AL/HR-TR-1996-0016

A COMPARISON AND INTEGRATION OF THREE TRAINING EVALUATION APPROACHES: EFFECTIVENESS, UTILITY, AND ANTICIPATORY EVALUATION OF TRAINING

George M. Alliger
Scott I. Tannenbaum

Executive Consulting Group, Inc.
409 Vesper Court
Slingerlands, NY 12159

Winston Bennett, Jr.

HUMAN RESOURCES DIRECTORATE
TECHNICAL TRAINING RESEARCH DIVISION
7909 Lindbergh Drive
Brooks AFB, TX 78235-5352

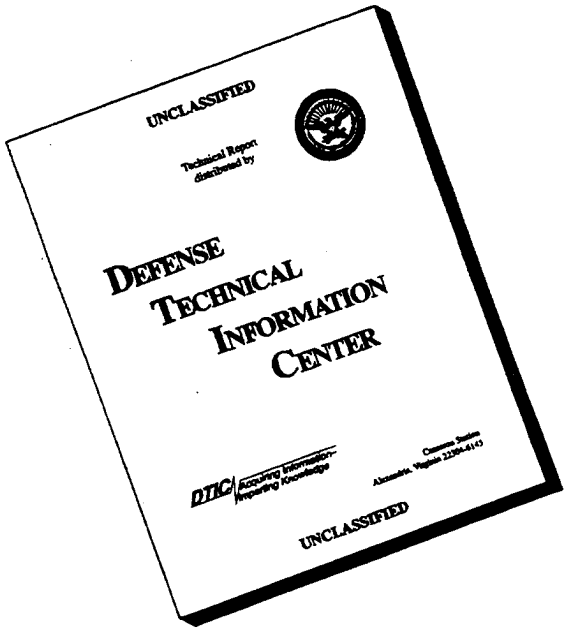
July 1996

Interim Technical Report for Period September 1993 - August 1995

AIR FORCE MATERIEL COMMAND
BROOKS AIR FORCE BASE, TEXAS

ARMSTRONG LABORATORY

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

NOTICES

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Office of Public Affairs has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This report has been reviewed and is approved for publication.

WINSTON BENNETT, JR
Program Manager
Instructional Systems Research Branch

R. BRUCE GOULD, Technical Director
Technical Training Research Division

JAMES B. BUSHMAN, LtCol, USAF
Chief, Technical Training Research Division

Please notify this office, AL/HRPP, 7909 Lindbergh Drive, Brooks AFB TX 78235-5352, if your address changes, or if you no longer want to receive our technical reports. You may write or call the STINFO office at DSN 240-3877 (or 240-2295) or commercial (210) 536-3877 (or 536-2295).

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	July 1996	Interim - September 1993-August 1995

4. TITLE AND SUBTITLE

A Comparison and Integration of Three Training Evaluation Approaches: Effectiveness, Utility, and Anticipatory Evaluation of Training

5. FUNDING NUMBERS

C - F41624-93-C-5011
PE - 62205F
PR - 1123
TA - C1
WU - 07

6. AUTHOR(S)

George M. Alliger
Scott I. Tannenbaum
Winston Bennett, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Executive Consulting Group, Inc.
409 Vesper Court
Slingerlands, NY 12159

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Armstrong Laboratory
Human Resources Directorate
7909 Lindbergh Drive
Brooks AFB, TX 78235-5352

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

AL/HR-TR-1996-0016

11. SUPPLEMENTARY NOTES

Armstrong Laboratory Technical Monitor: Winston Bennett, Jr. (210) 536-1981.

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This report provides a brief overview of three methods, or general approaches, to judging the usefulness of training. The first method or approach is the traditional evaluation approach of training effectiveness evaluation. This method or approach centers on estimations of training effect size and the determination of the statistical significance of those training effects. Thus, in this first category standard pre/post analyses and control group comparisons are included. Next, traditional training utility analysis, or training utility evaluation is reviewed. Here, costs and benefits of training are always contrasted in some way in determining traditional utility analysis. This branch of evaluation traces its roots back to Brogden & Taylor (1950), who discussed the need to examine the "dollar criterion." Finally, anticipatory training evaluation is discussed. Anticipatory training evaluation examines what kinds of training will have the greatest effectiveness and utility, given a variety of parameters and choices. The primary tool in anticipatory evaluation is Multi-Attribute Utility analysis (MAU). As opposed to training effectiveness and training utility evaluation, MAU is designed explicitly as a decision tool. It can be used most effectively as an anticipatory evaluation, the results of which facilitates planning for training. A detailed example of the development and application of MAU is described in this report, since it is the least well known of the three approaches to researchers and practitioners in the training area.

In explicating this expanded view of training evaluation, this report attempts to represent state-of-the-art understanding and research; thus current issues like risk and uncertainty in input and output evaluation and utility indices, and problems in transfer of learned skills to the job, are addressed in appropriate contexts. Recommendations for needed future research in training evaluation are discussed.

14. SUBJECT TERMS		15. NUMBER OF PAGES
Cost/Benefit analysis	Training effectiveness	56
Instructional system design	Training evaluation	16. PRICE CODE
Multi-attribute utility analysis	Utility analysis	

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

CONTENTS

CONTENTS (Continued)

List of Tables

List of Figures

PREFACE

The study of training effectiveness is undergoing great growth. It may even be said to be a time of upheaval, with old ideas being challenged and new ones proposed. This report is first dedicated to those trainers and decision-makers in U.S. Air Force who have influenced this work in many ways, not the least of which is the articulation of the difficulties and challenges of training faced every day. Those challenges include understanding how best to make strategic decisions about such emerging technologies as Intelligent Tutoring Systems.

This report is also dedicated to all of those who have contributed to our thinking about training evaluation, whether via published or unpublished writing, paper presentations, or simply fortuitous conversations. We have tried to recognize contributing individuals with citations wherever possible. Undoubtedly, however, individuals contribute to works like this in subtle ways, by influencing the very nature of state of the art training research. In particular, we would like to thank Dr. Mark Teachout for his comments on an earlier version of this report.

The opinions herein are those of the authors and do not necessarily reflect those of the Air Force. We thank Ms Gloria Koenig for her assistance in editing the final version of this report. Correspondence concerning this report should be addressed to Dr. Winston Bennett, Jr., Armstrong Laboratory, AL/HRTD, 7909 Lindbergh Drive, Brooks AFB TX 78235-5352.

A COMPARISON AND INTEGRATION OF THREE TRAINING EVALUATION APPROACHES: EFFECTIVENESS, UTILITY, AND ANTICIPATORY EVALUATION OF TRAINING

I. INTRODUCTION

Need for Improved Planning and Evaluating Air Force Training

Importance of Planning and Evaluating

That training plays a crucial role in the effectiveness and efficiency of the U.S. Air Force is indisputable. For this reason, the USAF dedicates a vast amount of people to the design and delivery of training. Moreover, the Air Force assigns individuals to evaluate training, so that training decisions can be made based on the best data available. In fact, a clear understanding of what kind of training is important for what jobs, and what factors influence training effectiveness, has historically been an emphasis in the USAF (Thorndike, 1947).

Not only must training design be optimized; the planning stage of training, prior to training design, must not be overlooked. Indeed, this is increasingly important since the technological sophistication required for many jobs is higher now than previously, job change is rapid, and new training technologies are available. Individuals involved in training decisions now face such a wide and complex range of possibilities, that understanding how to gather meaningful data and combine available information into a useful decision making framework is a daunting task.

An Example Training Problem. Consider the following case. The Air Force requires aircrew chiefs to master maintenance techniques for a new engine being deployed in an aircraft. The question is not whether there is a need for training. Instead, relevant questions include: Just how and where should training be delivered? Should a lecture format be adopted, or should performance and practice take precedence? How will trainers assure that maintenance mastery can be transferred from the schoolhouse to the field? Is investment in an Intelligent Tutoring System (ITS; Burns & Partlett, 1991) warranted? If so, just where should such technology be employed and how many ITS units would be required? Would such a plan be cost effective?

Problems like these are not rare, but increasingly common. The goal of this report is to begin to arrange some of the research facts and theory on training evaluation to permit decision makers to understand how the evaluation of training can help make their decisions the best possible.

The Training Design Process

This report argues that many kinds of questions are in fact not answerable given only the results of standard evaluations of training. In fact, the usual static model of training evaluation is simply obsolete. This obsolete model suggests that training evaluation is an end in itself (see Figure 1). Many training evaluations simply index the performance of trainees via paper and pencil or computer tests administered within the training environment. But, unless this information is integrated with the entire framework of training design and implementation, the evaluation is static and any advantage gained from it is relatively limited. Some more sophisticated training evaluations can yield a dollar training utility gain index, but are nevertheless still static and "dead-end." In such utility evaluations, the presumed cost of a training program is weighed against its benefits according to a formula which provides a gain or yield of one training approach over another or of training versus no training. Like the standard training evaluation, however, such approaches are also clearly short of what is needed for even a typical case like the one outlined above.

Acceptance of the model displayed in Figure 1 may not be clearly stated -- rather, it is often the implicit model in the minds of those who request or carry out training evaluations. Adherence to such a model, however, will fail to lead to solutions to the kinds of questions that are often important. Very often these questions are forward-looking, or anticipatory. As in the aircrew chief maintenance training example described above, we often need evaluations precisely to inform planning for training and implementation.

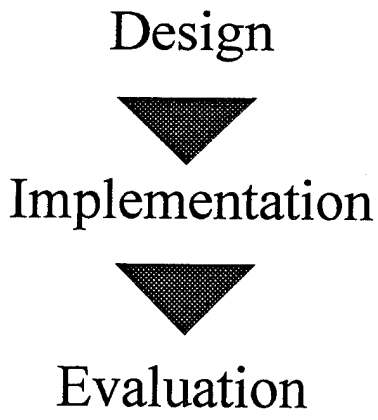


Figure 1: The Traditional View of Training Evaluation as an End in Itself

In order to consider training evaluation in such a way that it is most useful, several steps in evaluation model development need to be taken. First, it is necessary that training

evaluation be understood as one part of the entire process of training, as in Figure 2. This integrated view, referred to as the Instructional Systems Design (ISD) view, shows that training evaluation results feed back into training planning and design.

The ISD approach has a number of advantages. Not only is evaluation seen as only one part of the total training system, but it can feed back information useful for future training. Indeed, the ISD approach can be thought of as an implementation of the "Total Quality" or quality control philosophy. During each ISD cycle, discrepancies between the target and actual outcome are reduced and thus total quality is approached ever more closely. In training, the training target is presumably some immediate measurable training outcome or some more distal but still measurable job outcome, such as (for training in an immediate sense) number of washouts or washbacks, or (for training transferred to the job) percentage of mission aborts, aircraft available for sorties, or successful launches and recoveries. Over repeated training evaluation iterations, the discrepancy between the target and actual training outcomes can be reduced. Such a focused emphasis is in keeping with the concept of "continuous improvement."

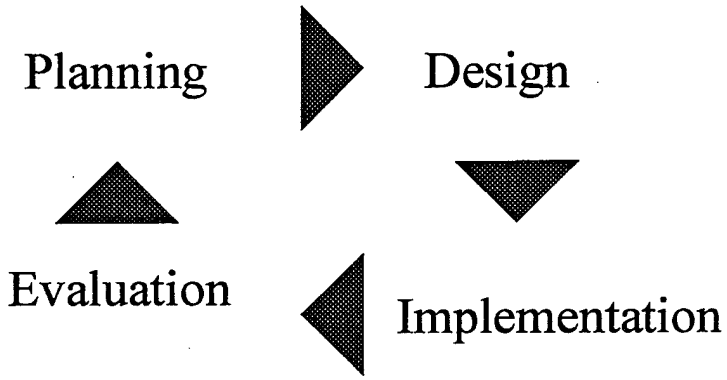


Figure 2: The Instructional Systems Design Process

Thus, an ISD view of training evaluation is better suited to address a situation like the aircrew chief maintenance question raised above. In particular, previous training evaluations might be able to provide information for current pending decisions on training design and implementation. Schematically, Figure 2 can be exploded to reveal explicitly what is implicit in the ISD systems approach: the repeating process of training planning, design, implementation, and evaluation (Figure 3). Figure 3 is meant to illustrate that each complete training design/evaluation process can be tied to other ones occurring later in time.

The interconnection between previous and later training design/evaluation cycles occurs because a training evaluation rarely produces information of use only to a given

training project. Some findings will generalize. For example, if a particular training format is found to be useful for a given type of content, then this finding should probably influence future training designs. As we will develop in this report, there are many other kinds of information that can be obtained from training evaluation, which can then be used in later training planning and design. Meta-analyses of various training issues (e.g., Alliger & Tannenbaum, 1995; Bennett, 1995) are one important way in which training results from many studies can be reviewed and synthesized, making them available for future researchers interested in estimating, for example, the likely effectiveness of a given training method.

We believe that many evaluation psychologists and instructional design experts have already made the transition from the traditional evaluation-as-an-end-in-itself view of training evaluation (Figure 1). These professionals now embrace, at least at the conceptual level, the ISD view (Figures 2 and 3). However, we also will argue in the pages to follow that the ISD view, while correct, is not yet fully articulated. Here the word "articulated" is used in both its major senses: the ISD view is not yet fully explained and described in sufficient detail, nor is it sufficiently faceted and evolved, to be as useful and prescriptive for both researchers and practitioners as it might be. Such an articulation is an important step in understanding training evaluation in its complete breadth and width.

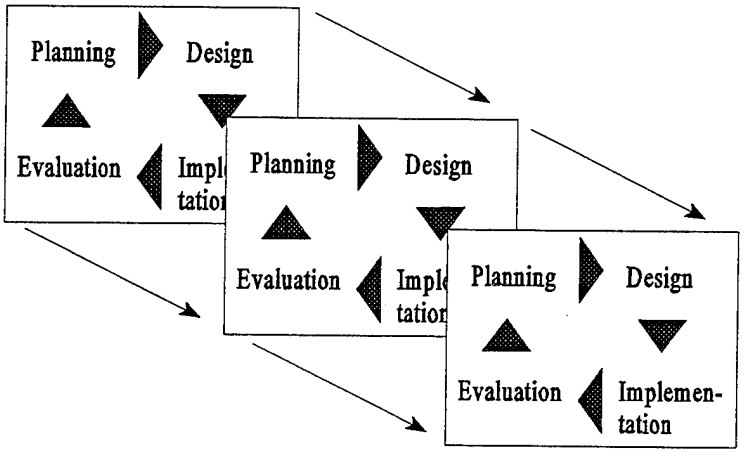


Figure 3: The Iterative Instructional Systems Design Process

Articulating the ISD model: Integrating three approaches to training evaluation

A useful way to expand the ISD model and to make it more practical is to consider how three relatively distinct approaches to training evaluation and planning might be integrated into this model. One purpose of this report, therefore, is to provide a brief overview of three methods, or general approaches, to judging the usefulness of training. First, we review traditional training evaluation. This we term *training effectiveness evaluation*, since it centers on estimations of training effect size and the determination of the statistical significance of those training effects. Thus, in this first category standard

pre/post analyses and control group comparisons are included. Second, we review traditional training utility analysis. We term this *training utility evaluation*, since costs and benefits of training are always contrasted in some way in determining traditional utility analysis. This branch of evaluation traces its roots back to Brogden & Taylor (1950), who discussed the need to examine the "dollar criterion." So, for example, a dollar figure representing the benefit of training after costs are considered would be a utility evaluation; similarly, any index (Return on Investment, Net Present Value, ratio of man/years saved to man/years spent) which explicitly compares costs and benefits is included in this second category of evaluation. Finally, we consider *anticipatory training evaluation*, which - examines what kinds of training will have the greatest effectiveness and utility, given a variety of parameters and choices. The primary tool in anticipatory evaluation is Multi-Attribute Utility Analysis (MAUA). As opposed to training effectiveness and training utility evaluation, MAUA is designed explicitly as a decision tool. It can be used most effectively as an anticipatory evaluation, the results of which facilitates planning for training. We include a detailed example of MAUA in this report, since it is the least well known of the three approaches to researchers and practitioners in the training area.

In explicating this expanded view of training evaluation, this report attempts to represent state-of-the-art understanding and research; thus, current issues like risk and uncertainty in input and output evaluation and utility indices, and problems in transfer of learned skills to the job, are addressed in appropriate contexts. We conclude the report with both recommendations and discussion of needed future research in training evaluation.

II. THREE GENERAL APPROACHES

This paper defines and explains three basic approaches to training evaluation. These are 1) training effectiveness evaluation, 2) training utility evaluation, and 3) anticipatory training evaluation.

Training effectiveness evaluation is the determination of the impact of training in terms of some dependent measure or measures, such that "impact" means a change or improvement in those measures. Training effectiveness is thus meaningful when questions such as the following are addressed. "Did the trainees learn anything?" "Did the trainees in this group learn more than the trainees in that group?" "Can the trainees perform X after training?" Thus, training effectiveness evaluation provides some indication that training has moved the trainees on some dependent variable such as the assessment of declarative or procedural knowledge, or indicators of on-the-job performance.

Training utility describes the benefits of training relative to the costs of training. Clearly, one benefit due to training is indexed by the degree to which training is effective. Thus, the first topic we cover, training effectiveness, is closely linked to the second, training utility. Also, training utility is often a central concern of planners and those who wish to consider what training should be implemented, or whether it should be continued. Thus, training utility is also linked to the third view of training evaluation we cover: anticipatory training evaluation.

The third method we term anticipatory training evaluation. It may use information from training effectiveness studies, training utility estimates, and as much other information as deemed important, to provide comparative ratings for several training alternatives. It is a particular implementation of an approach to decision making termed Multi-Attribute Utility.

These three approaches to the evaluation of training, far from being independent, are linked to each other and reside within the broader context of the entire instructional design model.

III. INTEGRATING THE THREE APPROACHES

As defined above, training effectiveness evaluation, training utility evaluation, and anticipatory evaluation, can be integrated into the traditional instructional design process. Such an integration is shown in Figures 4 and 5.

Integrated ISD model

Figure 4 shows the ISD model including the three approaches to training evaluation. We discuss below each aspect of the ISD model in Figure 4 prior to a more complete examination of the three training evaluation approaches. Figure 5 shows the iterative nature of the integrated ISD model.

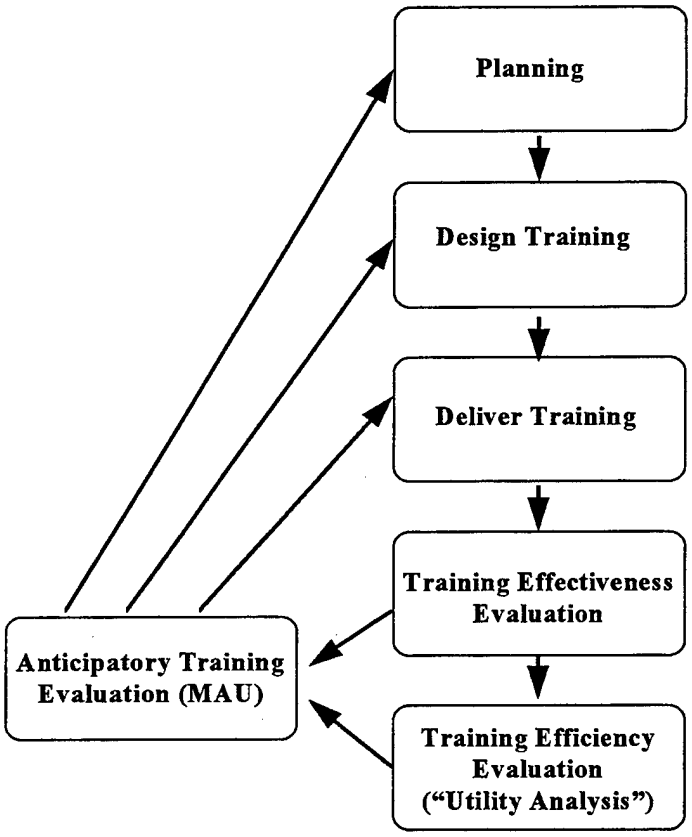


Figure 4. The Instructional Design Process Including Three Approaches to Training Evaluation

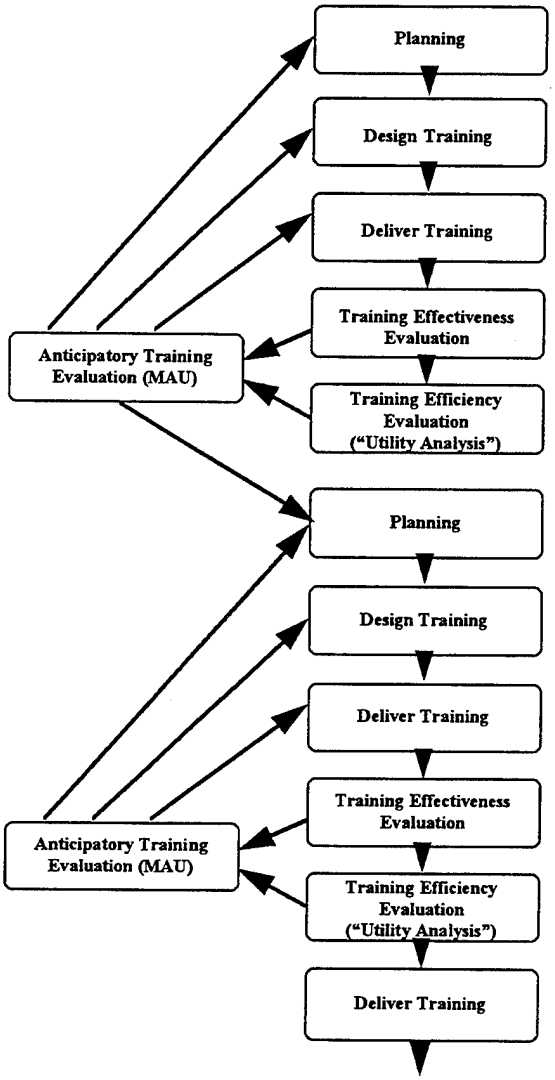


Figure 5. The Interative Instructional Design Process Including Three Approaches to Training Evaluation

Planning. In most cases, the process begins with a planning process. In this step some rational approach is carried out to determine what training would benefit the organization. Survey questionnaires (e.g., skills assessment or critical task identification); interviews with subject matter experts, job incumbents, or customers; content analysis of a particular topic; or other approaches to identifying training needs can be employed. The outcome is typically the ranking of tasks, job domains, or skill sets that are either important to the job in question, or in which the current incumbents are identified as lacking to some degree.

Planning is an important part of the design process. A meta-analysis of training found that training effectiveness (as indexed by *d*, standardized effect size) may be influenced by the degree to which planning has occurred (Bennett, 1995). More specifically, he found that studies reporting a training needs analysis also showed greater mean training effectiveness. In this meta-analysis, Bennett (1995) coded training effectiveness studies by whether they reported no training needs analysis information, or whether they reported conducting a person, task, or organizational level needs analysis (McGehee & Thayer, 1961). Studies reporting two of these forms of needs analysis had a higher mean *d* than did studies reporting none.

Some elegant method for analyses and display of training needs and training course data for purposes of examining training efficiency have been developed, such as the matching technique (Ford & Wroten, 1984). This procedure helps to identify which tasks may be over- or under-trained in a particular course. This technique has been applied (Teachout, Sego, and Olea, 1993; Teachout, Olea, Barham, and Phalen, 1993), and extended (Teachout, Sego, & Ford, 1995a, 1995b) for purposes of improving training content, training efficiency, training transfer, and for training re-design. That is, this approach can be important in the planning of training, as well as the design of training.

Training Design. Training design is based upon information generated in the planning step. At this point, the actual nature of the training is decided upon, including media to be used and content to be presented. Traditionally, it is thought that various "learning principles" and "transfer principles" should be followed in order to develop the most effective training (Cascio, 1991). Alliger, Tannenbaum, and Bennett (1995) question this assumption, partly because many of the so-called principles were discovered in laboratory settings, rather than in field research. In some cases, it seems that maximizing transfer to the job is done best by training precisely on the tasks that will be performed, with the training environment designed to replicate as closely as possible the work environment. In this way, trainees are not trained to maximize acquisition of the learning-specific task in a learning-specific environment, but rather are learning exactly what they will perform in the same kind of environment in which it will be performed (cf. Detterman, 1993). However, what we have just described as desirable reflects "high-fidelity" training. Fidelity is itself a principle which has its roots in laboratory research. Thus, it does not seem possible to make simple statements about the advisability of laboratory-derived training principles.

Ideally, evaluation planning should also begin at this time. While it is not uncommon for training evaluation to be first considered long after training is underway (or even finished!), the most effective approach is to plan training evaluation at the same time training itself is planned. In this way, evaluation can be integrated into the training rather than being, and feeling like, an "add-on." The advantages of integrated training evaluation can include early warning for trainees and trainers alike of trainee or curriculum problems, optimization in data collection, and the most ideal matching of evaluation tools to training content. Many organizational obstacles exist to effective training evaluation (Tannenbaum

& Woods, 1992). These are difficult to overcome unless they are considered early in the design process.

Indeed, the careful consideration of training evaluation in the planning phases of training can have an impact on the design of training itself. As designers struggle with the appropriate criteria of training success, the appropriate purposes and emphases of training will be reviewed and clarified as well. Good training evaluation principles make it necessary to consider the climate for transfer in the workplace in which the trainees are expected to apply their training (e.g., Tracey, Tannenbaum, & Kavanagh, 1995). For example, the extent to which individuals have opportunities to rehearse the actions they were trained to perform has a profound impact on training transfer (Ford, Quinones, Sego, & Sorra, 1992), as can trainee expectations (Tannenbaum, Mathieu, Salas, & Cannon-Bowers, 1994).

Careful training evaluation planning helps assure that the data gathered will be able to answer decision-makers' questions. The evaluation results will be more likely to indicate how closely the goals of the training, as understood by the customers of the training, have been met.

The identification of appropriate training success criteria is not a simple process. In the Air Force, some aspect of task performance is often the selected criterion. However, other criteria may be equally important. Semi-structured interviews with customers of training (e.g., supervisors of newly graduated trainees) can assist in the identification of proper criteria.

Part of the design of training is the development of course metrics. Video shoots, computer program coding, construction of exercises and trainee materials, and so forth are a major cost in the training process. Development of training evaluation tools will also occur at this stage. Presumably the criteria are clearly understood, having been defined in the previous stage. Knowledge tests can be written, skill simulations constructed, processes for collection of archived data instituted, and field performance data collection planned.

Training. Training is now conducted. Training can, of course, take many forms, from intelligent tutoring to lecture. It can vary on many dimensions (e.g., duration, intensity, location, difficulty, etc.).

Training Effectiveness Evaluation. This is the step where training evaluation results are obtained. Data, including all pre-training and post-training data, are analyzed according to whatever analysis plan best addresses the questions the training evaluation design was meant to answer.

As can be seen from Figures 4 and 5, training evaluation fits into the instructional design process as the effectiveness assessment approach most immediate to the training itself. Often, results will show whether certain parts of the training appear to work in terms

of immediate learning and/or in terms of transfer to job performance. One of the major outcomes at this point is the computation of a *training effect size*, which indicates the degree to which the trainees are superior after training, relative either to themselves prior to training (a within-subjects training effect size) or to others either not trained or otherwise trained (a between-subjects effect size).

Training Utility Analysis. Once training effectiveness is determined, this information can be fed into the next step, which is the assessment of the utility of the training. In essence, utility is some index that compares training costs and benefits.

Anticipatory Training Evaluation (MAU). The third evaluation aspect of the articulated ISD model is anticipatory evaluation. Using inputs from each of the prior evaluation approaches, anticipatory evaluation can incorporate other additional information as well. It is this aspect of evaluation which truly permits the completion of the ISD model, so that previously conducted training is very useful in helping decide among alternatives for future training.

IV. TRAINING EFFECTIVENESS EVALUATION

Evaluation of the effectiveness of training can be accomplished in different ways. This aspect of training evaluation is probably the one most studied and written about. For example, articles on training evaluation design or discussions of training criteria usually focus on the detection of training effectiveness. In this section we review training effect size, training evaluation designs, and training criteria.

Longitudinal Evaluation

There is increasing agreement that development of longitudinal approaches to training evaluation is crucial. Indeed, there may be somewhat of a paradigm shift occurring in this regard (cf. Tannenbaum & Yukl, 1992). If we are ever to develop scientifically verifiable evidence of nonlinear trends across time, longitudinal designs that go beyond simple pre-post designs are necessary. Educational theorists have recognized this for some time; individual learning curves can be studied only when multiple sampling of student performance occurs across time (Rogosa, 1988; Rogosa, Brandt, & Zimowski, 1982). Training researchers have recently been discussing phenomena that are not less needful of longitudinal measurement. Concepts like "transfer curves" (Baldwin & Ford, 1988) and generalized retention or decay curves also require many points in time. The expansion of the concept of longitudinal utility also requires multiple samples of performance so that trends in utility increment or decrement can be plotted.

An emphasis on longitudinal evaluation also highlights the importance of environmental factors in determining training effectiveness (Tannenbaum & Yukl, 1992). Many inhibitors (e.g., lack of peer or management support) and facilitators (e.g., job aids, opportunity to practice) exist in the workplace and can help augment or undermine training transfer.

Training Effect Size

The determination of training effectiveness centers around the determination of an effect size. Although this topic is important in its own right, it will be made evident later that training effect size has implications for the other two arms of training evaluation, namely training utility estimation and anticipatory training evaluation.

Technically, an effect size is a sample-size independent index of experimental effect magnitude. Thus effect size is often estimated, using some dependent measure of interest, by the standardized difference between two groups (e.g., trained and control group). "Standardized" here means that effect sizes from different samples can be directly compared without further scaling.

Equation (1) can be used for computing the *d*, the standardized mean difference between a trained and a control group. The control group may be a group receiving no training, or a different kind of training. It should be noted that the effect size *d* is a biased estimator of δ, the population mean difference in standard deviation units. The amount of

$$d = \frac{(\bar{X}_{Trained} - \bar{X}_{Control})}{sd_{pooled}} \qquad \textbf{(1)}$$ Standardized mean difference

bias is positive but relatively small, to the order of 1 - 3/(4N-9) (Hedges & Olkin, 1985). Thus, given a d of .5, and an N of $n_{Trained} + n_{Control} = 100$, the total bias is 1 - 3/(400-9) = 0.992327. Thus the best estimator of δ is not .5 but .5(.992327) = 0.496163. We can conclude that for practical purposes, the bias of equation (1) is small, and that the use of this equation to estimate training effect size is reasonable. Note too that "*d* = .5" is shorthand for "*d* = .5 standard deviations." This is one sense in which standardization occurs: two *d*'s from entirely different studies can be directly compared because both are in the same standard deviation metric. The second, and corollary, sense in which *d* represents a standardized statistic is that it does not reflect sample size in the way that a *t* - statistic, for example, does.

It should be noted that equation (1) presupposes two independent groups, a trained and a control group. The two means which are entered into equation (1) are thus in most cases means of post-training performance on some dependent variable of interest. Or, a *d* could be computed as the comparison between groups of pre-training to post-training gains. (Contrary to popular understanding, the direct comparison of gain scores from two independent samples is completely appropriate. In fact, a *t*-test computed on such gain scores is equivalent to an analysis of covariance on the same data, where pre-training scores are the covariate; Maxwell & Howard, 1981.) Realistically, it is rare that trainees can be randomly assigned to training classes and control groups (Tannenbaum & Woods, 1992). Fortunately some alternatives exist. For example, nothing prevents the computation of a *d* on a single trained group; that is, the means compared can come from the pre-training and post-training measures. Such a *d* would differ from that of equation (1) in the same way that a dependent *t*-test differs from the independent samples *t*-test. A *d* may also be computed from quasi-experimental designs. So, for example, a *d* can be calculated by subtracting from the post-training measure for one cohort the pre-training score for another cohort, and dividing by the pooled SD (e.g., Tannenbaum & Woods, 1992). Of course, the standard threats to validity need to be kept in mind whenever a quasi-experimental design is used (Cook & Campbell, 1979). The different ways a *d* may be computed will be highlighted when we discuss training effectiveness designs.

Some authors (e.g., Alliger & Scherer, 1995; Cohen, 1977; Rosenthal & Rubin, 1982) have written on the best way to interpret *d*, or explain its meaning to others. A *d* is, in simplest terms, the mean difference between the control and trained groups in standard

deviation units. Thus, for example, a *d* of .5 indicates that the mean of the trained group is .5 standard deviations above the mean for the control group. Such an effect size can, however, be described in other terms. Actually, using an additional description of a given *d* can be very useful, since a simple effect size, standardized or not, may not give a clear picture of the impact of training. Additional ways to describe the difference between a trained and control group include a) the percent of trained group scoring higher than mean control; b) percent of trained group correctly classified -- this reflects how well trained/control group membership could be predicted given only scores on training criteria; c) percent of non-overlap between control and experimental groups (Cohen, 1977); d) percent of times a member drawn randomly from the trained group will have a higher score than a member drawn randomly from the control group -- this is called the "Common Language Effect Size" (McGraw & Wong, 1992); e) the success rate of the trained group compared to the success rate of the control group -- termed the "binomial effect size," this requires a success cutoff on the training criterion (Rosenthal & Rubin, 1982); and f) percent of shared variance between group membership and training criterion -- commonly called "variance accounted for," this index is based on square of correlation between group membership (trained or control) and the training criterion measure. Please note that each of these ways of looking at effect sizes are first, all based on *d*, and second, each uses a percent metric. Figure 6 (Alliger & Scherer, 1995) further illustrates each effect size interpretation.

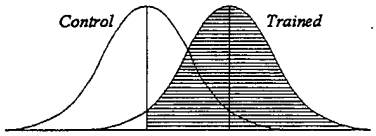
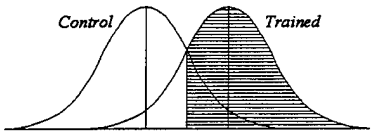
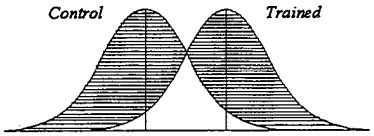
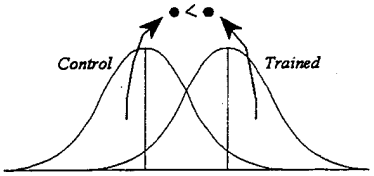
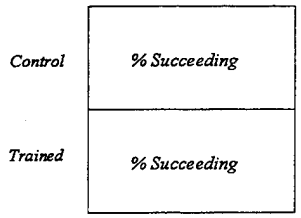
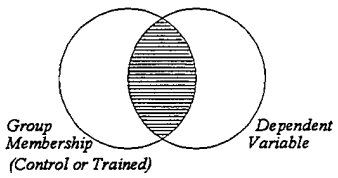
Definition	Diagram showing training effect.
1. Percent of trained group scoring higher than mean control group score. Value is 50% if no effect (null is true).	
2. Percent of trained group correctly classified. Reflects how well trained/control group membership could be predicted given only scores on DV. Value is 50% if no effect.	
3. Percent of non-overlap between control and trained groups. Value is 0% if no effect.	
4. Percent of times a member drawn randomly from trained group will have a higher score than a member drawn randomly from control group. Value is 50% if no effect.	
5. Success rate of trained group compared to success rate of control group. Requires success cutoff (dichotomization) on DV. Value is 50% if no effect.	
6. Shared variance between IV and DV; commonly called "variance accounted for," based on square of correlation between group membership (trained/control) and DV. Value is 0% if no effect.	

Figure 6. Comparison of Six Ways to Interpret Effect Size. Adapted from Alliger & Scherer (1995).

In order to provide a sense of how each of these effect size interpretations relate to each other in terms of magnitude, consider an example. Suppose the U.S. Air Force introduces a new type of training method for mechanics, such as requiring training experience on operational ("hot") aircraft rather than decommissioned ("cold") aircraft. These two types of training could be compared for effectiveness. Perhaps on-the-job readiness simulations are performed by graduates of both "hot" and "cold" training. These simulations provide the dependent variable (e.g., number of errors, time to perform, performance confidence, etc.). The group identity of the trainees is either "hot" or "cold" training. After the data on the dependent variable are collected, a *d* is computed, according to formula (1). Suppose this *d* is equal to .5. As mentioned, this simply indicates that the "hot" group is ½ of a standard deviation higher than the "cold" group (which in this example serves as the control group). The seven interpretations of this *d* previously defined can then be computed. Table 1 shows the results. It is clear from Table 1 that not every interpretation of the same underlying effect size will have the same psychological or perceptual impact. In fact, this very difference in impact underlies the development of some of these effect size interpretations. The binomial effect size, for example was designed to illustrate that some effects seem important when interpreted as success rates even when the correlational interpretation seems weak (Rosenthal & Rubin, 1982).

Table 1
Comparison of Effect Size Interpretations

	d		
Interpretation of *d*	.2 (small effect)	.5 (moderate effect)	.8 (large effect)
Percent of trained group A above mean of trained group B	57.9%	69.1%	78.8%
Percent of trained group A correctly classified	54.0%	59.9%	65.5%
Percent of non-overlap between A and B	14.8%	33.0%	47.4%
Percent of random pairs where score of A will be higher than score in B	55.6%	63.8%	71.4%
Success rate in A versus that in B	55.0% vs. 45.0%	62.1% vs. 37.9%	68.6% vs. 31.4%
Percent variance accounted for	1.0%	5.9%	13.8%

As a final comparison of training effect size interpretations, we offer two considerations. First, Figure 7 (from Alliger & Scherer, 1995) shows the percent magnitude of these interpretations across a range of *d*. Note that for the same underlying *d*, the interpretation can vary widely. For example, except for $d = 0$, the Common Language and Percent Above Mean Control (perceptual or psychological) interpretations of training effect are always larger for any

given underlying *d* than are other interpretations. Indeed, by interpreting the underlying training effect by r^2, a researcher might inadvertently convey that the training effect is much less powerful than other interpretations would convey.

A second consideration that is of serious interest is the typical magnitude of effect sizes found in by training evaluation studies. Bennett (1995) searched the scientific literature for training effectiveness studies for which *d* was available or could be recovered. This study reports meta-analyses on 466 independent *d*'s from 177 training evaluation studies. Overall types of training, and with a total cumulated sample size across the 466 *d*'s of N =47,605, Bennett (1995) found a mean weighted *d* of .76. Of course, he reported extensive moderator analyses, for this index of central tendency should not be taken as indicative of a single underlying population effect size.

Training Designs

The nature of effect size in general and how it is encountered in the training literature has been discussed. A point still to be made, however, is that indices of training effectiveness, as indicated by an effect size, can be obtained from different designs. "Design" here is meant to signify the kind of study the researcher employs -- whether within or between subjects, for example, whether a control group is used, and so forth. We will discuss several general designs. These are probably best thought of as prototypes, or ideals, which may be more or less realized in practice.

1. Pre-training measures only. No effectiveness estimation possible, even if a control group is available. Any computation of effect requires some post-training information.

2. Post-training measures only. If only post-training measures are available, a training effectiveness index can be computed, but only if there are two groups, a trained and a control. Such an effect size index is subject to various threats to validity (Cook & Campbell, 1979), but may still be a valid index of training effect if those threats can be reasonably ruled out. In the case of a single group, a training effect size cannot be computed.

3. Pre-training and post-training measures. A training effect size can be computed, whether you have two groups (control and trained) or only one group (trained). In the case of a single, trained group, dependent effect estimation is possible. That is, the pre-training mean is subtracted from the post-training mean, and the difference is divided by the pre-post pooled standard deviation. Again, there are threats to validity in such a single group design which need to be addressed.

In the case of two groups (control and trained) a training effect size can definitely be computed. This design represents the classic case where a control group is available, and measures of knowledge or performance can be assessed both before and after training. An effect size can be computed in at least two different ways. First, the mean pre- to post-gain can be

computed for each group, and the difference between the two can be computed and standardized by the appropriate pooled standard deviation term. Or, second, the difference between the post-training measures can be entered into the effect size computation.

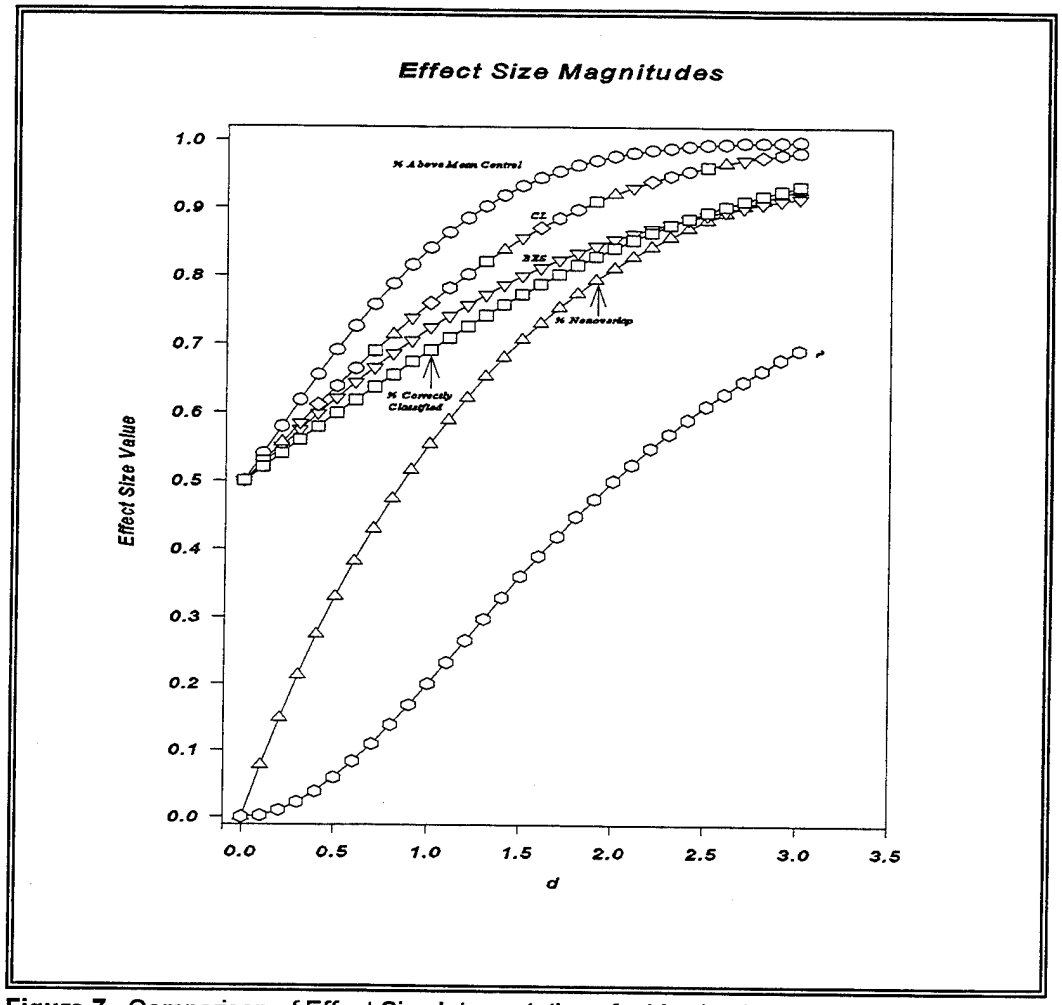


Figure 7. Comparison of Effect Size Interpretations for Varying Levels of Actual Underlying Effect. From Alliger & Scherer (1995).

Table 2 illustrates a simple comparison between designs and when training effectiveness can be calculated. Actually, things are not as simple as indicated in the table. If, for example, one has only a trained group, but multiple cohorts are available, then effectiveness may be computed in additional ways.

Table 2
Feasibility of Computing Training Effectiveness for Three Training Designs

	Single Group	Two Groups
Pre-training measures only	Not feasible	Not feasible
Post-training measures only	Not feasible	Feasible
Both pre- and post-training measures	Feasible	Feasible

Training criteria

A third important consideration in the discussion of training effectiveness relates to the topic of training criteria. "Training criteria," as we use it here, is a phrase that encompasses what in an experimental setting are termed "dependent variables." In the case of training evaluation, the dependent variables of interest usually include knowledge or skill performance of some kind. Of course, the development of appropriate measures are guided by the goals of the training. Criteria can be temporally proximal (measured close to training) or distal (measured some time after training). They can be specific or global, multiple or composite. In 1959 and 1960, Kirkpatrick (1959a, 1959b, 1960a, 1960b) proposed a four-fold taxonomy of training criteria. Each had a different purpose. Level 1 was termed Reactions, and was defined as learners' "liking of" and "feelings for" a training program. Level 2, Learning, was defined as "principles, facts, and techniques understood and absorbed" by learners. Level 3 was Behavior, defined as "using [learned principles and techniques] on the job." Level 4, Results, was simply the ends, goals, or "results desired... reduction of costs; reduction of turnover and absenteeism; reduction of grievances; increase in quality and quantity of production; or improved morale." This now well-known scheme should not be taken as the last word in training criteria, and indeed carries serious flaws when taken too literally. For example, the hidden assumptions of linear causal links from reactions to results is probably mistaken (Alliger & Janak, 1989).

Alliger and Tannenbaum (1995) presented an augmented version of Kirkpatrick's taxonomy. They suggested that Kirkpatrick's Level 2, learning, could be divided into learning measured just after training (Level 2a), and Level 2b, which is learning measured some time after training (retention). They further suggested that behavior (Kirkpatrick's Level 3) could be divided into performance assessed just after training (Level 3a) and performance on the job (Level 3b). Table 3 presents this expanded view of Kirkpatrick's training criteria, as specified by both time of measurement and what is measured.

Table 3 A Taxonomy of Training Criteria				
	Attitudes toward training	Learning	Behavioral Performance	Effects of learning
Temporally proximal (just after training)	Affect and Utility Reactions (Levels 1a, 1b)	Learning (Level 2a)	Immediate Performance (Level 3a)	N/A
Temporally distal (on-the-job)	Affect and Utility Reactions (Levels 1a, 1b)	Retention (Level 2b)	Application/ Transfer (Level 3b)	Organizational impact (Level 4)

Warr and Bunce (1995) added a third category to reaction-level training assessment. In addition to training satisfaction and perceptions of the usefulness of training, they suggested that perceptions of training difficulty would be important to assess. Warr and Bunce (1995) indeed found that difficulty reactions were negatively related to later measures of training effectiveness.

V. TRAINING UTILITY ANALYSIS

The assessment of the utility of training is really the assessment of its input-to-output efficiency. The term "efficiency" is used here in a broad sense, to refer to the totality of training utility in terms of cost effectiveness. This use of the term "efficiency" needs to be distinguished from the efficiency with which training delivery is matched to training needs (cf. Teachout, Sego, & Ford, 1995).

Utility analysis is a collection of methods for assessing the cost-effectiveness of many interventions, including training (e.g., Mathieu & Leonard, 1987). Actually, the applications of utility analysis to training have been relatively few -- most applications have been to personnel selection. While the estimation of training utility is relatively straightforward, as we proceed to show, two preliminary points should be made.

The first point is that there is no one, or even best, way to estimate the utility of training. This is in part because utility can serve so many different purposes, and address so many different audiences. For one group, a simple Return-on-Investment index may be sufficient. Others may wish to incorporate the opportunity costs of money (inflation); in this case an index like Net Present Value may be useful. More complex variants of utility which also use dollars as the main metric include extensions of the Brogden-Cronbach-Glaser models. It is also possible that one might wish to estimate training utility, but that dollars are not the central focus for the audience of interest. In such cases, relative gains in productivity (e.g., numbers of sorties flown, numbers of successful launches and recoveries) for fixed development efforts (e.g., number of man/months) could be estimated. Thus, so long as there is some comparison of training costs (money or work or time invested) to training benefits (money or work or time saved or gained), utility is being estimated.

The second point is that utility estimates are virtually always attended by a large degree of uncertainty. That is, even though equations may provide a point estimate of dollars, work, or time saved, the point value needs to be hedged. For example, a Return on Investment (ROI) of 150% seems very persuasive, since it indicates that for every dollar invested a dollar and a half is realized. But, there is clearly some uncertainty associated with either the costs or benefits that are used to calculate the ROI. This is particularly true when there are "expert judgments" being made with regard to costs, benefits, or both. Such expert judgments are, in fact, very common, because it is rare that objective indicators of all costs and benefits are available. Often, benefits are projected from initial data of some kind, and the bases of the projection may be very much open to debate. But even costs, which can seem to be "harder" and more certain than benefits, may be hard to document when examined carefully. For these reasons, the best practice is to compute different utility estimates under different conditions. "Breakeven" analyses (where the minimum benefit is defined in order to balance investment) and "sensitivity" analyses (where one considers how changes in various cost and benefit estimates effect the utility outcome) are two systematic ways to understand how robust a utility estimate is. Still other authors have developed methods which permit the construction of actual confidence intervals around utility estimates, much like

the standard confidence intervals which are placed around a mean and which help illuminate the stability of that mean and whether it can be assumed to be greater than zero over repeated sampling.

ROI models

For simple ROI utility models, there are three basic steps to the determination of training utility. First, training costs should be calculated. Training costs, as defined here, may include costs associated with any number of factors, including development, administration, salary, housing, and delivery costs. The simple rule when computing costs is to consider thoroughly and logically the hidden and evident costs associated with the entire training effort. Checking for completeness with as many people involved in the training as possible can lead to the greatest assurance that all (or at least most) relevant factors have been considered and estimated with at least reasonable accuracy.

Second, benefits need to be estimated. The costing of benefits is often difficult. Sometimes actual dollar values are available, as in the case of sales commissions. For the military, however, translations of productivity benefits or performance improvements into dollars may be required for an ROI, if that ROI is to be based on dollars.

Third, ROI is calculated. Since ROI = (Benefits/Costs) X 100, this is a not a difficult procedure. Any ROI which is greater than 100 is considered to indicate positive utility of training.

It is also possible, with some additional assumptions, to calculate an index termed "Net Present Value" (NPV). Since NPV additionally takes into account the "cost of money," computations for it are more complex than simple ROI models. That is, inflationary forces over time can affect the prediction of return, and NPV includes these. The formula for net present value is:

$$NPV = \text{Total Present Value} - \text{Costs} \qquad \textbf{(2)} \text{ Net Present Value}$$

In order to calculate NPV, the total present value of the training must be calculated. This requires computing the present value (PV) for as many years as the training is projected to be effective, or some fewer number of years. PV is the value of training for a given year, divided by 1 plus the expected interest rate, raised to a power equal to the number of years since training:

$$PV = \frac{\text{Value of Training for Year X}}{(1 + \text{Expected Interest Rate})^{\text{Number of Years Since Training}}} \qquad \textbf{(3)} \text{ Present Value}$$

The total present value (TPV) is the sum of all present values:

$$TPV = \text{PV for Year 1} + \text{PV for Year 2} + \ldots + \text{PV for Final Year} \qquad \textbf{(4)} \text{ Total Present Value}$$

The net present value is then computed over the expected life of the training effects, minus costs as in equation (2).

SDy models

Estimating the return on training can also be performed with a variant of the Brogden-Glaser-Cronbach general utility model. A calculation of annual net payoff (so-called "ΔU") can be performed by the following equation:

$$\Delta U = (n)(d)(SDy), \qquad \text{(5) Annual net payoff}$$

where n is the number of trainees, *d* is the effect size of the training implemented, and SDy is the standard deviation of the trainees' job performance in dollars (Cascio, 1987). There are several important points to be made about this equation. The first is that *d*, the training effect size, is the same *d* that we discussed under training effectiveness. That is, utility of training has as one important parameter the effectiveness of the training -- everything else being equal, the utility, or payoff, of the training is greater as training effectiveness increases. Of course, the same point holds for standard ROI or NPV equations, but it is in equation (2) that the effect size index itself becomes clearly part of the utility calculation. This *d* may, in fact, be a mean *d* calculated from two or more studies. That is, it may be a meta-analytically derived *d*. A second, important point about the calculation of ΔU by equation (2) is that SDy, the standard deviation of job performance in dollars, is not necessarily easy to obtain. Some research indicates that using 40% of salary for the position in question is appropriate (Schmidt & Hunter 1983, Cascio, 1987), although this is subject to some debate. A third point about equation (2) is that it generates a point estimate for annual return on training. However, there is likely to be uncertainty associated, not with *n* perhaps, but with both *d* and SDy. This means that in order to estimate utility appropriately, some form of sensitivity analysis (to see how vulnerable estimates are to changes in values of parameters), breakeven analysis (to compare derived ΔU to the minimum value necessary to break even), or other kind of analysis incorporating uncertainty is important. Cascio (1987) presents a discussion combining NPV and ΔU to provide a breakeven analysis for a hypothetical training scenario.

Limitations to utility analysis

Focus on justification. In theory, current approaches to utility analysis should help decision makers by pointing out the financial implications of various training options. However, most approaches to understanding the utility of various personnel decisions have focused on providing a justification for decisions, rather than providing input for decision-making. The original Taylor-Russell tables (Taylor & Russell, 1939), for example, provided a means for psychologists and others to show that using a selection device with above-zero validity should provide a boost to average job performance over random selection in most cases. Development of more recent models of utility (e.g., cost-accounting) have generally had the same basic purpose -- to show that the impact of various human resource interventions are positive and substantial.

Recent reviews (e.g., Boudreau, 1991) have suggested that virtually any utility analysis method designed to output an index in terms of dollars saved, standard deviation in performance gained, and so forth will, almost invariably, show a substantial impact of virtually any intervention. One reason for the common finding of intervention effectiveness is that most utility analyses have examined fairly obvious issues. Does a selection device with a validity greater than zero demonstrate positive utility in comparison to random selection? If enough people are hired, in most cases the answer will be a clear "yes." Do most training interventions also yield a positive utility in comparison to a control training condition? In most cases, the answer is again, "yes." Thus, in both the selection and training literatures there is increasing evidence that some systematic intervention is better than nothing, or chance. For example, training has been shown to be generally effective in most cases (Burke & Day, 1986; Tannenbaum & Yukl, 1992; Kulik & Kulik, 1991). But the decisions that most organizations face are not training versus no training, or systematic selection versus random selection. Instead, the issue is more likely to be which of several training alternatives is the best choice given the specific training needs and organizational realities.

Focus on financial metrics. Furthermore, while financial costs and benefits are essential parameters to consider when making training decisions, there are other factors to consider such as the ease of deploying the training, the adaptability of the training for other needs, and the acceptability of the training. Traditional utility analyses have focused strictly on financial factors. In fact, the biggest problem with most utility models may be that they require a direct translation of performance into dollars. This has proven to be the Achilles' heel of utility analysis due to the many ways in which it can be conceptualized (e.g., Cascio, 1987; Cascio & Morris, 1990; Hunter, Schmidt, & Coggins, 1988). Moreover, utility in a dollar metric may be particularly difficult to generate in not-for-profit organizations like the military. So-called new approaches to traditional utility analysis (e.g., Raju, Burke, & Normand, 1990) do little to address this problem.

Focus on non-longitudinal utility. Finally, most utility analyses are based on one, or at the most two, waves of data collection. The problem with this is that although training will clearly have effects that may increase or decrease over time, traditional utility analysis is based on effectiveness studies that are a snapshot of a given time period.

Three conclusions of this section are that: a) training evaluations and utility analyses both need to move towards more of a decision-making framework, b) utility analysis, as traditionally practiced, needs to go beyond a straight cost-benefit perspective to incorporate other decision factors, and c) longitudinal issues in training utility need to be more seriously considered. As the next section will show, anticipatory training evaluation to some extent addresses each of these concerns.

VI. ANTICIPATORY TRAINING EVALUATION (MULTI-ATTRIBUTE UTILITY ANALYSIS)

Fortunately, a framework exists for considering tradeoffs among several alternative organizational strategies, to deal in non-financial metrics, and to address utility across time, namely Multi-Attribute Utility Analysis (MAUA; Brownlow & Watson, 1987; Edwards & Newman, 1986; Huber, 1980). MAUA is usually not referenced by those studying traditional I/O utility analysis. An exception is Boudreau (1991), who suggests that it is helpful to think of traditional I/O utility as a special case of MAUA. Boudreau argues that like MAUA, traditional utility analysis could be usefully adapted to the needs of decision makers. But rather than pursuing an MAUA approach, Boudreau studies and modifies traditional utility analysis to become more of a decision aid system. He does this by endorsing cost-accounting methods (Cronshaw & Alexander, 1985; 1991) and various methods to incorporate imperfect parameter certainty estimates into utility analysis. Boudreau rationalizes this approach of widening and honing traditional utility analysis by noting, correctly, that "little theoretical or empirical research has approached [traditional] utility analysis from this decision making perspective" (p. 625).

We propose a different strategic vision for developing a useful decision assistance methodology for the U.S. Air Force to meet its goals of evaluation and planning for the use of Intelligent Tutoring Systems (ITS). This strategy is to adapt MAUA to address the unique demands of this particular problem. This permits the use, when appropriate, of more traditional utility analyses as an input to a much larger decision-making context. But MAUA can function well even when traditional utility analysis approaches would not be applicable.

For example, consider the following scenario: the Air Force needs to decide on which of three training methods to use in a particular situation.

Traditional utility analysis would suggest the following steps: a) determine the effectiveness of the training (i.e., perform a training evaluation study or estimate the results); b) examine the costs of developing and delivering each training option; c) estimate the benefits from training in dollars; d) compute the payoff (costs - benefits) for each method in dollars; and e) choose the method with the highest payoff. Multi-Attribute Utility Analysis, on the other hand, would a) require careful stipulation of (probably multiple) criteria for making a decision; b) permit operationalization of several indices (attributes) which can characterize each criteria; c) permit differential weighting of both criteria and their attributes (perhaps based on previous training evaluation or utility studies); and d) suggest which alternatives are most favorable based on the criteria selected.

The difference between these methods is, therefore, that traditional utility analysis is not as well-equipped as Multi-Attribute Utility Analysis to act as a decision aid from beginning to end of the decision process. MAUA can handle many criteria and many aspects of each and it is not strictly a dollar-based approach. It can, however, readily incorporate findings from traditional cost-benefit analyses, training evaluation studies, and other data sources such as expert judgment.

In summary, Multi-Attribute Utility Analysis shows potential for application to important Air Force decision-making processes, such as those required for evaluations and planning of optimum placing of Intelligent Tutoring Systems. It will however require modification and customization to meet the needs of this decision context.

Description of the MAUA approach

The Multi-Attribute Utility approach (henceforth referred to as "MAUA") can be reasonably characterized as a procedure which has several steps. These are described below in a generic way. Following this, the application of MAUA to ITS research is described. What follows is our description of the MAUA approach, but adapted and changed to reflect our understanding of the needs of the military community for practical utility analysis that benefits decision making.

It should be emphasized that MAUA approaches do not make decisions or remove control from decision makers. MAUA methods possess sufficient flexibility that the characteristics of good decision makers (e.g., perceptiveness, communication skills, self-confidence, and creativity under stress; Shanteau, 1988) can have full play. At the same time it forces an appropriately weighted consideration of all aspects of the alternatives that are judged necessary as input to or consideration for a decision.

Steps of MAU

There are eight steps in our MAU process:

1. Define decision context.

2. Identify relevant decision factors.

3. Identify how the decision factors are to be measured (i.e., determine "attributes").

4. Determine proportion weights of the factors and attributes.

5. Determine attribute values for each decision alternative.

6. Identify any constraints and eliminate any unacceptable alternatives.

7. Calculate the relative favorability (i.e., "utility") of each alternative.

8. Determine the robustness of the conclusions.

Below we describe each of the MAUA steps in some detail.

1. Define decision context

The first step in MAUA is to stipulate the context of the decision to be made. This is important because the context will influence various parameter estimates used in making a decision. That is, without explicit consideration of context, effective generation of attributes and attribute values (described below) is difficult.

The *purpose* and *content* of the training is one such contextual variable. For example, the estimated effectiveness of an ITS would be different for training mechanical trouble-shooting to jet engine mechanics than training leadership skills to officers. Another contextual variable could be the *number of people* likely to be trained. The estimated development cost per trainee for an ITS would be lower if 1000 people will use the system than if only 100 will be trained.

A partial list of contextual variables that should be considered would include the following.

- Training options being considered

  As the number of options multiply, the decision process will necessarily be more complex and a systematic approach like MAU will be more important.

- Number of people likely to be trained

  The estimated costs of development and implementation per trainee will differ according to this number.

- Number of people per class or per tutor

  The estimated effectiveness of a method and/or the estimated costs will depend upon the number of trainees per class (or per system) who are to be taught.

- Readiness of the targeted trainees for learning the material

  If trainees are highly "ready" for a given course it may be predicted to be more effective for them than for trainees who have had less preparatory work. For example, if trainees are permitted to enter a training program only after they have successfully completed a set of rigorous prerequisite programs then the training can be expected to be effective, all other things held constant. Similarly, low trainee readiness may adversely affect the potential effectiveness of one training method more than another.

- Supportiveness of the job environment to training efforts

  It is becoming more and more evident (Tracey, Tannenbaum, & Kavanagh, 1995) that the climate for training transfer is a critical variable in training effectiveness. For example, if the transfer environment does not allow for immediate opportunities to practice, then the capability of a training method to encourage skill retention should increase in importance. In extreme cases, an unfavorable environment can place a ceiling on the potential effectiveness of any training option -- reducing the likelihood that a more expensive training alternative will be worth the investment.

- Objectives and content of the training

  As suggested earlier, training content can be important -- as in the example comparing leadership training versus training mechanical trouble-shooting. Results from Bennett (1995) suggest that the match between content of training and the method(s) used to deliver that content is critical to the observed effectiveness of the training. Similarly the objectives of training can be crucial to understanding decision context. For example, if training is designed as a broad-based introduction to a topic, appropriate indices of training effectiveness could be substantially different than those chosen for assessment of a program designed to proceduralize a highly complex and specific skill.

- Proportion of job that training will cover

  Sometimes a particular job training implementation, like an ITS system, will cover only a portion of some larger course, which itself only covers a part of the job. What proportion of the job proposed training will cover, or what proportion of current training will be replaced by a new training implementation, can be important considerations in understanding the decision context.

- Degree of scrutiny the training will receive

  This is a "political" variable. It can be important for decision makers to consider various organizational implications of their decision, such as how much scrutiny a program will receive. This may influence the appropriateness of trying a "risky" or cutting-edge alternative. It may further dictate the inclusion and measurement of additional criteria so that benefits can be measured and reported in terms meaningful to "people at the top." Stated another way, it may mean that both qualitative and quantitative methods and metrics are used.

2. Identify relevant decision factors

Decision factors are those aspects of the alternatives that decision-makers consider to be important or influential. For example, consider a decision about what type of training to deploy: ITS or traditional lecture training. Potential decision factors could include cost, ease of field deployment, degree of proven effectiveness, and unique capabilities. ITS can sometimes enable a certain aspect of a job to be trained which was not possible with other methods. For example, it might not be possible to place experts in all field locations where a particular complex skill is needed. ITS, however, might be deployed with the capability to teach expert mental models.

In identifying factors, it is important to poll experts carefully. The problem should be examined from all angles, so that all relevant factors are identified and characterized correctly. Factors will vary for each decision context (see Step 1). Portability of training, for example, may be an important factor for some training implementation decisions, but not others.

3. Identify how the decision factors are to be measured (i.e., determine "attributes")

An "attribute," in MAUA terms, is a measurable property which is used to assess how well a problem alternative fares on a decision factor. One or more attributes will relate to each factor. Actually, the "MA" in MAUA stands for "*multi*-attribute" because it is envisioned that most factors will have more than one attribute.

In terms familiar to I/O psychologists, an attribute is an "operationalization" of a decision factor. For example, the factors "ease of field deployment of training" might be operationalized as (be assigned the attributes) "time to set up training equipment" and "independence from given instructor(s)."

At this point in the process we are simply identifying the attributes. Step 5 is where we determine the attribute values for each alternative, e.g., how much time it takes to set up the equipment for training option A, option B, etc.

4. Determine proportion weights of the factors and attributes

Here each of the factors are first ranked in terms of their relative importance. Then the factor with the highest importance is assigned a raw value of 100, and each of the others are assigned values of importance between 0 and 100. These values reflect the relative importance of the various factors. Finally, these raw importance weights are converted to proportion weights by dividing each weight by the sum of the weights. Proportion weights will range from 0 to 1.0, with the sum of the proportion weights totaling 1.0.

These proportion factor weights are then used to calculate attribute proportion weights. For a single-attribute factor, we use the proportion weight for that factor as the proportion weight for the attribute. For multi-attribute factors, we first use the same procedure described above to

obtain proportion weights for the attributes: attributes are ranked within factors; the highest value within a factor is assigned a value of 100; the other attributes within that factor are assigned values between 0 and 100; these raw weights are converted to proportion weights within factors. Thus, at this stage each factor will have attribute weights in the range of 0 to 1.0; attributes weights within factors will sum to 1.0. Finally, proportion weights across factors are obtained by multiplying each within-factor attribute weight by the proportion weight of the factor it describes. When this process is complete, attribute proportion weights will sum to 1.0 across all factors.

5. Determine attribute values for each decision alternative

For each alternative we would then determine its relative "value" on each attribute. For example, in the case of "training system portability" (which is one attribute, or operationalization, of the decision factor "ease of field deployment of training"), values would be derived for each alternative.

Experts would rank the training alternatives in terms of their portability. Usually, the top ranked alternative is assigned a score of "100" for that attribute. Then, a structured process is used to determine relative values for the remaining alternatives. Generally, the expectation is that the values for a given attribute will be ordered linearly, but need not be a strict linear function. In fact, accelerated curving value estimates are common.

Attribute values can come from a number of sources . Frequently in MAUA, values are judged, that is, obtained from subject matter experts. Often when this is the case, value anchors can be used to help experts generate attribute values. For example, job performance values could first be considered in terms of whatever standard performance appraisal terms are currently used (e.g., "Excellent," "Good," "Fair," and "Poor"). Then values could be assigned to these, with "Excellent" being set at 100.

Attribute values may also be based on more objective data (e.g., the actual costs associated with each option) or based on historical information. For example, suppose that one decision factor is "training effectiveness" and one attribute of training effectiveness is "average improvement in test scores after training". Historical data on training effectiveness may be available for several training methods in the form of a meta-analysis (cf. Burke and Day, 1986; Kulik and Kulik, 1991; Bennett, 1995). If so, the attribute values could be based on the results of the meta-analysis. Alternatively, data from a specific training evaluation study may also be used to help determine the attribute values.

6. Identify any constraints and eliminate any unacceptable alternatives

Sometimes specific levels of certain attributes can be considered unacceptable, and a given alternative can be eliminated as soon as attributes are rated. For example, perhaps some degree of field deployment of a training system is required. In this case, any alternative for which the attribute "system portability" is extremely low (e.g., a full motion flight simulator) might be ruled

out. That is, a constraint on a level of an attribute has been imposed which eliminates an alternative.

Early elimination of alternatives via constraints is usually discouraged, because the idea behind an MAUA approach is compensability of factors. That is, for a given alternative, a low value on one factor may be compensated by a higher level on another. However, "reality" should rule, and sometimes decision makers should flatly reject an alternative at this stage.

7. Calculate the relative favorability (i.e., utility) of each alternative

The utility of each alternative is obtained by first multiplying the proportion weight for each attribute by its attribute value for a given alternative and then summing these. This should yield a score for each alternative. Alternatives with higher scores represent "better" alternatives given the assumptions captured by the approach. This does not mean that a decision has been made. It does however tell the decision maker(s) what is the "rational" choice *based on the judgments they provided.* Thus, a systematic process is enforced which will cause the decision maker(s) to seriously consider the choice which is most logical, given their own inputs.

There are several ways of comparing costs and benefits and depicting the relative utility of the alternatives in a graphical fashion at this point. These comparisons are designed to assist the decision makers in understanding the trade-offs among the various alternatives.

8. Determine the robustness of the conclusions

This step could also be labeled "sensitivity analysis." The goal is either to understand whether changes in *weights* assigned to factors or their attributes, or whether changes in *values* assigned to each attribute for each alternative would significantly change the relative favorability of the alternatives. Because there is usually some uncertainty connected with the weights, it is hoped that a moderate change in weights would not change the decision reached. As a matter of fact, research indicates substantial robustness for MAUA outcomes in many cases. As Goodwin & Wright (1991) point out, "Large changes in these figures are often required before one option becomes more attractive than another: a phenomenon referred to as 'flat maxima'" (pp. 25-26).

Robustness determination, or sensitivity analysis, can be done by hand but can easily be automated on computer. In the latter case, such useful approaches as "breakeven" analyses can be undertaken -- a breakeven point for a given factor, for example, would be when its value is changed to the extent of yielding a different final utility decision.

This MAUA approach, as outlined above, might be reasonably augmented or assisted by a number of "job aids." For example, a questionnaire could be developed to help decision makers frame a decision context (Step 1). A list of likely decision factors and attributes could provide the core decision tree for many decisions (Steps 2 and 3). Surveys could be administered to training professionals and those serviced by training to help identify decision factors (Step 2) and how to

measure and weight these (Steps 3 and 4). Monte Carlo studies might consider the relationship between cohort size and training effectiveness, helping gauge the values of training effectiveness attributes (Step 4) within a given context (Step 1). A list of results of relevant meta-analyses, reviews, and training evaluation studies could provide input for attribute values (Step 5). Worksheets for estimation of training (including ITS) development and implementation costs and identification of historical ITS development/implementation costs could be useful to decision makers when estimating costs (Step 5). Various graphical alternatives (along the lines of classical "utility curves") could assist decision makers in understanding relative alternative favorability (Step 7). Targeted training evaluation and effectiveness studies could provide values (Step 5) in those cases where no data were to be found. Computer programs could be written to assist in automating most of the MAUA steps, and permit immediate results and sensitivity analysis (Steps 7 and 8).

Hypothetical example using the MAU approach to determine utility

This section describes a hypothetical example of how MAUA could be used to compare several alternative training options. All the weights and values are purely hypothetical and only for illustration purposes.

1. Define decision context

Decision: Given ITS, CBT, field lecture-based training, and schoolhouse training, which method would be best for a course on C-130 engine trouble-shooting?

Context: The number of people to be trained will range from 100 to 2500 in the next three years; for initial decision purposes, the number of trainees is estimated at 1000. Prospective trainees will have completed basic training and been screened for interest and general readiness (so the effectiveness of the training will likely be higher than for a different context). The transfer environment does not always offer immediate opportunities to practice what they have learned in training (so retention becomes more important). In terms of organizational "politics," management is supportive of this training effort and substantial scrutiny is being given to this training. For the field lecture-based training the number of trainees per class is estimated at 10; for schoolhouse training the number of trainees per lecture is estimated at 25.

2. Identify relevant decision factors

A key to identifying decision factors is to avoid conditioned thinking, so that all relevant factors are identified. A group of SMEs could be convened, and asked to generate the list of factors. A structured elicitation method would be used, so that issues that may initially seem secondary or unimportant are nonetheless considered. The final list of factors might also be circulated among other SMEs later, with a request for feedback on completeness and accuracy.

For this training problem, assume that the factors generated were: a) financial costs of training; b) psychological costs; c) ease of field deployment; d) training effectiveness; e) scientific value; f) upgrading/adaptability; and g) acceptability.

3. Identify how the decision factors are to be measured (i.e., determine "attributes")

Attributes are, in effect, "operationalizations" or measurable aspects of factors. As in Step 2, this step usually involves a facilitated and structured knowledge elicitation approach. Multiple attributes per factor are permitted.

Figure 8 shows the hypothetical MAUA tree, with attributes subordinate to their factors. The factors are also shown under two traditional higher-order factors, Costs and Benefits.

Note that training the factor termed "training effectiveness" includes retention and transfer as two attributes. This emphasizes the ability of MAUA easily to incorporate longitudinal emphases of utility.

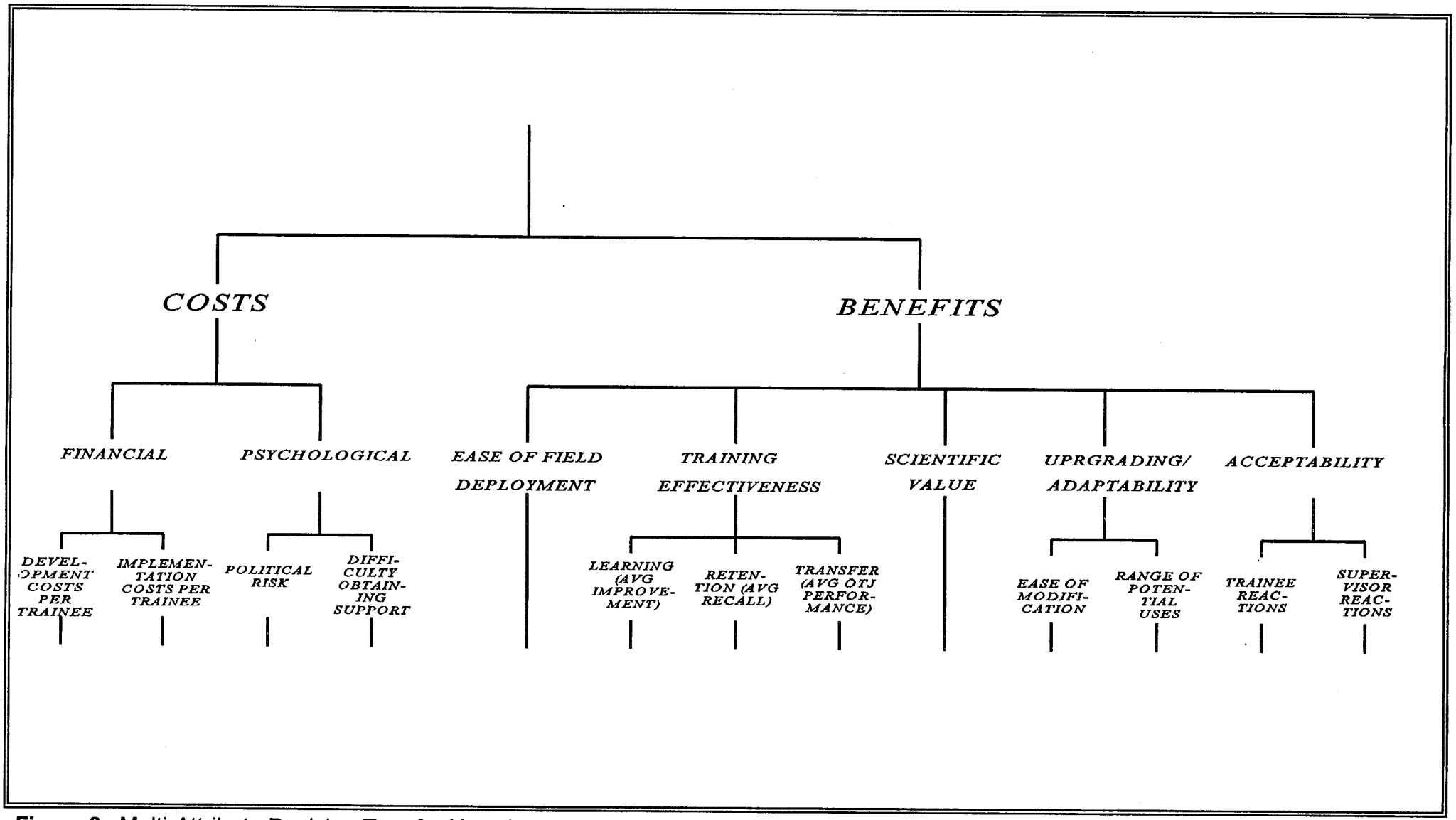


Figure 8. Multi-Attribute Decision Tree for Hypothetical Training Decision

4. Determine proportion weights of the factors and attributes

Factor weights. Presume that the seven factors are given an importance ranking of (high to low): Training Effectiveness, Financial Costs, Psychological Costs, Ease of Field Deployment, Acceptability, Upgrading, and Scientific Value. Effectiveness then gets a raw weight of 100, and expert judgments yield weights of, 90, 60, 55, 55, 40, and 25 for the remaining six factors. Proportion weights are then determined by dividing each raw weight by the sum of raw weights. So, for example, the proportion weight for Effectiveness is: 100/(100+90+60+55+55+40+25) or 100/425 = .24.

Attribute weights. As described in section B, for a single-attribute factor, we use the proportion weight for that factor as the proportion weight for the attribute. For multi-attribute factors, we obtain, in the manner described immediately above, proportion weights for the attributes within factors. Proportion weights across factors are then obtained by multiplying each within-factor attribute weight by the proportion weight of the factor it describes. So, for example, assume the attributes of training effectiveness are given within-factor values of 100 (Transfer), 85 (Retention), and 60 (Learning). These weights could reflect the context (Step 1) in that transfer is the most important aspect of learning for this task, while retention is weighted more importantly than immediate learning because decay has been shown to be an issue for this particular skill set.

The within-factor proportion weights for these attributes of training effectiveness are then transfer: 100/(100+85+60) = 100/245 = .41, retention: 85/245 = .35; and learning: 60/245 = .24. Then, each of these within-factor weights is multiplied times the factor proportion weight, which in this case is .24. The final across-factor proportion weight for these attributes is then: (.24)(.41) = .10, (.24)(.35) = .08, and (.24)(.24) = .06.

Figure 9 shows the hypothetical MAUA tree with the factor proportion weights, and within-factor and across-factor proportion weights for attributes.

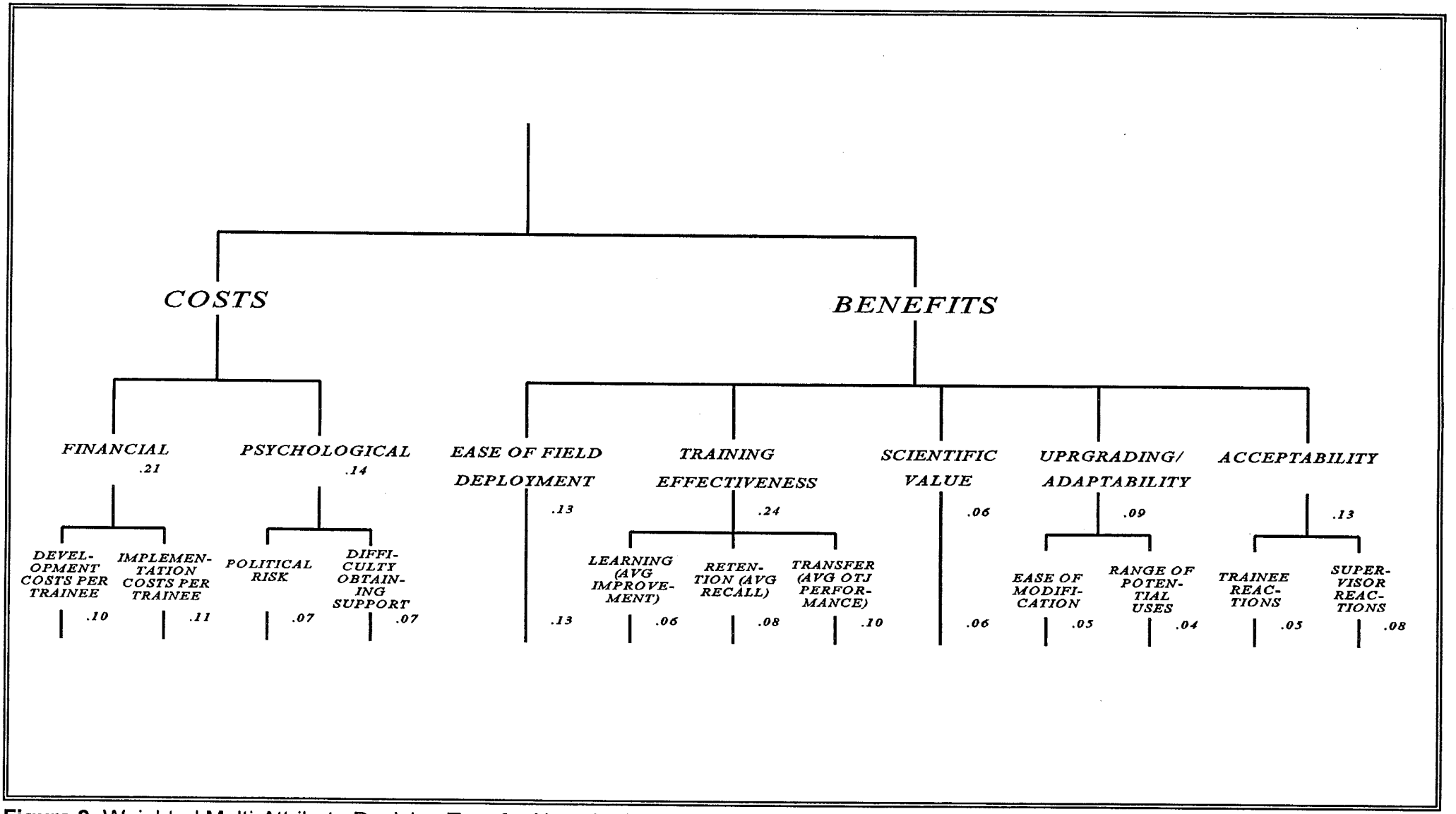


Figure 9. Weighted Multi-Attribute Decision Tree for Hypothetical Training Decision

Note that what has been captured to this point are the experts' opinion as to which factors should influence their decision and the relative importance of those *factors*. We have not discussed the relative advantages of any of the *alternatives* yet. Table 4 shows the hypothetical value data/anchors values assigned for each attribute for each alternative. Note that the lowest anchor for each attribute is assigned a zero. This is common practice in MAUA approaches. Assigning an above-zero value to the lowest anchor is also possible, however.

Table 4. Illustration of Factors, Attributes of Factors, and Values of Attributes

Factor	Attribute	Value Data/Anchors	Values
Financial Costs	Development Costs Per Trainee	0 - $1,000	100
		$1,000 - $2,000	95
		$2,000 - $5,000	80
		$5,000+	0
	Implementation Costs Per Trainee	$95 (ITS, CBT)	100
		$110 (Field-Based Lecture)	80
		$150 (School House Lecture)	0
Psychological Costs	Political Risk	None	100
		Moderate	80
		Great	0
	Difficulty Obtaining Support	None	100
		Slight	80
		Moderate	60
		Great	0
Ease of Field Deployment		Easy	100
		Fairly easy	70
		Moderately Difficult	40
		Difficult	0
Training Effectiveness	Learning (Average Immediate Posttraining Improvement, in SD units)	1+	100
		.5-1	80
		.1 -.5	40
		0	0
	Retention (Average Delayed Post-training Improvement)	.75+	100
		.4-.75	70
		.1 - .4	40
		0	0

Factor	Attribute	Value Data/Anchors	Values
Training Effectiveness (Continued)	Transfer (Average OTJ Performance)	Excellent	100
		Good	80
		Fair	50
		Poor	0
Scientific Value (potential for contributing to knowledge)		Excellent	100
		Good	90
		Fair	50
		Poor	0
Upgrading/ Adaptability	Ease of Modification	Very Easy	100
		Fairly Easy	70
		Moderately Hard	40
		Very Hard	0
	Range of Potential Uses	Wide Range	100
		Fairly Wide	90
		Moderately Wide	80
		Small Range	0
Acceptability	Trainee Reactions	Excellent	100
		Good	95
		Fair	90
		Poor	0
	Administrator Reactions	Excellent	100
		Good	80
		Fair	70
		Poor	0

5. Determine attribute values for each alternative

Having determined how each attribute should be weighted, the values of each alternative are obtained. Presume meta-analytic research indicates that the most effective training method (in terms of immediate learning) for mechanical trouble-shooting is ITS, followed by field lecture training and schoolhouse training, with CBT being the least effective. ITS would then get a value of 100 for this attribute, and suppose, based on the meta-analysis results and expert judgment, the others are given ratings of 80, 80 and 40, respectively. Similarly, assume that in anticipation of this decision we collected historical cost data for the four methods as they were used in previous Air Force training contexts. This information could be used when determining the values for the cost related attributes in the model.

Table 5 shows an alternatives X values table. As can be seen from this table each alternative has been rated in terms of its value for each attribute. The first numerical row of Table 5 shows the attribute weights.

Table 5. Illustration of Alternatives X Attribute Values Table

	Costs				Benefits								
	Financial		Psycho-logical		Ease of Field De-ploy-ment	Training Effectiveness			Sci-en-tific Value	Upgrading		Acceptability	
	Develop-ment Costs/ Trainee	Implemen tation Costs/ Trainee	Political Risk	Difficulty in Obtaining Support		Learning	Reten-tion	Transfer		Ease of Modifica-tion	Range of Potential Uses	Trainee Reac-tions	Super-visor Reac-tions
	0.10	**0.11**	**0.07**	**0.07**	**0.13**	**0.06**	**0.08**	**0.10**	**0.06**	**0.05**	**0.04**	**0.05**	**0.08**
ITS	100	100	80	60	100	100	100	100	100	100	100	100	100
CBT	95	100	80	60	100	40	40	80	50	100	80	90	80
Field-based	80	70	80	80	70	80	70	100	50	40	90	90	80
School house	80	30	100	80	0	80	70	80	50	40	80	90	70

6. Identify any constraints and eliminate any unacceptable alternatives

All alternatives are judged to be within theoretically acceptable ranges on all attributes.

7. Calculate the relative favorability (i.e., "utility") of each alternative

Finally, the relative favorability of each alternative is calculated. This is done by multiplying, for each alternative, its value by the attribute weight. The result will be a score between 0 and 100. Thus, the favorability of ITS is calculated by:

$$.11(100) + .10(100) + .07(60) + .07(100) + \ldots + .08(100).$$

Table 6 shows the alternative X attribute value table with results computed. Below the double line in Table 6 are the weighted value terms. The last column in this table shows the final summed "score" for each alternative (note too that the proportion weights for all attributes sum to 1.0).

In this hypothetical example, ITS has achieved the highest score: 85.8. This is the best choice, according the MAU process. The other alternatives were ranked as Field-based lecture, CBT, and Schoolhouse lecture, in that order.

8. Determine robustness of the conclusions

One way to determine how stable this ranking of alternatives is to submit the top choice to altered values. For example, we could change values for a given factor for ITS in Table 5 or 6 to the next lowest level. So, for example, we might wonder whether ITS would still score highest if its acceptability values were 80 rather than 100. It turns out that this would drop the final favorability value for ITS from 85.8 to 83.2 -- still ahead of the nearest competitor at 77.9. Similar "sensitivity" analyses can be performed for each factor or combination of factors.

Another type of sensitivity analysis examines the effects of changing weights, either at the factor or the attribute level. The general approach is the same as the method described above.

Table 6. Alternatives X Attribute Values Table with Results

	Costs				Benefits									
	Financial		Psychological		Ease of Field Deployment	Training Effectiveness			Scientific Value	Upgrading		Acceptability		
	Development Costs/ Trainee	Implementation Costs/ Trainee	Political Risk	Difficulty in Obtaining Support		Learning	Retention	Transfer		Ease of Modification	Range of Potential Uses	Trainee Reactions	Supervisor Reactions	Row Sum
	0.10	**0.11**	**0.07**	**0.07**	**0.13**	**0.06**	**0.08**	**0.10**	**0.06**	**0.05**	**0.04**	**0.05**	**0.08**	**1.00**
ITS	0	100	80	60	100	100	100	100	100	100	100	100	100	
CBT	80	100	80	60	100	40	40	80	50	100	80	90	80	
Field-based	100	70	80	80	70	80	70	100	50	40	90	90	80	
Schoolhouse	95	30	100	80	0	80	70	80	50	40	80	90	70	
ITS	0	11	5.6	4.2	13	6	8	10	6	5	4	5	8	**85.8**
CBT	8	11	5.6	4.2	13	2.4	3.2	8	3	5	3.2	4.5	6.4	**77.5**
Field-based	10	7.7	5.6	5.6	9.1	4.8	5.6	10	3	2	3.6	4.5	6.4	**77.9**
Schoolhouse	9.5	3.3	7	5.6	0	4.8	5.6	8	3	2	3.2	4.5	5.6	**62.1**

Summary

MAUA processes would appear to have good potential to meet the needs of the U. S. Air Force in decision-making for Intelligent Tutoring Systems. MAU can be adapted to include traditional utility analysis as one input. Other reasonable inputs are results of meta-analytic searches, traditional training studies, and surveys.

MAUA has the obvious strength of permitting decision makers to consider formally as many decision factors as they think necessary. Indeed, MAUA actually forces SMEs to consider all angles of a decision, including ones which they might not necessarily have considered. Its formal process, therefore, can be seen as one potential counter-balance to the shared stereotypes decision makers may have. At the same time it can be implemented in a fashion flexible enough to empower decision makers rather than "control" them, or to make decisions for them. Overall, MAUA should lead to efficient decision making and optimum decisions for planning and use of ITS.

VII. CONCLUSIONS

This technical report has considered three different yet related approaches to the evaluation of training. The first, training effectiveness evaluation, is in some sense the foundation for the other two, since it provides fundamental information regarding how well training is working. This in turn is input to the second approach to training evaluation: training utility analysis, or cost-to-benefit analysis. Third, Multi-Attribute Utility analysis was considered under the phrase "anticipatory training evaluation." This type of training evaluation may also use training effectiveness information, and even training utility information, as input. In this sense it bears an nested relationship to the first two types of training evaluation. But, as we have attempted to make clear, anticipatory training evaluation may also be previous or exterior to training effectiveness and utility evaluation. This is true because it is *anticipatory* -- it helps those charged with making decisions about training evaluation make those decisions. Thus, it is both the receiver of input from previous training evaluations and also informs training decisions and influences future and further evaluations. In this sense, anticipatory training evaluation is broader than simple training effectiveness evaluation and training utility analysis.

Thus, training effectiveness evaluation is tied clearly into the ISD process described in the first part of this report. It is tied in not only at the end of this process, but in the beginning and indeed throughout. As a result, the ISD process is not best conceived as representing a single endeavor such as the development of a single course. Instead, the ISD process, and with it the three approaches to training evaluation outlined here, is an ongoing process of continuous improvement in training.

REFERENCES

Alliger, G.M., & Scherer, G. (1995). The psychological impact of different ways to interpret experimental effect size. Manuscript submitted for publication.

Alliger, G.M., & Tannenbaum, S.I. (1995). A meta-analysis on the relations among training criteria. Paper presented in symposium, (M. Teachout, chair), Meta-analyses in training evaluation. American Psychological Association, New York.

Alliger, G.M., Tannenbaum, S.I., & Bennett, W. (1995). Transfer of training: Comparison of paradigms. Paper presented at the annual meeting of the Society of Industrial and Organizational Psychology, Orlando.

Alliger, G.M., & Janak, E. (1989). Kirkpatrick's levels of training criteria: Thirty years later. Personnel Psychology, 42, 331-342.

Baldwin, T.T., & Ford, J.K. (1988). Transfer of training: A review and directions for future research. Personnel Psychology, 41, 63-105.

Bennett, W., Jr. (1995). A meta-analytic review of factors that influence the effectiveness of training in organizations. Unpublished doctoral dissertation.

Boudreau, J.W. (1991). Utility analysis for decisions in human resource management. In M.D. Dunette & L. Hough (Ed.), Handbook of Industrial and Organizational Psychology, Second Edition, Volume 2. Palo Alto: Consulting Psychologists Press.

Brogden, H.E., & Taylor, E.K. (1950). The dollar criterion - Applying the cost accounting concept to criterion construction. PP, 3, 133154.

Brownlow, 0., & Watson, R. (1987). Structuring multi-attribute value hierarchies. Journal of the Operational Research Society, 38, 309-317.

Burke, M.J., & Day, R.R. (1986). A cumulative study of the effectiveness of managerial training. Journal of Applied Psychology, 71, 232-245.

Burns, H., & Partlett, J.W. (1991). The evolution of intelligent tutoring systems: Dimensions of design. In H. Burns, J.W. Partlett, & C.L. Redfield (Eds.), Intelligent Tutoring Systems: Evolutions in Design, Hillsdale, NJ: Erlbaum.

Cascio, W.F. (1991). Applied Psychology in Personnel Management. Englewood Cliffs, NJ: Prentice Hall.

Cascio, W.F. (1987). Costing human resources: The financial impact of behavior in organizations (2nd ed.). Boston: Kent.

Cascio, W.F., & Morris, J.R. (1990). A critical reanalysis of Hunter, Schmidt, and Coggin's (1988) "Problems and pitfalls in using capital budgeting and financial accounting techniques in assessing the utility of personnel programs.. Journal of Applied Psychology, 75, 410-417.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences. New York: Academic Press.

Cook, T.D., & Campbell, D.T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Chicago: Rand McNally.

Cronshaw, S.F., & Alexander, R.A. (1985). One answer to the demand for accountability: Selection utility as an investment decision. Organizational Behavior and Human Decision Making, 35, 102-118.

Cronshaw, S.F., Alexander, R.A. (1991). Why capital budgeting techniques are suited for assessing the utility of personnel programs: A reply to Hunter, Schmidt, and Coggin (1988). Journal of Applied Psychology, 76, 454-457.

Detterman, D.K. (1993). The case for the prosecution: Transfer as an epiphenomenon. In D.K. Detterman & R.J. Sternberg (Eds.), Transfer on trial: Intelligence, cognition, and instruction (pp. 1-24). Norwood, NJ: Ablex Publishing.

Edwards, W., & Newman, J.R. (1986). Multiattribute evaluation. In H.R. Arkes & K.R. Hammond (Eds.), Judgment and Decision Making, Cambridge: Cambridge University Press.

Ford, J.K., & Wroten, S.P. (1984). Introducing new methods for conducting training evaluation and for linking training evaluation to program design. Personnel Psychology, 37, 651-665.

Ford, J.K., Quinones, M.A., Sego, D.J., & Sorra, J.S. (1992). Factors affecting the opportunity to perform trained tasks on the job. Personnel Psychology, 45, 511-527.

Goodwin, P., & Wright, G. (1991). Decision analysis for managerial judgment. New York: John Wiley.

Hedges, L.V., & Olkin, I. (1985). Statistical methods for meta-analysis. Orlando: Academic Press.

Huber, G.P. (1980). Managerial decision making. Dallas, TX: Scott, Foresman, & Co.

Hunter, J.E., Schmidt, F.L., & Coggins, T.D. (1988). Problems and pitfalls in using capital budgeting and financial accounting techniques in assessing the utility of personnel programs. Journal of Applied Psychology, 73, 522-528.

Kirkpatrick, D.L. (1959a). Techniques for evaluating training programs. Journal of ASTD, 13, 3-9.

Kirkpatrick, D.L. (1959b). Techniques for evaluating training programs: Part 2 - Learning. Journal of ASTD, 13, 21-26.

Kirkpatrick, D.L. (1960a). Techniques for evaluating training programs: Part 3 - Behavior. Journal of ASTD, 14, 13-18.

Kirkpatrick, D.L. (1960b). Techniques for evaluating training programs: Part 4 - Results. Journal of ASTD, 14, 28-32.

Kulik, C.C., & Kulik, J.A. (1991). Effectiveness of computer-based instruction: An updated analysis. Computers in Human Behavior, 7, 75-94.

Mathieu, J.E., & Leonard, R.L. (1987). An application of utility concepts to a supervisor skills training program: A time-based approach. Academy of Management Journal, 30, 316-335.

Maxwell, S., & Howard, G. (1981). Change scores -- necessarily ANATHEMA? Educational and Psychological Measurement, 41, 747-756.

McGehee, W., & Thayer, P.W. (1961). Training in business and industry. New York: Wiley.

McGraw, K.O., & Wong, S.P. (1992). A common language effect size statistic. Psychological Bulletin, 111, 361-365.

Raju, N.S., Burke, M.J., & Normand, J. (1990). A new approach for utility analysis. Journal of Applied Psychology, 75, 3-12.

Rogosa, D. (1988). Myths about longitudinal research. In K. W. Schaie (Ed.), Methodological issues in aging research. New York: Springer.

Rogosa, D.R., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. Psychological Bulletin, 92, 726-748.

Rosenthal, R., & Rubin, D.B. (1982). A simple, general purpose display of effect size magnitude. Journal of Educational Psychology, 74, 166-169.

Schmidt, F.L., & Hunter, J.E. (1983). Individual differences in productivity: An empirical test of estimates derived from studies of selection procedure utility. Journal of Applied Psychology, 68, 407-414. Schmidt,

Shanteau, J. (1988). Psychological characteristics and strategies of expert decision makers. 11th Conference on: Subjective probability, utility and decision making (1987, Cambridge, England). Acta Psychologica, 68, 203-215.

Tannenbaum, S.I., Mathieu, J.E., Salas, E., & Cannon-Bowers, J.A. (1991). Meeting trainees' expectations: The influence of training fulfillment on the development of commitment, self-efficacy, and motivation. Journal of Applied Psychology, 76, 759-769.

Tannenbaum, S.I., & Woods, S.B. (1992). Determining a strategy for evaluating training: Operating with organizational constraints. Human Resource Planning, 15, 63-81.

Tannenbaum, S.I., & Yukl, G. (1992). Training and development in work organizations. Annual Review of Psychology, 43, 399-441.

Taylor, H.C., & Russell, J.T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. Journal of Applied Psychology, 23, 565-578.

Teachout, M.S., Sego, D.J., & Ford, J.K. (1995a). Extending the Training Efficiency and Effectiveness Methodology (TEEM) with Training Transfer Data. Brooks AFB, TX: Technical Training Research Division, Armstrong Laboratory, Human Resources Directorate.

Teachout, M.S., Sego, D.J., & Ford, J.K. (June, 1995b). Application of the Training Efficiency and Effectiveness Methodology (TEEM) to Aerospace Ground Equipment technical training (AL/HR-TP-95-0013). Brooks AFB, TX: Technical Training Research Division, Armstrong Laboratory.

Teachout, M.S., Sego, D.J., & Olea, M.M. (November, 1993). Assessing training efficiency and effectiveness for Aerospace Ground Equipment training. Proceedings of the Military Testing Association (445-450), Williamsburg, VA.

Teachout, M.S., Olea, M.M., Phalen, W.P., & Barham, B.S. (November, 1993). Improving the validity and efficiency of Aerospace Physiology Instructor Training. Proceedings of the Military Testing Association (451-455), Williamsburg, VA.

Thorndike, R.L. (1947). Research problems and techniques (Report No. 3) AAF Aviation Psychology Program Research Reports, U.S. Government Printing Office.

Tracey, J.B., Tannenbaum, S.I., & Kavanagh, M.J. (1995). Applying trained skills on the job: The importance of the work environment. Journal of Applied Psychology, 80, 239-252.

Warr, P., & Bunce, D. (1995). Trainee characteristics and the outcomes of open learning. Personnel Psychology, 48, 347-375.