this approach to the measurement of training effectiveness. The chief strengths of the experimental approach are that randomized designs do not require extraordinary statistical sophistication to analyze the results, do not require extensive information about training that

Table IV-3. Comparison of Empirical Research Design Approaches

	Characteristics				
Design Approaches	**Statistical Sophistication**	**Adjunct Training Information**	**Internal Validity**	**Control/ Intrusiveness**	**External Validity**
Randomized (True) Experiments	+	+	+	–	–
Quasi-Experiments	?	?	?	?	?
Correlational Studies	–	–	–	+	+

Note for Table IV-3. *Plus and minus signs indicate that the design presents few (+) or many (–) problems with regard to selected characteristics. The question mark (?) indicates that the design is somewhat problematic with respect to those characteristics.*

the participants had before the evaluation, and are relatively easy to interpret, given their internal validity. At the same time, experiments are more intrusive to implement and may be impossible for the types of large-scale evaluations required to assess LSTSs. Furthermore, the restrictions to the data collection limit the generality of findings (i.e., external validity). However, these latter two attributes make the correlational studies particularly relevant to LSTS evaluations. Theses studies are perhaps the only viable alternative in large-scale situations, and they potentially provide data that are relevant to real-world training situations. Furthermore, to the extent that correlational studies allow the measurement of training effects on complex phenomena (e.g., simulated battle outcomes), they most closely approach the goals of the ideal Level IV evaluations. However, the strengths of correlational designs must be tempered by their weaknesses. They are particularly susceptible to threats to internal validity, demand considerable information about training and other variables that impact performance, and require considerable technical sophistication to analyze the resulting data.

C. TRAINING ANALYSES

The previously described survey and performance-based designs may not be feasible or applicable as methods for determining the training effectiveness of LSTSs in certain situations. For example:

- Decision-makers need to estimate the effectiveness of a training system during the process of development. However, if a training system exists in concept only, it obviously cannot train tasks nor can users react to it.

- Some systems (particularly LSTSs) train relatively large numbers of people on an extensive set of tasks. In that regard, employing performance-based designs and even training surveys to evaluate such systems may be too costly.
- Some training systems may be designed to train tasks and objectives (e.g., emergency procedures and security-related situations) that cannot be readily or safely evaluated using traditional methods.

In such cases, an alternative approach is to use some form of analytic model for estimating or forecasting training system effectiveness. A training analysis model is a method or procedure for estimating the training effectiveness of training systems. The key characteristic that distinguishes training analyses from the previously described approaches is reliance of training analyses on nonempirical data (i.e., data other than those derived from actual performance or user input).

1. State of the Art

A large number of training analysis models have been developed for estimating the effectiveness of training devices and simulations. Muckler and Finley (1994) identified 35 formal models, 21 of which are capable of providing quantitative predictions of transfer effectiveness (i.e., outcomes from transfer experiments). Simpson (1995) later examined several analytic models and identified over 50. The models are based on a variety of methods, but they share the common central assumption "… that a training system can be evaluated in terms of its elements and that the benefits of a training system can be understood by studying those elements" (Pfeiffer and Horey, 1988, p. 7).

Muckler and Finley (1994) identified strengths or advantages of these training system estimation models, including the following:

- To the extent that training estimation models require precise statements of training objectives, they have ensured that training development is focused properly on relevant training goals.
- The submodels for media selection have provided effective procedures for choosing appropriate media within cost constraints.
- The models promote systematic thinking during training development and provide the means to consider alternative methods for delivering training and for assessing the consequences of these methods on performance.
- The models have provided quantitative indexes of effectiveness that equal or exceed the sophistication of most models in the domain of human resource management.

Despite these advantages, Muckler and Finley (1994) indicate that few models have been institutionalized or have even survived the early stages of research. Part of the problem can be traced to the lack of familiarity with and general resistance to technology-based solutions. However, some specific weaknesses or disadvantages of these models have also caused the lack of acceptance. Some of the more serious problems can be briefly summarized as follows:

- **The models are too complex and too costly to run and maintain.** Even the earliest models required some form of computer support to execute. Muckler and Finley estimate that such systems would typically require $100K yearly to set up and maintain.
- **The models require extensive training efficiency, effectiveness, and cost data.** Some of these data can only be obtained through empirical research, thereby defeating the purpose of the models.
- **As presently configured, the models require weeks to set up and process data.** This response time is too slow to support the design and development of training systems.

In addition to these formal models, Simpson (1999) identified several training analysis methods or strategies frequently employed to evaluate training. Although these methods are analytic in nature, they are often combined with other approaches to TEAs, such as survey or performance-based methods. Simpson differentiated among the following five different analytic methods:

1. **Modeling.** Modeling is a method common to most forms of TEAs. In the context of TEAs, it refers to the analyst's attempts to represent accurately and economically the key characteristics of the training system. Examples of models range from simple lists of system components to abstract flow charts of system functions. Simpson (1999) stressed that modeling is an especially important method for evaluating a system under development.
2. **Analogy.** Analogy is the method by which the analyst infers the effectiveness of a training system by examining the effectiveness of a similar system. Analogy is the central method underlying the formal comparison-based prediction (CBP) model described later in this section.
3. **Extrapolation.** Simpson (1999) defined extrapolation as "… prediction based on some understanding about how a process works" (p. 52). Extrapolation can be based on a theory. For instance, a theory could specify what learning processes or training features are required to produce certain training outcomes. To use this method, a training system would then be evaluated for those processes or training features.

4. **Task list analysis.** Devices can be evaluated by the tasks that they train, and a list of tasks that defines the domain of interest is prepared. For instance, LSTSs can be evaluated against a list of tasks that define a particular area of interest, such as command and control (C2). The training systems are then systematically operated to determine the extent to which tasks can be realistically performed. This method assumes that the system actually exists, at least in some prototype form.
5. **Historical data.** Historical data from published and unpublished sources can be compiled to evaluate LSTSs. Simpson (1999) summarized results from 10 studies that used historical data to evaluate devices. He characterized these studies as "more narrowly focused, involve less data, and cover a shorter time frame than the typical academic review or meta-analysis" (p. 53).

2. Examples of Formal Models

A surprisingly large number of training analysis models have been developed over the last 20 years. Refer to the following for comprehensive reviews of this literature:

- **Goldberg and Khattri (1987).** This review of the literature laid the groundwork for the development of the Training Effectiveness and Cost Iterative Technique (TECIT) model for evaluating the cost effectiveness of a training device or simulator.
- **Knerr, Nadler, and Dowell (1984).** The authors compared extant methods for predicting training device effectiveness on five dimensions: objective, units of analysis, components, metrics, and level of development.
- **Muckler and Finley (1994).** These authors reviewed 36 "training system estimation models" to determine strengths and deficiencies of each.
- **Pfeiffer and Horey (1988).** The authors categorized 18 methods and found that training device effectiveness belonged to one of four underlying techniques: index techniques, magnitude techniques, proximity techniques, and interlocking techniques.
- **Rosen, Berger, and Matlick (1985).** The authors provided a general review of the literature as part of an effort to develop methods for determining cost and training effectiveness that can be performed manual or on a hand calculator.
- **Simpson (1995).** The author reviewed methods related to the cost effectiveness analysis of training to assess the status of technology and to develop a general conceptual model.

- **Sticha et al. (1988).** The authors reviewed the analytical procedures, psychological theory, and empirical findings that related to the development of the Optimization of Simulation-Based Training Systems (OSBATS), a model for performing the tradeoff analyses required to design training devices and simulators.
- **Tufano and Evans (1982).** The authors examined four models to provide a basis for selecting or refining one or more for incorporation into a defined set of procedures for specifying the effectiveness of Army training devices or simulators.

Instead of providing a comprehensive review of the literature, the following discussion provides three examples specifically chosen to illustrate the range of approaches that have been used to analyze training effectiveness:

1. Multi-attribute utility measurement (MAUM)
2. Comparison-based predictions
3. Simulated transfer.

a. Multi-Attribute Utility Measurement (MAUM)

Several training analysis models use methods derived from the MAUM concept, including the CTEA model described by Dawdy, Chapman, and Frederickson (1981), the OSBATS model described by Sticha et al. (1988), and the TECIT model developed by Goldberg (1988). For the most part, training analysis models use a straightforward MAUM model, such as SMART (Simple Multi-Attribute Rating Technique) developed by Edwards (1977). For evaluating training simulations, SMART is basically a four-step process:

1. Determine the criteria against which simulations will be evaluated (dimensions of value)
2. Determine the importance of each criterion
3. Rate the simulations relative to these criteria
4. Aggregate the ratings across criteria to determine the overall utility of the simulations.

To illustrate these four steps, which are described in more detail below, we use data from Pfeiffer and Siegel (1974) summarized in Table IV-4. In this study, the researchers evaluated the utility of two different simulation methods: human interactive vs. Monte Carlo simulation.

Table IV-4. Calculated Utilities for Two Simulation Models (Source: Pfeiffer and Siegel, 1974)

Dimension of Worth	w_j	Human Interactive Simulation	Monte Carlo Simulation
		u_{ij}	u_{ij}
Construct validity	.18	50	28
Repeatability of output	.16	81	100
Degree of error/low variability	.14	29	50
Feasibility of use and application	.13	62	92
Content validity/real-world detail	.12	83	58
Low confounding of variables	.10	30	50
Sensitivity/input-output	.08	63	75
Parsimony/low number of assumptions	.04	100	50
Generality/flexibility of application	.03	68	33
Modifiability	.02	50	100
U_i = aggregate utility = $\Sigma\, w_j\, u_{ij}$		59	63

1. **Determine the criteria against which simulations will be evaluated (dimensions of value).** In consultation with SMEs, the analyst determines a comprehensive set of criteria to evaluate the simulation. These criteria (dimension of value) should describe the goals of the simulation and should be determined independently of the characteristics of the particular simulation(s) being evaluated. These dimensions may comprise a list of abstract properties that the training simulation should have (the approach embodied in the example, Table IV-4) or a set of tasks or skills that the simulation should train. Regardless, the number of dimensions should be relatively small: at least 8 but no more than 15. The normalized importance rating on the j^{th} dimension is listed in the column headed w_j in Table IV-4.

2. **Determine the importance of each criteria.** SMEs then rank the dimensions of value in order of importance. Once the rank order is established, the SMEs rate the dimensions so that the ratios among ratings are preserved. This can be accomplished by arbitrarily assigning the most important dimension a rating of 100 and then rating other dimensions against that standard, being

careful to preserve the ratios. That is, a dimension that is rated 50 should be half as important as the highest rated dimension but 5 times more important than a dimension rated 10. Once the ratings are agreed upon, each is divided by the sum of all ratings, converting the ratings into probability-like values. The utility of the two simulations on dimension *j* is listed in the column headed u_{ij} in Table IV-4.

3. **Rate the simulations relative to these criteria.** Edwards (1977) asserts that he uses the term "rate" loosely to refer to subjective and objective measurement. For instance, if dimensions refer to tasks, the measure might be a relatively objective count of subtasks that can be trained by the simulation. Alternatively, the simulation might be subjectively rated for ease of use. In either case, the responses must be scaled so that the maximum plausible value (i.e., not necessarily values achieved with the simulations under consideration) is equal to 100, whereas the minimum plausible value is equal to 0. Once those points are established for the actual measured scale, intermediate points are determined by linear interpolation, although more complex curvilinear utility functions are sometimes used.

4. **Aggregate the ratings across criteria to determine the overall utility of the simulations.** Aggregate utility of the i^{th} training system over *j* dimensions is calculated from the following formula:

$$U_i = \Sigma\ w_j\, u_{ij} \quad ,$$

which is no more than the formula for a weighted average. This number can be interpreted as a percentage, varying from 0 to 100.

Assuming that the evaluation dimensions are valid components of training effectiveness, the resulting utility measure, U_i, can be interpreted as an overall index of training effectiveness. However, it should not be interpreted as reflecting a specific empirical measure, such as the transfer effectiveness ratio (TER). This measure is most meaningful when compared with alternative training systems or versions of the same systems. In that regard, the example results from Pfeiffer and Siegel (1974) are instructive. They show a slight advantage for the Monte Carlo simulation over the human interactive model. However, the results on the individual dimensions were quite mixed. The Monte Carlo simulation was superior on six of the evaluative dimensions, whereas the human interactive simulation was superior on the remaining four dimensions.

b. Comparison-Based Predictions (CPBs)

Gary Klein and Associates (e.g., Klein et al., 1985) developed the CBP method of evaluating training effectiveness. The CPB method is a formalized procedure for reasoning

by analogy. It estimates the training effectiveness of proposed training systems by comparing these systems to similar systems for which training effectiveness data exist. The CBP developers frequently explain their method by referring to the approach used by real estate agents to determine the proposed asking price of a house just put on the market. Agents or appraisers generate this figure by systematically comparing the target (i.e., the to-be-appraised) home to similar homes that have recently been sold in the neighborhood. They use the selling prices of the recently sold homes and adjust the proposed selling price of the target house upwards or downwards based on key factors known to drive real estate prices (e.g., square footage, the number of bathrooms, and the location).

According to the CBP guidebook for determining the cost and effectiveness of training devices (Klein et al., 1985), the method is divided into four phases, which are further subdivided into a number of individual steps:

- **Phase I. Set up the problem.** During this initial phase, the analyst decides what he or she wants to predict.
  - Describe the training system under development, which is designated as A.
  - Specify the target measure of effectiveness, which is designated T(A) and is the quantitative index that the analyst intends to predict.
  - Identify the major causal factors (CF) for T(A) that are affected by A. The final list of factors should be short (5 to 7 factors) and should include only those that account for the most variance in T(A) (i.e., only the "high drivers").
  - Determine a context, or scenario, for the prediction—specifying the conditions under which A operates and how T(A) will be measured.
- **Phase II. Select resources.** After defining the problem, the analyst selects the cases, methods, and experts that he or she will use to make the final prediction.
  - Identify comparison training system(s), which are designated as $B_{1...n}$.
  - Choose appropriate SMEs to make the judgments.
- **Phase III. Collect and analyze data.** In consultation with the selected SME, the analyst makes the prediction.
  - Determine the value for the measure from the comparison training system, which is denoted T(B).

- Using the high drivers, determine the differences between systems A and B. Estimate the effects of the differences on T(B).
- Adjust T(B) so that it provides a valid estimate of T(A).

- **Phase IV. Document the process.** In the final step, the analyst writes up the methods and results from the process.

Table IV-5 summarizes the results from an example in the CBP guidebook (Klein et al., 1985) in which the authors proposed to estimate the effectiveness of a device used to train crews on a new howitzer. The measure chosen was the performance on the final round of a howitzer-crew qualification test. Two comparison cases were identified: a tank gunnery trainer that has similar features to the proposed howitzer trainer and an older howitzer panel gunnery trainer that is currently being used for howitzer instruction. Accordingly, the analysts chose two SMEs: one with experience on the tank gunnery trainer (SME_1) and one with experience on the howitzer panel gunnery trainer (SME_2). The SMEs provided performance data that results from training on their respective devices.

SME_1 had actual performance that he maintained. SME_2 did not and, consequently, had to estimate performance. They compared their respective devices with the proposed device and determined that the key difference was in the performance data recording system. Whereas SME_1 regarded the key feature as an asset and determined that tank gunnery performance would improve about 5 percent, SME_2 thought that the same feature was a detriment and would decrease performance about 5 percent. The resulting adjusted performance values (73.5 percent and 71.25 percent) became the initial estimates of T(A). However, instead of taking the simple average of these two values (72.4 percent), the analysts biased the estimate toward SME_1 because the initial estimate of SME_1 was based on actual data and was therefore considered more valid and reliable. The resulting estimate was 73 percent. However, to ensure that the final estimate includes the initial values, the analysts added a range of ±2 percent to the estimate to provide subjective confidence limits to the data.

One of the advantages of the CBP method is that the estimates are based on actual data rather than on theoretical models. Another advantage is that the method requires relative—as opposed to absolute—judgments from SMEs, and these kinds of judgments are generally more valid and easier for humans to make. Also, the CPB outcomes are highly interpretable. Finally, compared with most training analytic models, CBP is simple and requires a relatively small amount of data.

Table IV-5. Example Results of a Comparison-Based Prediction of Training Effectiveness

Phase	CBP Element	Symbol	Result
I. Set Up the Problem	Target case	A	To-be-developed training device for a new howitzer
	Target value measure	T(A)	Average number of hits on final round
	Causal factors	CF	Physical fidelity Feedback potential Performance data recording
	Context	–	Howitzer crew qualification testing
II. Select Resources	Comparison cases	B_1 B_2	Tank gunnery trainer Howitzer panel gunnery trainer
	SMEs	SME_1 SME_2	Tank gunnery training supervisor Howitzer panel gunnery training supervisor
III. Collect and Analyze Data	Comparison target values	$T(B_1)$ $T(B_2)$	70% final test round hits (based on performance data) 75% final test round hits (based on SME estimates)
	Adjustments to T(B)	–	SME_1: +5% for better data recording SME_2: –5% for deterrence to drill and practice
	Initial estimates of target value	T(A)	SME_1: 73.50% SME_2: 71.25%
	Final	T(A)	73%±2%
IV. Document the Process	–	–	1- to 2-page summary of methods and results

On the other hand, the kinds of data required (i.e., actual training effectiveness data from analogous systems) could potentially be difficult to find. The authors of the CBP maintained that if such data were not available, SMEs could estimate values for both T(B) and T(A). However, once the analyst is forced to rely solely on estimations, the advantages of the CBP are not as clear.

c. Simulated Transfer

Pfeiffer and Horey (1988) described a third approach for estimating training effectiveness. They referred to this approach as "simulated transfer." In this remarkably straightforward approach, SMEs are carefully selected for their familiarity with the effects of and performance on the to-be-evaluated simulation system and the corresponding operational system. They then use their knowledge to predict the outcomes of transfer experiments. In the example developed by Hagin et al. (1982), SMEs were asked to estimate the number of trials required to reach proficiency on a particular aviation maneuver, first in the aircraft and then in the simulator under evaluation. Then, they were asked to estimate the number of aircraft trials that would be required to reach proficiency *after* an aircrew had received simulation training to reach this proficiency. From these data, analysts could calculate typical TEA measures (e.g., percent transfer savings and CTER).

Hagin et al. (1982) argued that the simulated transfer method possesses two types of distinct advantages over standard rating scales. First, it provides a structured and easily understood format to which SMEs can respond. Second, the resulting data provide a direct, quantitative estimate of training effectiveness that is commensurate with the results from actual learning and transfer experiments. Hoffman and Morrison (1992) also recognized that simulated transfer methods could be easily modified to estimate results from complex transfer experiment, particularly ones that would be difficult to implement.

Hoffman and Morrison (1992) examined the validity of simulated transfer by asking armor gunnery trainers to estimate of average performance of armor crews after various fixed amounts of practice using live-fire techniques on the tank, laser-simulated techniques on the tank, and a computer-based simulation. The results indicated gross differences among training media that were congruent with expectations. However, the details were disappointing because the estimated learning curves were nearly linear and did not show the characteristic negative acceleration shown in actual curves. This latter result brought into doubt the ability to use estimated data to make precise quantitative tradeoffs among training methods.

V. RECOMMENDATIONS

The previous sections discussed a number of methods that can be used to measure training effectiveness. The choice depends on many considerations, including the goals of the evaluation, the training objectives of the system under evaluation, and the system's stage of development. Furthermore, guidance concerning TEA is not highly codified, so most of the decisions still depend on the professional judgment of evaluation proponents and technical specialists. Despite the vagaries of TEA methodologies, several general recommendations for measuring the effectiveness of LSTS systems can be inferred.

At the most general level, evaluators should document a plan for measuring training effectiveness. A written plan provides the evaluator an opportunity to think through the process and allows potential users a chance to comment on the potential utility of the training effectiveness measure(s). More specific recommendations, as described below, can be included in this plan.

A. IDENTIFY SPECIFIC MEASUREMENT ISSUES

The analyst must decide among several approaches for measuring the training effectiveness of an LSTS. The first step in this process is to determine and document the exact issues that should be addressed. The following provides some example issues that directly relate to the measurement of training effectiveness:

1. Determine the Types of Decisions That the Measure Is Intended To Inform

For instance, the measure may be used to support decisions related to the design of LSTSs under development, the procurement of a prototype that has been initially assessed, or the design of a strategy for employing a procured LSTS. If the LSTS has been deployed, the evaluator can consider using empirical performance measures and a transfer design to determine training effectiveness. On the other hand, if the LSTS is under development, the empirical approach is inappropriate, and the evaluator must consider using one of the training analysis models described previously.

2. Identify the Audience for Whom the Measure Is Intended

The audience could include scientific/technical personnel, unit and training manager, or training system users. Clearly, a more technical audience would expect a precise quantitative answer. The evaluator must be prepared, however, to address a diverse audience. The evaluator must seek the most precise quantitative answer possible for scientific/technical audiences but then translate that answer to more operational terms for training managers/users.

3. Specify Whether Absolute or Relative Effectiveness Measures Are Required

Training effectiveness measures can be stated in absolute or relative terms. Absolute measures may be required to justify expenditures or to allocate training to ensure specific performance levels. For instance, the Operational Requirements Document (ORD) may specify that an LSTS produce acceptable performance within N hours of use or $X of usage costs. In contrast, relative measures of effectiveness are required when two or more training alternatives are compared. This approach is required, for example, when one wishes to compare the results from training on an LSTS to the results from training with other devices or media that train the same tasks.

B. DEVISE A MEASUREMENT PLAN THAT SPECIFIES PERFORMANCE MEASUREMENT AND RESEARCH DESIGN

From the issues, the analyst should devise a plan for measuring the training effectiveness of the LSTS in question. An essential aspect of this plan is the specification of the performance measures and the design for collecting these measures. These two aspects of measuring training effectiveness are clearly interrelated.

Too often, TEA plans specify either detailed lists of performance measures or the precise experimental design. The measurement plan should list the exact performance measures that will be collected, and it should also indicate the conditions under which these measures will be collected.

C. USE VALID AND RELIABLE PERFORMANCE MEASURES

TEAs must be based on valid and reliable performance measures, and, to the extent possible, analysts should employ measures that have known validity and reliability.

However, many performance measures used in TEAs are developed for that particular analysis, and, consequently, the reliability and validity of such measures are not known.

In the case of LSTSs, the UJTL provides several candidate performance measures. Although these measures are relatively well defined, their reliability and validity are not known. When reliability and validity characteristics of measures are not known, evaluators should design the TEA so that it provides the opportunity to measure reliability and validity, particularly for those measures that they suspect are problematic.

Suppose, for example, that evaluators use SMEs to develop a task checklist designed to assess the performance of an LSTS. They could measure the content validity of the checklist by having a different set of SMEs verify items in the checklist. Once the content of the checklist is validated, the evaluators develop detailed instructions for using the list to assess behavior. To assess the reliability of this performance measurement system, the evaluator should arrange for two or more observers to evaluate some, if not all, of the same performers independently. Then, the inter-rater reliability of the observations can be calculated and reported as part of the evaluation results. Finally, the empirical validity of the observations can be assessed by correlating the scores on processes and the task outcomes, such as weapon system effects or the ratio of enemy-to-friendly losses.

D. OBTAIN THE MOST VALID MEASURE OF PROFICIENCY POSSIBLE WITHIN THE CONSTRAINTS OF THE EVALUATION

The most relevant data for measuring training effectiveness is job or task proficiency. The analyst's goal is to seek the most valid measure of this conceptual criterion. In many situations, the most valid measure of job proficiency is on-the-job performance using the weapons systems and related equipment as they would actually be used in combat situations. At the same time, several potential problems can rule out performance measurement on actual equipment:

- It may be too costly in terms of manpower, equipment, technology, and other resources. Given the sheer scope of LSTSs, testing their effectiveness in a field exercise would be prohibitively expensive.
- It may involve tasks that are inherently unsafe or damaging to equipment and cannot be performed routinely. Examples include procedures for operating damaged equipment.
- In some cases, performance processes and outcomes on actual equipment cannot be observed reliably. For instance, detailed observations of tank crews are difficult because the crew is often unobservable.

If performance on the actual equipment is infeasible, the next best approach is to seek a surrogate measure (e.g., a virtual simulation). Most often, analysts consider performance on a simulation as a surrogate for performance on the actual equipment. In this sort of transfer experiment, two different configurations of simulators are used: a "test configuration" that embodies the key characteristics of the simulator under consideration and a "criterion version" that closely corresponds to the actual equipment. The basic assumption is that transfer to simulators can predict transfer to performance on the actual equipment (Taylor, Lintern, and Koonce, 1993). This approach would be most useful for evaluating LSTSs when the evaluator has a prototype version of the simulation that can be reconfigured to test alternative simulation approaches for the production version.

In the early development of LSTSs, when important decisions must be made, obtaining performance data from primary or surrogate measures is not possible. In this case, the analyst should seek the best input from an SME using some systematic model for estimating training effectiveness (see Section IV). Selecting appropriate SMEs who have the relevant experience and do not have vested interests in the outcome of the analysis is probably more important than choosing the correct analytic model. Whenever possible, analysts should use more than one SME, allow the SMEs to make their judgments independently, and measure their agreement.

E. TO THE EXTENT POSSIBLE AND REASONABLE, IMPOSE EXPERIMENTAL CONTROL ON THE RESEARCH SITUATION

In general, analysts should always opt for the approach that allows the greatest degree of control over the experimental situation. Thus, a randomized experiment is generally preferred to a quasi-experiment, and a quasi-experiment is generally preferred to a correlational study. Greater control allows stronger inferences about the relationships between causes and effects. Also, if properly designed, randomized experimental designs require less extensive and sophisticated statistical analyses than quasi-experiments require. Similarly, quasi-experiments are less difficult to analyze than correlational studies.

At the same time, the preference for control over the situation should be balanced with concerns for realism. For example, the analyst should not choose levels of training that cannot be duplicated in the operational context. The essential consideration is whether the conditions being tested fairly represent—or at least do not distort—the conditions under which the system will actually be used. For instance, the amount of training on the training system that can be accomplished during a typical short-term learning or transfer experiment

could seriously underrepresent the amount of training time that could be achieved once the system is fielded. This qualification to the general recommendation appears particularly appropriate for LSTS TEAs. A single short-term experiment is probably inappropriate to evaluate an LSTS because of the scope and diversity of its training objectives. The alternatives are to evaluate the LSTS through a series of limited-objective experiments or use quasi-experiments or correlational studies to determine the overall effects of the LSTS once it has been fielded and is in regular use.

F. USE ANALYTIC MODELS IN THE EARLY STAGES AND CONTINUE USING THESE MODELS THROUGHOUT SYSTEM DEVELOPMENT

LSTSs can be expensive to develop. To support the early design and resourcing decisions, the evaluator must be able to measure the potential payoffs from the systems during the concept development phase. The only approach for measuring the training effectiveness of devices under development is to use analytic models of system capabilities. As the system is being developed, analysts should seek an appropriate model for estimating training effectiveness. The development and use of such a model can be costly in time and other resources; however, if the model is maintained and updated throughout all stages of development, the use and further development of the model should become more cost effective.

Although the evaluator should be encouraged to use and reuse the same training analysis model throughout development, the model should be updated and revised periodically to reflect any performance-based results from later evaluations. The point of the updates is not to confirm what analysts already know; rather, the idea is to perform excursions from the known cases and test conditions and circumstances that are not known. This approach is particularly appropriate for evaluations of LSTSs, where the costs of updating or improving the systems are likely to be substantial.

Training analysis models can also be used to test the effectiveness of alternative improvements to the LSTS that have been proposed but not implemented. By so doing, the evaluator would have the benefit of testing the alternatives without incurring the costs of gathering empirical performance data.

G. INCLUDE USER REACTIONS AS AN ADJUNCT TO PERFORMANCE DATA AND ANALYSIS

Including user reactions as TEA data is sometimes described in disparaging terms, such as those in Kirkpatrick's schema where these reactions are relegated to the lowest level of analysis. However, most analysts acknowledge the importance of user reactions by agreeing that a training device cannot be effective if trainers and trainees dislike it and refuse to use it.

User reactions are an easy measure to obtain. The only requirement is that users have sufficient experience with the system to form some stable impressions. That said, user reactions provide the weakest evidence for training effectiveness and should not be used as sole or even primary data for the TEA of an LSTS. Rather, these judgment-based measures should be used to supplement performance- or analytic-based data.

REFERENCES

Alluisi, E.A. (1991). The development of technology for collective training: SIMNET, a case history. *Human Factors*, *33*, 343-362.

Babbitt, B.A., and Nystrom, C.O. (1989a). *Questionnaire construction manual* (ARI Research Product 89-20). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Babbitt, B.W., and Nystrom, C.O. (1989b). *Questionnaire construction manual annex. Questionnaires: Literature survey and bibliography* (ARI Research Product 89-21). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

BDM Federal, Inc. (1995, June). *Assessment Package*. Monterey, CA: Author.

Bessemer, D.W. (1991). *Transfer of SIMNET training in the Armor Officer Basic course* (ARI Technical Report 920). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Bickley, W.R. (1980). *Training device effectiveness: Formulation and evaluation of a methodology* (ARI Research Report 1291). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Boldovici, J.A. (1987). Measuring transfer in military settings. In S. Cormier and J. Hagman (Eds.), *Transfer of learning*. New York: Academic Press.

Branson, R.K., Raynor, G.T., Cox, J.L., Furman, J.P., King, F.J., and Hannum, W.H. (1975). *Interservice procedures for instructional systems development* (AD A019 486). Tallahassee, FL: The Center for Educational Technology, Florida State University.

Campbell, D.T., and Stanley, J.C. (1963). Experimental and quasi-experimental designs for research. In N. L. Gage (Ed.), *Handbook of research on teaching*. Boston: Houghton Mifflin Company. (Also published as *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally, 1966.)

Campshure, D.A., and Drucker, E.H. (1990). *Predicting first-run gunnery performance on Tank Table VIII* (ARI Research Report 1571). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Cook, T.D., and Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin Company.

Dawdy, E.D., Chapman, W.A., and Frederickson, E.W. (1981). *REMBASS Final report: Vol. 1. Executive summary and Vol. 2. Final report* (Contract No. DABT-60-80-C-0056). Ft. Huachuca, AZ: Applied Science Associates.

Delbecq, A.I., Van de Ven, A.H., and Gustafson, D.H. (1975). *Group techniques for program planning: a guide to nominal group and delphi processes*. Glenview, IL: Scott Foresman.

Department of Defense (1996, August). *Training simulators, simulations, and devices* (DoD Instruction 1430.13). Washington, DC: Author.

Department of Defense (1999, October). *Training simulators, simulations, and devices* (DoD Draft Directive 1430.13). Washington, DC: Author.

Dynamics Research Corporation (1996, March). *Guidelines for the development of tasks, conditions, and measures to support the joint or service training systems*. Andover, MA: Author. Retrieved February 22, 1999 from the World Wide Web: http://www.drc.com/FAST/guidtask.doc.

Edwards, W. (1977). How to use multiattribute utility measurement for social decision-making. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-7*, 326–340.

Fletcher, J.D. (1988). *Responses of the 1/10 cavalry to SIMNET*. IDA Analysis Memorandum No. M-494. Arlington, VA: Defense Sciences Office. (ADA 200499)

Fletcher, J.D., and Chatelier, P.R. (2000). Training in the military. In S. Tobias and J.D. Fletcher (Eds.), *Training and retraining: A handbook for business, industry, government, and the military*. New York: Macmillan.

Fowlkes, J.E., Lane, N.E., Salas, E., Franz, T., and Oser, R. (1994). Improving the measurement of team performance: The TARGETs methodology. *Military Psychology*, *6*, 47–61.

Goldberg, I. (1988). Training effectiveness and cost iterative technique (TECIT). Vol. I: Training effectiveness analysis (Research Note 88-35). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Goldberg, I., and Khattri, N. (1987). *A review of models of cost and training effectiveness analysis* (ARI Research Note 87-58). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Hagin, W.V., Osborne, S.R., Hockenberger, R.L., Smith, T.H., and Gray, T.H. (1982). *Operational test and evaluation handbook for aircrew training devices: Operational effectiveness evaluation* (AFHRL-TR-81-44(II)). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

Hiller, J. H. (1987). Deriving useful lessons from combat situations. *Defense Management Journal*, *2nd & 3rd Quarter*, 29-33.

Hoffman, R.G., Graves, C.R., Koger, M.E., Flynn, M.R., and Sever, R.S. (1995). *Developing the reserve component virtual training program: History and lessons learned* (ARI Research Report 1675). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Hoffman, R.G., and Morrison, J.E. (1992). *Methods for determining resource and proficiency tradeoffs among alternative tank gunnery training methods* (ARI Research Product 92-03). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Holding, D.H. (1991). Transfer of training. In J. Morrison (Ed.), *Training for performance: Principles of applied human learning*. Chichester: John Wiley and Sons.

Joint Advanced Warfighting Program (2000, June). *The Joint Experiment J9901: Attack operations against critical mobile targets*. Alexandria, VA: Institute for Defense Analyses.

Jones, M.B., and Kennedy, R.S. (1996). Isoperformance curves in applied psychology. *Human Factors*, *38*, 167–182.

Jones, M.B., Kennedy, R.S., and Bittner, A.C., Jr. (1981). A video game for performance testing. *American Journal of Psychology, 94,* 143–152.

Kirkpatrick, D.L. (1976). Evaluation of training. In R. L. Craig (Ed.), *Training and development handbook: A guide to human resource development*. New York: McGraw-Hill.

Klein, G.A., Johns, P., Perez, R., and Mirabella, A. (1985). *Comparison-based prediction of cost and effectiveness of training devices: A guidebook* (ARI Research Product 85-29). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Knerr, C.M., Nadler, L.B., and Dowell, S.K. (1984). *Training transfer and effectiveness models* (HumRRO FR-TRD(VA)-84-1). Alexandria, VA: Human Resources Research Organization.

Krueger, A.R. (1994). *Focus groups: A practical guide for applied research*. Thousand Oaks, CA: Sage.

Lane, N. E. (1986). *Issues in performance measurement for military aviation with applications to air combat maneuvering* (Final Report DAAG29-81-D-0100, EOTR86-37). Orlando, FL: Essex Corporation.

Lickteig, C.W. (1996). *Research methods for advanced warfighting experiments* (ARI Technical Report 1047). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Muckler, F.A., and Finley, D.L. (1994). *Applying training system estimation models to Army training, Volume 1: Analysis of the literature and Volume 2: An annotated bibliography* (ARL-TR-463). Aberdeen Proving Grounds, MD: U.S. Army Research Laboratory.

Newell, A., and Rosenbloom, P. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition*. Hillsdale, NJ: Erlbaum.

Office of the Inspector General, DoD (1997, April 30). *Requirements planning and impact on readiness of training simulators and devices* (Report No. 97-138). Washington, DC: Author.

Orlansky, J., Taylor, H.L., Levine, D.B., and Honig, J.G. (1997, March). *The cost and effectiveness of the Multi-Service Distributed Training Testbed (MDT2) for training close air support* (IDA Paper P-3284). Alexandria, VA: Institute for Defense Analyses.

Pfeiffer, M.G., and Horey, J.D. (1988). *Analytical approaches to forecasting and evaluating training effectiveness* (Technical Report 88-027). Orlando, FL: Naval Training Systems Center.

Pfeiffer, M.G., and Siegel, A.I. (1974). Model development and the assessment of competing models. *Proceedings of the Eighteenth Annual Meeting of the Human Factors Society*. 425-428.

Povenmire, H.K., and Roscoe, S.N. (1973). Incremental transfer effectiveness of a ground-based general aviation trainer. *Human Factors, 15,* 534–542.

Roscoe, S.N. (1971). Incremental transfer effectiveness. *Human Factors, 13,* 561–567.

Rosen, M.H., Berger, D.C., and Matlick, R. K. (1985). *A review of models for cost and training effectiveness analysis* (ARI Research Note 85-34). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Simpson, H. (1995, June). *Cost-effectiveness analysis of training in the Department of Defense* (Technical Report 95-004). Seaside, CA: Defense Manpower Data Center.

Simpson, H. (1999). *Evaluating large-scale training simulations. Volume 1: Reference manual* (DMDC Technical Report 99-05). Seaside, CA: Defense Manpower Data Center.

Smith, M.D., and Hagman, J.D. (1998). Enhancing the resource efficiency of live-fire tank gunnery evaluation. (ARI Technical Report 1088). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Space and Naval Warfare Systems Center (2000). *Prairie Warrior focus group analysis summary*. San Diego: Author. Retrieved July 31, 2000 from World Wide Web: http://www-code44.nosc.mil/cpof/pwfg.html.

Sticha, P.J., Singer, M.J., Blacksten, H.R., Morrison, J.E., and Cross, K.D. (1988). *Research and methods for simulation design: State of the art* (HumRRO Final Report FR-PRD-88-27). Alexandria, VA: Human Resources Research Organization.

Taylor, H.L., Lintern, G., and Koonce, J.M. (1993). Quasi transfer as a predictor of transfer from simulator to airplane. *Journal of General Psychology*, *120*, 257–276.

Trochim, W.M. (1999). The Research Methods Knowledge Base, 2nd Edition. Cornell, NY: Cornell University, Cornell Custom Publishing. Retrieved July 11, 2000 from the World Wide Web: http://trochim.human.cornell.edu/kb/index.htm (version current as of June 29, 2000).

Tufano, D.R., and Evans, R.A. (1982). *The prediction of training device effectiveness: A review of Army models* (ARI Technical Report 613). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

U. S. Army Training and Doctrine Command (1994, September). *The TRADOC Training Effectiveness Analysis (TEA) system* (TRADOC Regulation 350-32). Fort Monroe, VA: Author.

Waag, W.L., Raspotnik, W.B., and Leeds, J.L. (1992). *Development of a composite measure for predicting engagement outcome during air combat maneuvering* (AL-TR-1992-0002). Brooks Air Force Base, TX: Armstrong Laboratory.

Yerushalmy J. (1947). Statistical problems in assessing methods of medical diagnosis, with special reference to x-ray techniques. *Public Health Reports*, *62*, 1432–1449.

GLOSSARY

ACC	Air Combat Command
AETC	Air Education and Training Command
AFB	Air Force Base
AFHRL	Air Force Human Resources Laboratory
AFMC	Air Force Materiel Command
AFQT	Armed Forces Qualification Test
AFRC	Air Force Reserve Command
AFRL	Air Force Research Laboratory
AFSOC	Air Force Special Operations Command
AFSPC	Air Force Space Command
AL	Armstrong Laboratory (Brooks Air Force Base)
AMC	Air Mobility Command (AMC)
AOB	Armor Officer Basic
ARI	U.S. Army Research Institute for the Behavioral and Social Sciences
ASD(C3I)	Assistant Secretary of Defense for Command, Control, Communications, and Intelligence
BLUFOR	friendly (Blue) forces
C2	command and control
CAF	Combat Air Forces
CBP	comparison-based prediction
CCF	Critical Combat Functions
CNET	Chief of Naval Education and Training
CTEA	cost and training effectiveness analysis
CTER	cumulative transfer effectiveness ratio
DIS	distributed interactive simulation
DMDC	Defense Manpower Data Center
DoD	Department of Defense

DoDD	Department of Defense Directive
DoDI	Department of Defense Instruction
FY	fiscal year
GED	general education degree
HITL	human-in-the-loop
HumRRO	Human Resources Research Organization
IG	Inspector General
IPISD	Interservice Procedures for Instructional Systems Development
LSTS	large-scale training simulation
MAIS	major automated information system
MAUM	multi-attribute utility measurement
MDT2	Multi-Service Distributed Training Testbed
NAVMSO	Navy Modeling and Simulation Management Office
NAWCTSD	Naval Air Warfare Center Training Systems Division
NTC	National Training Center
O/C	observer/controller
OIG	Office of the Inspector General
OPFOR	opposing forces
OSBATS	Optimization of Simulation-Based Training Systems
OSD	Office of the Secretary of Defense
PACAF	Pacific Air Forces
R&D	research and development
ROI	return on investment
SIMNET	Simulation Networking
SMART	Simple Multi-Attribute Rating Technique
SME	subject matter expert
SYSCOM	systems command
TARGETs	Targeted Acceptable Responses to Generated Events or Tasks
TEA	training effectiveness analysis
TECIT	Training Effectiveness and Cost Iterative Technique

TER	transfer effectiveness ratio
TRAC–WSMR	TRADOC Analysis Center at White Sands Missile Range
TRADOC	Training and Doctrine Command
U-COFT	Unit Conduct-of-Fire
UJTL	Universal Joint Task List
UPAS	Unit Performance Assessment System
USAFE	U.S. Air Forces in Europe
USD(A&T)	Under Secretary of Defense, Acquisition and Technology
VV&A	verification, validation, and accreditation

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	October 2000	Final — January 1999–October 2000

4. TITLE AND SUBTITLE
On Measuring the Effectiveness of Large-Scale Training Simulations

5. FUNDING NUMBERS
DASW01 98 C 0067
BE-2-1709

6. AUTHOR(S)
John E. Morrison, Colin Hammon

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Institute for Defense Analyses
1801 N. Beauregard St.
Alexandria, VA 22311-1772

8. PERFORMING ORGANIZATION REPORT NUMBER
IDA Paper P-3570

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
ODUSD(R)R&T, PP
OUSD (P&R)
The Pentagon, Room 1C757
Washington, DC 20301

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
Review of this material does not imply Department of Defense endorsement of factual accuracy or opinion.

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Public release/unlimited distribution.

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 180 words)***

This paper identifies and evaluates methods for measuring the benefit or utility of large-scale training simulations (LSTSs). The intended audience is those people who are directly responsible for assessing the training effectiveness of LSTSs. We examine literature related to the measurement of LSTS training effectiveness [including guidance provided by the individual military Services and the Office of the Secretary of Defense (OSD)], the body of research on performance measurement, and research practices related to experimental design and analysis. We then synthesize the findings from these literature sources to provide seven recommendations for performing training effectiveness analyses of LSTSs: (1) identify specific measurement issues, (2) devise a measurement plan that specifies performance measurement and research design, (3) use valid and reliable performance measures, (4) obtain the most valid measure of proficiency possible within the constraints of the evaluation, (5) to the extent possible and reasonable, impose experimental control on the research situation, (6) use analytic models in early stages and continue using these models throughout system development, (7) and include user reactions as an adjunct to performance data and analysis.

14. SUBJECT TERMS
large-scale training simulations, training effective analysis, training effectiveness measures

15. NUMBER OF PAGES
78

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAR