DOCUMENT RESUME

ED 255 577 TM 850 221

AUTHOR Tufano, Daniel R.; Evans, Robert A.
TITLE The Prediction of Training Device Effectiveness: A Review of Army Models.
INSTITUTION Army Research Inst. for the Behavioral and Social Sciences, Alexandria, Va.
REPORT NO ARI-TR-613
PUB DATE Apr 82
NOTE 80p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS *Armed Forces; Evaluation Methods; Mathematical Models; Postsecondary Education; *Predictive Validity; *Program Effectiveness; Program Evaluation; Simulation; Statistical Analysis; *Training Methods; Transfer of Training
IDENTIFIERS TRAINVICE Models

ABSTRACT

This is a critical review of the analytical models developed by the U.S. Army to predict training device effectiveness. The Simulation Systems Technical Area of the U.S. Army Research Institute for the Behavioral and Social Sciences performed the research. Special consideration is given to a family of models known collectively as TRAINVICE. These four models differ in several important ways, although they are all intended to provide an index of transfer potential. The variables considered in the calculation of the indices are assigned different degrees of emphasis in each model. In selected cases a variable is completely ignored: e.g., the instructional features of a training device. The preparatory analyses employed to estimate the values of each variable change from model to model. The combinatorial procedures used to calculate an index of effectiveness from the variable values are different in each model. The principal points of difference among the various predictive models are presented to facilitate a comparative assessment of their merit. Results of the analysis and subsequent validation will provide a basis for the selection and refinement of one or more of the models for incorporation into a defined set of procedures for specifying training device effectiveness. (Author/DWH)

Technical Report 613

The Prediction of Training Device Effectiveness: A Review of Army Models

Daniel R. Tufano and Robert A. Evans

Simulation Systems Technical Area
Training Research Laboratory

U. S. Army
Research Institute for the Behavioral and Social Sciences

April 1982

Approved for public release; distribution unlimited.

U. S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the Deputy Chief of Staff for Personnel

JOSEPH ZEIDNER
Technical Director

L. NEALE COSBY
Colonel, IN
Commander

NOTICES

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-TST, 5001 Eisenhower Avenue, Alexandria, Virginia 22333.

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Technical Report 613	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) The Prediction of Training Device Effectiveness: A Review of Army Models		5. TYPE OF REPORT & PERIOD COVERED -- 6. PERFORMING ORG. REPORT NUMBER --
7. AUTHOR(s) Daniel R. Tufano Robert A. Evans		8. CONTRACT OR GRANT NUMBER(s) --
9. PERFORMING ORGANIZATION NAME AND ADDRESS US Army Research Institute for the Behavioral & Social Sciences (PERI-IE) 5001 Eisenhower Ave., Alexandria, VA 22333-5600		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 2Q263744A795
11. CONTROLLING OFFICE NAME AND ADDRESS US Army Research Institute for the Behavioral & Social Sciences (PERI-IE) 5001 Eisenhower Ave., Alexandria, VA 22333-5600		12. REPORT DATE April 1982 13. NUMBER OF PAGES 64
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) --		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE --

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

--

18. SUPPLEMENTARY NOTES

--

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Training Device
Effectiveness
Transfer of Training
Prediction
Analytic Model
TRAINVICE

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report is a critical review of the analytic models which have been developed by the Army to predict training device effectiveness. In particular, special consideration is given to a family of models known collectively as TRAINVICE. Although they are all intended to provide an index of transfer potential, these four models differ in several important ways. The variables considered in the calculation of the indices are given different degrees of emphasis (or weight) in each model. In fact, in some cases a variable is ignored completely; e.g., the instructional features of a training device. (over

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

i SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

The preparatory analyses employed to estimate the values of each variable change a great deal from model to model. Moreover, the combinatorial procedures used to calculate an index of effectiveness from the variable values are very different in each model.

A major goal of this report is to present the principle points of difference among the various predictive models, and thus facilitate a comparative assessment of their merit. The results of the analysis and subsequent validation will provide a basis for the selection/refinement of one or more of the models for incorporation into a defined set of procedures for specifying training device effectiveness.

Technical Report 613

The Prediction of Training Device Effectiveness: A Review of Army Models

Daniel R. Tufano and Robert A. Evans

Submitted by:
Frank J. Harris, Chief
Simulation Systems Technical Area
Training Research Laboratory

Approved by:
Joseph Zeidner
Technical Director

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES
5001 Eisenhower Avenue, Alexandria, Virginia 22333

Office, Deputy Chief of Staff for Personnel
Department of the Army

April 1982

Army Project Number
2Q263744A795

Training and Simulation

Approved for public release; distribution unlimited.

ARI Research Reports and Technical Reports are intended for sponsors of R&D tasks and for other research and military agencies. Any findings ready for implementation at the time of publication are presented in the last part of the Brief. Upon completion of a major phase of the task, formal recommendations for official action normally are conveyed to appropriate military agencies by briefing or Disposition Form.

FOREWORD

The Simulation Systems Technical Area of the Army Research Institute for the Behavioral and Social Sciences (ARI) performs research and development in the areas of training devices and simulators in the Army. Of special interest is research concerning the evaluation of training device effectiveness.

Throughout the acquisition of a simulator or training device, training effectiveness must be evaluated. Ideally, an empirical transfer of training test would provide the data needed for an evaluation. However, when empirical data cannot be obtained, training device effectiveness can only be estimated using analytic methods.

This report provides a critical review of analytic methods recently developed by the Army for the evaluation of training device effectiveness. The results of this report have implications for training developers in PM TRADE and TRADOC and for researchers in the field of training device effectiveness.

JOSEPH ZEIDNER
Technical Director

THE PREDICTION OF TRAINING DEVICE EFFECTIVENESS: A REVIEW OF ARMY MODELS

EXECUTIVE SUMMARY

Requirement:

To review the analytic models and methods developed by the Army for the prediction of training device effectiveness; and to recommend procedures for the development, validation and application of improved models.

Procedure:

Four predictive models, known collectively as TRAINVICE, were compared in terms of their implicit assumptions, analytic procedures, validity, and utility for training-device acquisition.

Findings:

Despite their common purpose, the four TRAINVICE models differ considerably in: the task, equipment, and personnel variables; and the mathematical formulae used to calculate training effectiveness indices. The major limitation shared by all of the TRAINVICE models is that they yield overall indices of effectiveness. The utility of such an index is strongly questioned. The recommendation was made that a model be developed which would permit a more detailed assessment of training device effectiveness. Ideally, such a model would generate effectiveness indices for individual skills, and would provide procedures for aggregating the skill indices into separate task indices. It was concluded that separate skill and task indices would yield effectiveness predictions of sufficient detail to be of use to the training developer in the design, evaluation, and implementation of training devices.

Model application and development will require research to be done in two areas: field validation of the TRAINVICE models in various task domains; and longer range investigation of the models' underlying assumptions. The latter area should include a refinement of the learning guidelines contained in the models, and specification of behavioral criteria which are suitable to analytic as well as empirical evaluation of training device effectiveness.

Utilization of Findings:

The review and recommendations will be of use to the training developer wishing to use one of the existing TRAINVICE models as well as to the model developer trying to improve the prediction of training effectiveness.

THE PREDICTION OF TRAINING DEVICE EFFECTIVENESS:
A REVIEW OF ARMY MODELS

CONTENTS

LIST OF TABLES

I

Introduction

This report examines analytic methods and models for the evaluation of training device effectiveness. The need for such non-empirical evaluation procedures has been a persistent concern of military training developers since the 1960's. In particular, the Army has recently developed a series of models, known as TRAINVICE, which attempt to predict the degree to which training on a particular training device will transfer to performance on operational equipment. These models, which have evolved from a history of military training research, are the principle focus of this paper.

Jeantheau (1971) reported an attempt by the Navy at the "qualitative assessment" of training device effectiveness. The forms and guidance included in this document permit the cataloging of training device features and expert opinions on those features. These procedures do not, however, result directly in the evaluation of a particular training device. Rather, the method simply provides a format for collecting and using information on training devices.

In a later effort, done for the Army, Caro (1970) developed the Task Commonality Analysis (TCA) method for the prediction of transfer of training from a device to operational equipment. The predictions were based on Realism ratings[1] of the stimulus (display) and response (control) properties of the training device. In deducing which tasks would be trained well (i.e., high transfer) and which would not, Caro adhered to Osgood's principles of transfer. He assumed that if both the stimuli and responses in the training situation were similar to those in the operational situation, then positive transfer would result. Further, he assumed that if the stimuli were similar, but the associated responses were different, then negative transfer would occur. Caro's choice is not surprising since these assumptions are ubiquitous in the field of training evaluation and are well represented in the TRAINVICE models which are discussed below.

Caro's TCA method represented the state of the art when it was published. It provided the impetus and much of the groundwork for the development of the TRAINVICE models. Although TCA is similar to the TRAINVICE models in its goal and in some of its assumptions, it will not be treated more fully here for the following reasons. The realism ratings were rudimentary (see footnote 1) and were not based on clearly articulated criteria. The transfer predictions consist of simple, qualitative statements about whether or not a task will be trained well. Furthermore, the judgmental operations required to generate the predictions have not been reduced to a formal algorithm. That is, there are no fixed procedures for transforming or combining data to arrive at a clear prediction.

---

1 Raters were simply asked to judge whether or not a display or control was "realistic". The realism score for a piece of equipment was the percentage of raters who said that that piece of training equipment was "realistic".

It was not until 1976 that the shortcomings of earlier approaches were addressed. Between this time and 1980, the Army developed a family of predictive models known collectively as TRAINVICE. In their attempt to generate quantitative predictions of effectiveness through formal procedures these four models represent the most ambitious steps taken to date in the field of analytic evaluation. The level of sophistication and the potentially great utility of these models warrant a very close examination of the procedures, assumptions and validity of TRAINVICE.

The original method, developed in 1976, is referred to as TRAINVICE-A (TV-A) in this report (Wheaton, Fingerman, Rose, and Leonard, 1976). In 1979, the Honeywell Corporation modified TV-A as part of an effort to develop detailed guidance for user application (PM-TRADE, 1979). This modified approach is referred to as TRAINVICE-B (TV-B) in this report. Other modifications to TV-A were developed by the U.S. Army Research Institute (Narva, 1979a, 1979b) and are reported herein as TRAINVICE-C (TV-C). Finally, in an effort to develop a user guidebook for applying TV-C, additional revisions were made. (Swezey and Evans, 1980) This approach is referred to as TRAINVICE-D (TV-D) in this report.

Although each model purports to provide an index of effectiveness, or transfer of training potential for a device, these models differ in several important ways. For example, the variables considered in the calculation of the indices are given different degrees of emphasis or mathematical weight in each model. The procedures used to estimate the values for each component vary considerably from model to model. Moreover, the procedures used to calculate an index of effectiveness from the variable values are also very different in each model.

The TRAINVICE models do, however, share a common data collection method. This method consists of a structured interrogation of a subject matter expert. As such the models place a very high premium on the judgment of an expert. The method focuses decision-making on a specific set of issues for each task or part of a task. In the first of the TRAINVICE models, for example, one of the issues considered is the similarity between the equipment on a training device and that on the operational equipment to perform a particular subtask. This issue is further delineated into physical similarity (appearance, location, etc.) and functional similarity (amount of information flow between the human operator and the controls and displays). For each of these, (i.e. physical and functional similarities) the expert assigns values from a rating scale which ranges from 0 to 3. Guidance is provided by a description of the criteria associated with each value, (e.g., a "3" means identical to operational equipment). This procedure continues until all equipment (i.e. displays/controls) associated with all sub-tasks have been rated. An analogous rating procedure is performed for all variables in the model pertaining to each subtask. In this manner, the subject matter expert can estimate numerical values for each predictor variable considered by the model (e.g., similarity, training techniques, task difficulty, etc.). These estimated values for the variables are then entered into a general formula which results in an overall figure of merit (index of training effectiveness) for the training device in question.

The judgments of the subject matter expert and the index of effectiveness ly on many assumptions, both theoretical and mathematical in nature. The

theoretical assumptions include: a) what is being predicted (e.g., a particular measure of transfer of training); and b) which task and equipment variables have the predictive power to generate such a measure of effectiveness. The mathematical assumptions concern: a) the manner in which all the values are combined (e.g., weighting strategies, etc); and b) the numerical properties of the rating scales used to estimate those values. As indicated earlier, the four analytic models reviewed in this report differ considerably in the assumptions made and in the forms in which the assumptions are manifested.

Section II of this report contains a detailed description of each model, taken individually. Section III is a general summary and critique of all four models, in which differences among the models are discussed in detail. Finally, based upon the results of the critical review, future directions are discussed in Section IV.

II
Models

TRAINVICE-A (TV-A)

Overview

The Wheaton, Fingerman, Rose, and Leonard (1976) approach, TV-A, is an attempt to predict and evaluate training device effectiveness, specifically transfer of skills from training to operational settings, by combining judgments about a variety of factors. Judgments are transformed into values related to the interactions among device design and use, trainee ability, and training strategy. Effectiveness, therefore, is assumed to be a function of the:

1. Transfer Potential: potential for transfer of training using a particular device which is determined by the:

    a. overlap or communality of the skills taught on a device and those necessary to perform on the operational equipment, and

    b. physical and functional similarity between a device and the operational equipment

2. Learning Deficit: differences (i.e., deficits) between a trainee's knowledge before training on a device and what must be known about the operational equipment, weighted by the difficulty of acquiring such knowledge, and;

3. Training Techniques: appropriateness of training techniques or device features incorporated into a device, and how well these features adhere to accepted principles of learning.

The Wheaton, Fingerman, Rose, and Leonard (1976) model combines values of judgments made for each of the above factors into an index reflecting the effectiveness of a device.

Process

Generating forecasts about the effectiveness of a training device, according to Wheaton, Fingerman, Rose, and Leonard (1976) requires analyses of the components: Transfer Potential, Learning Deficit, and Training Techniques for a training device. These components are subsumed under one of the three categories in the structural and functional model of this training device effectiveness approach. The process of analyzing each of the components (i.e., Transfer Potential, Learning Deficit, Training Technique) requires judgments[2] to be made for five basic analyses of:

---

2 See Appendix for the rating scales.

1. Task Communality
2. Physical Similarity
3. Functional Similarity
4. Learning Deficit
5. Training Technique

Values for these analyses are derived for a device under evaluation, then where appropriate, compared to the operational equipment for which a device was developed.

Inputs

Before performing the procedures by which values of the TV-A variables are estimated, a user needs a list of training objectives and relevant task analytic information for both an operational setting and a training device being evaluated. TV-A requires that most analyses be conducted at the subtask level and some analyses at the level of the skills and knowledges which comprise each subtask. A Subtask, according to Folley's (1964) definition is ". . . an activity that is performed by one person and bounded by two events" (Wheaton, Fingerman, Rose, and Leonard, 1976, p. 16). The value of each TV-A variable is estimated for each subtask identified in the operational task analysis.

Procedures

Task Communality Analysis (C)

Task Communality Analysis (C) assesses the overlap between training subtasks and those in the operational equipment. The value of C is determined by comparing operational and training device task analyses with each other. In this procedure, a training device is given a rating of "1" for each "i" operational subtask it covers or "0" for those it fails to cover. Since a C_i value of "0" decreases the sum in the numerator of the final prediction formula, the task communality rating serves to penalize a training device for each operational subtask not covered. A training device is not, however, penalized for including subtasks which are not in the operational environment (i.e., additional subtasks).

In the overall device effectiveness prediction formula, the sum of C values for a device is compared to C_i values for subtasks on the operational equipment. Since this comparison is made against the operational equipment, C_i always = "1" for the operational equipment.

Physical Similarity Analysis (PSA)

The Physical Similarity Analysis (PSA), and the Functional Similarity Analysis (FSA), discussed below, combine to form the Similarity (S) component in the predictive equation. The degree of similarity (S_i) between a training device and the operational equipment is the average of values assigned to the fidelity variables: physical and functional similarity. The PSA allows for judgments concerning the physical characteristics (i.e., appearance, size,

location, etc.) of displays and controls used in training specific behavioral performance on a device. The Functional Similarity Analysis is concerned with the information processing activities of the human who is viewing the displays and operating the controls.

The information required to perform the PSA is a list of all displays and controls on the operational and training equipment relevant to each subtask. The displays and controls corresponding to each subtask are given a rating by judging how well the operational equipment is represented in a training device: from "0" (not represented) to "3" (identical to operational equipment). These ratings are averaged across controls and displays and divided by "3" to yield a physical similarity index ranging between "0" and "1".

Like the C analysis, the PSA for the operational equipment is "1". The rationale is that the operational equipment represents the maximum degree of physical similarity. A PSA value of "1", therefore is assigned to all displays and controls on the operational equipment corresponding to the subtasks.

Functional Similarity Analysis (FSA)

Like the PSA, performing the Functional Similarity Analysis (FSA) requires a list of operational subtasks and corresponding displays and controls. A flow diagram for each subtask is then generated indicating the [illegible], amount, and direction of information to and from the operator for each control and display. The amount of information (in "bits") is determined by the number of stimulus (i.e. information transmitted from a display to an operator) and response (i.e. information transmitted from an operator to a control) states which displays or controls can assume. The remainder of this analysis consists of rating differences between the amount of information in an operation setting (H_{OS}) and that in a training setting (H_{TS}).[3] For each control and display, a training device is given a rating: from "0" (missing) to "3" (identical: $H_{OS} = H_{TS}$). Ratings for controls and displays are then averaged and divided by "3" to give a functional similarity index for each subtask which ranges between "0" and "1". The overall similarity index for each subtask (S_i) is the average of the Physical and Functional Similarity Indices $\frac{(P + F)}{2}$.

The FSA for each display and control on operational equipment, similar to other analyses discussed thus far, is always a "1".

Learning Deficit Analysis (D)

The Learning Deficit(D) index, for every subtask, requires each skill and knowledge be given two ratings (rating scales adapted from Demaree, 1961; see

---

3 The FSA analysis requires the rater to compare amounts of information in $\log_2$ units; a potentially difficult procedure for users who are unfamiliar with information theory. This shortcoming was addressed in TV-B's revision of the Functional Similarity Analysis.

Appendix). The first rating, repertory scale (RS), assesses the degree to which trainees are already proficient in the skills and knowledges to be taught. A rating from "0" (no experience) to "4" (complete understanding) is assigned to each skill and knowledge. An estimate is then made of the levels of proficiency required of a trainee, for each skill and knowledge, in order to perform a particular subtask to criterion. Accordingly, a criterion scale (CS) value is assigned to each skill and knowledge: from "0" (no experience) to "4" (complete understanding). The criterion scale value minus the repertory scale value (CS-RS), then represents the learning deficit for each skill and knowledge. The learning deficit index (LD) for each subtask is simply the average of the learning deficit values of all skills and knowledges involved:

$$LD = \frac{\sum_{i=1}^{S+K} CS_i - RS_i}{\#\text{ skills and knowledges}}$$

LD ranges between "0" and "4".

The Learning Deficit value for each subtask is then weighted by the difficulty of training the skills and knowledges necessary for that subtask (i.e., how hard it is to overcome the learning deficit). To do this, each subtask is ranked according to the amount of time required to train that subtask on the operational equipment (a rank of "1" for the easiest subtask; higher ranks for subtasks requiring more training time). The learning deficit value for each subtask is multiplied by its rank, then divided by "4" times the total number of subtasks. This procedure yields a weighted learning deficit value (D_i) for each subtask which ranges between "0" and "1". A D value is computed once for the operational subtasks because these values are applicable to both a training device and operational equipment.

Training Techniques Analysis (T)

In the Training Techniques (T) analysis a training device is rated on how well it implements established learning principles. The first step is to assign one or more task taxonomic labels, (after US Naval Training Device Center, 1972), to each operational subtask, using the skills and knowledges comprising each subtask. Associated with each of the thirteen task categories in the taxonomy are three sets of learning principles which are related to stimulus, response, and feedback aspects of these tasks (after Willis and Peterson, 1961; and Micheli, 1972). For each subtask, ratings are given on how well a training device implements each of the relevant learning principles: "-3" (complete violation of principle); "0" principle not implemented or violated); "3" (optimal implementation of principle). The lowest ratings given to learning principles in each category (i.e., stimulus, response, and feedback) are then averaged, to yield a T score for each subtask. In order to scale T down to between "0" and "1", "3" is added to the averaged score, and the sum is divided by "6".

As Wheaton, Fingerman, Rose, and Leonard (1976) pointed out, the determination of T values is rather conservative since only the poorest implementation of training techniques on a device is considered. In the TV-A procedure, a training device does not get credit for having a few, especially good instructional features.

The operational equipment is assumed to make optimal use of training techniques, therefore, $T_i = 1$.

Outputs

Indices

Each of the TV-A analyses can be calculated and collapsed across subtasks to derive a separate index purporting to assess Transfer Potential, Learning Deficit or Training Techniques. Calculation of these indices may serve as a diagnostic function to locate deficiencies or assets in a training device. For a detailed discussion on these indices, the reader is referred to Research Memorandum 76-16 (Wheaton, Fingerman, Rose, and Leonard, 1976). Such a presentation is beyond the scope of this effort.

Overall Device Effectiveness Prediction

The developers of the TV-A model tried to predict the Gagne, Foster, and Crowley (1948) measure of transfer:

$$\tau = \frac{C - E}{C}$$

In this classic transfer of training paradigm, both C and E are measures of practice (time, trials, errors) required on operational equipment, in order to meet a performance criterion. C represents a control group, which practiced only on operational equipment. E represents an experimental group, which practiced on a simulator or training device first, then transferred to operational equipment. The question which this transfer equation attempts to answer is: How much training time (i.e., on operational equipment) can be saved by providing practice on a simulator?

τ is, therefore, a measure of savings. It equals the amount of training time on operational equipment saved by practicing on a simulator first (C-E), as a proportion of training time required when operational equipment alone is used (C). τ varies between $-\infty$ and +1. In theory, the closer τ is to +1, the greater the transfer of skills acquired with a simulator to operational equipment.

Wheaton, Fingerman, Rose, and Leonard (1976) attempted to predict τ directly by estimating values for C and E, and substituting these into the original transfer equation. In order to do this, it was assumed that training time (regardless of experimental conditions) is a function of: (1) how well a training setting represents the operational (real world) situation, both in terms of tasks covered in training and fidelity of the training setting; (2) the difficulty inherent in the tasks which must be learned to some criterion; and (3) the appropriateness (or value) of the instructional techniques used to train the tasks. The first factor is represented in TV-A by two variables, a coverage variable, C (task communality), and a similarity variable, S (physical and functional similarity). The second is represented by the learning difficulty variable, D, and the third by the training variable, T. As Wheaton, Fingerman, Rose, and Leonard (1976) stated, "The time, trials, or errors to a criterion on subtask i is assumed to be a linear function of $C_i \times S_i \times D_i \times T_i$" (p. 48).

Since the training setting for a control group, however, is the operational equipment, it is clear that all operational subtasks are covered by the equipment ($C_i = 1$), and the physical and functional similarity is identical for each subtask ($S_i = 1$). It is also assumed in a TV-A application that, when training takes place on operational equipment, the instructional techniques used are optimal ($T_i = 1$). These assumptions mean that the amount of practice required by a control group (C) is determined solely by the difficulty of each subtask (D_i) summed over all subtasks:

$$\sum_{i=1}^{N} D_i$$

In order to estimate E for the experimental group, the amount learned on a training device must be subtracted from the amount learned on operational equipment. Since a training device is assumed not to be identical to the operational equipment, the values of the coverage (C_i) and similarity (S_i) variables will not always be "1", and must be estimated by the procedures just discussed. Likewise, the training techniques employed to teach each subtask are assumed to be less than optimal when a training device is used, T_i must also be estimated. The amount learned on operational equipment is $\sum_{i=1}^{N} D_i$.

Therefore E is assumed to be equivalent to: $\sum_{i=1}^{N} D_i - \sum_{i=1}^{N} C_i \times S_i \times D_i \times T_i$.

Given these estimated values of C and E, the predicted value of f is calculated by the equation:

$$f = \frac{\sum_{i=1}^{N} C_i \times S_i \times D_i \times T_i}{\sum_{i=1}^{N} D_i}$$

Summary

The Wheaton, Fingerman, Rose, and Leonard (1976) model purports to generate a prediction of transfer of training potential for training devices based on an analysis of both operational and training equipment. The model aggregates values for a series of factors assumed to be related to a device's effectiveness. The factors identified are task communality, similarity, learning deficits of the trainees, difficulty of each task to be trained and the training techniques incorporated into a device. The final evaluation index or figure of merit is a value ranging from "0" to "1.0", with values approaching "1" indicating greater transfer potential and, therefore, greater effectiveness.

In reviewing the TV-A model it is important to note that the theoretical assumptions and specific methodology were based on previous efforts (e.g., Wheaton, Rose, Fingerman, Korotkin, and Holding, 1974, 1976). Some of these assumptions may be questioned, and one might consider some elements missing. TV-A, however, represents one of the most systematic and complete methods for assessing device effectiveness. In fact, Wheaton, Fingerman, Rose, and Leonard (1976) have themselves begun a critical assessment and have suggested directions for future efforts. For example, they recommended consideration

of some important external variables. These include the amount of training and practice provided and user acceptance of a device. While these considerations are external to a device, they represent variables which can influence device effectiveness.

An additional device related variable that may be considered for inclusion in a model is what Wheaton, Fingerman, Rose, and Leonard (1976) call Environmental Fidelity Analysis (EFA). The EFA would potentially account for special or adverse conditions which may affect performance. These conditions may include extreme temperature, reduced visibility, etc. (Wheaton, Fingerman, Rose, and Leonard, 1976). It might be possible to obtain judgments or estimates of degradation of performance, probability of occurrence and subtasks affected by such conditions. A future model may, for example, include an assessment or estimate of how well a device prepares for such contingencies. The difficulty, of course, is that a device developer may not be able to replicate such conditions, assuming they are known, and a researcher may not be knowledgeable about human behavior under the same circumstances.

A future revision of TV-A might include a less laborious approach to the Training Techniques Analysis. Wheaton, Fingerman, Rose, and Leonard (1976) suggested that perhaps this analysis could be conducted at the subtask rather than the skill/knowledge level. This possibility becomes more realistic when considering recent evidence of the utility of such approaches as cluster analysis in ranking job related tasks (see Boldovici, Boycan, Fingerman, and Wheaton, 1979; Wheaton, Fingerman, and Boycan, 1978). From such analyses, it should be possible to form clusters of skills/knowledges or subtasks which can generalize to entire tasks. Applications of a T_i analysis, therefore, would be conducted on a restricted number of subtasks, and thus make analysis easier.

The evaluation methodology presented in this section remains to be validated both in terms of predictive ability and the constructs within the method. As one reviews the literature in this area, this criticism applies to other revisions of TV-A as well as to alternative approaches. It has become apparent, and will be discussed in the last section of this paper, that evaluations of the various approaches have been long overdue and represent a situation that must be remedied.

The methodology discussed thus far is based on a variety of assumptions; some of which are accepted while others may require further justification. TV-A assumes a linear relationship, for example, between the component variables and transfer of training potential of a device. This assumption is presently accepted particularly in the absence of any compelling reason to do otherwise. Another assumption made is that equipment similarity (i.e., fidelity) is monotonically related to transfer, and, therefore, a valid predictor variable. This is also related to the assumption that operational equipment represents an optimal training setting against which a device may be compared. There is presently no evidence to support these notions. An opposing perspective may assume that training devices are typically built with instructional features which are not present when operational equipment is used for instruction. In addition, training devices can be built to simulate the range of conditions a trainee may encounter on the job; this may not be possible when using operational equipment.

Criticisms which are not unique to the TV-A model include the detailed input requirements (e.g., task analytic data) and the premise that device effectiveness seems limited to transfer of training. In addition, all approaches reviewed for this paper mathematically combine a number of variables into a final, overall index. These criticisms remain unresolved and must be addressed in the near future.

TRAINVICE-B (TV-B)

Overview

The TRAINVICE-B (TV-B) model assumes that a device is the appropriate medium for training based on the media selection decision procedures specified in the Training Device Requirement Documents Guide (1979). Within the media selection decision procedures, a training developer previously analyzed and organized tasks, skills and knowledges, and objective data formulating a training device concept. The TV-B approach is purported to insure that established training requirements, incorporated into a device, were emphasized.

TV-B provides an approach to analyze and evaluate the effectiveness of a training device, typically in comparison to alternative device concepts or already existing devices. Applying the TV-B approach results in an effectiveness score for each alternative device concept, which is then used to decide which concept should be developed further.

The TV-B approach, therefore, is embedded in a series of administrative procedures designed to establish the need for a device, determine if a device which potentially may serve a training function already exists and to evaluate either existing devices or device concepts in terms of effectiveness.

For example, in deciding whether a device is an appropriate training medium, a developer would have already collected information regarding tasks, task elements, and controls and displays. The application of TV-B, in effect, becomes a trade-off analysis, because a device is not expected to meet all task training requirements. To the extent that a device does not address all the requirements, a developer is provided with a methodology to assess alternative concepts.

The TV-B methodology is similar to the TV-A approach. A rating of the correspondence between the operational equipment and a training device is combined with an index of the extent of training required and ability level of the trainees. The product of these values becomes the training device effectiveness index. In TV-B, however, when an existing device is compared to a training concept or requirement, the effectiveness index may be adjusted for providing additional training beyond that required. The assumption is that training additional skills represents unnecessary costs which lead to a loss of effectiveness.

Procedures

The TV-B methodology allows values to be assigned to components which comprise two basic subdivisions: (1) device characteristics and (2) personnel

and training requirements. These sub-divisions are further divided into the following components:

Device Characteristics	o Task Commonality o Physical Similarity o Functional Similarity
Personnel & Training Requirements	o Skills and Knowledges Requirements o Task Training Difficulty

Values for these components are combined to form an index of training device effectiveness.

The information required to perform a TV-B analysis includes the:

1) list of tasks and elements (i.e. sub-tasks) to be trained (operational tasks);

2) tasks and task elements which can be trained with a particular device. This list includes task elements covered by a training device, which are not contained in the training objectives (i.e. unique elements);

3) skills and knowledges required to meet the training objectives;

4) controls and displays used to perform the tasks in the operational setting; and

5) controls and displays in the training device.

Task Commonality Analysis (TC)

The Task Commonality (TC) analysis in TV-B is different from the C analysis in TV-A. In TV-A, each subtask is given a"1" or a "0" depending on whether it was covered by a training device. In TV-B, a TC value is determined for each task, by rating whether or not task elements which require training are covered on a device ("1" covered, "0": not covered). The TC value for a task is calculated by adding all task element ratings, and dividing this sum by a combination of the total number of required task elements plus the number of task elements which do not require training but are still covered by the training device (i.e. unique elements).

Physical Similarity Analysis (PS)

In the Physical Similarity (PS) analysis the controls and displays on a training device and on the operational equipment are compared in terms of their appearance, size, location, etc. The comparison is made only for device characteristics which are directly involved in performing those task elements which require training. Each control or display on a training device is rated on the degree of physical similarity (i.e., fidelity) between it and the corresponding control or display on the operational equipment. The rating scale, used for this purpose, ranges from "0" (missing) to "3" (identical).

The scale values and criteria for judgments are very similar to those in TV-A. There are, however, changes in phrasing; e.g., "small noticeable differences" in place of the more traditional and, perhaps, technical "just noticeable differences" (see Appendix).

In order to derive a Physical Similarity index for each task, the ratings given to controls and displays on a device are totalled. This sum is then divided by a combination of "3" times the total number of required controls and displays plus the number of "unique" controls and displays. The unique pieces of equipment on a device are those used for task elements or skills which are associated with the task in question, but do not require training. Thus, the resulting index varies between "0" and "1", representing the physical similarity adjusted for extra or "unique" equipment.

Functional Similarity Analysis (FS)

The Functional Similarity (FS) analysis in TV-B, like that in TV-A compares the controls and displays of a training device to those in the operational equipment in terms of amount of information conveyed from or to the human operator. Just as in the PS analysis, each of the "required" controls or displays relevant to a particular task receives a rating from "0" to "3". The rating scale used, though similar to that in TV-A, includes less technical language. A "2" on the TV-A scale, for example, means that the amount of information in the operational and training settings are "within one $\log_2$ unit of each other." The corresponding description in TV-B is "the number of states in the training situation is less than half of the number of states in the operational setting." The only time the two scales are equivalent, is when there is less information in a training setting. The $\log_2$ in TV-A can also mean greater information in a training setting. This distinction, however, is consistent with the TV-B approach in adjusting for unique skills.

In order to calculate the functional similiarity index for each task, the ratings given to all controls and displays on a device are summed and the total is divided by the number of required controls and displays plus the unique ones. This results in an index ranging from "0" to "1". The last operation, (i.e. the inclusion of the unique displays and controls in the denominator) is the cost adjustment for extra training device features.

Skills and Knowledges Requirements Analysis (SKR)

In TV-B, there are two separate preparatory analyses which correspond to the Learning Deficit Analysis in TV-A. In TV-A, the Learning Deficit variable represents an estimate of how much the trainees have to learn, weighted by the amount of time it takes to train them to overcome a deficit, on the operational equipment. The procedures involved in both TV-B and TV-A are performed independently of the characteristics of the training device under evaluation.

In Skills and Knowledges Requirements Analysis (SKR), each skill or knowledge required to perform a task receives two ratings. The first rates the level of proficiency trainees have before training. The second rates

the proficiency level required after training. The rating scales used range from "0" (no experience) to "4" (complete understanding) and are almost identical to the Repertory and Criterion Scales used in TV-A (see Appendix.) The difference in before and after proficiency levels is determined for each skill or knowledge by subtraction. For each task, a SKR index is calculated by taking the mean of the proficiency differences and scaling it down to between "0" and "1" by dividing by 4.

Task Training Difficulty Analysis (TTD)

The TTD is quite different from the corresponding procedures in TV-A. The first step in this analysis is to determine how much time would be required to train the most difficult task element of all those in the training objectives (i.e., across all tasks). Training time here means time to train on the operational equipment. A TTD index is derived for each task by rating each required task element on how much time is needed to train it on the operational equipment, relative to the training time required by the most difficult task element. The ratings are made using a scale which ranges from "0" (requires no training) to "4" (requires as much time to train as the most time consuming task element) (see Appendix.) The index given each task is the average of the difficulty ratings given each task element, scaled to between "0" and "1".

Index of Training Device Effectiveness

The analyses just presented are used to calculate an overall index of effectiveness for a training device or concept. The values for TC, PS, and FS are summed and divided by 3. This value represents the degree of correspondence between a training device and the operational equipment. Next, the SKR and TTD values, for each task, are added and divided by 2. This value represents the amount of training required. In order to calculate the Index of Training Device Effectiveness, the value representing the degree of correspondence and the amount of training required are multiplied for each task. These products are then summed with the final index obtained by dividing by the amount of training required (i.e. $\frac{SKR + TTD}{2}$).

The final index formula is:

$$\frac{\sum_{i=1}^{N} \left(\frac{TC + PS + FS}{3} \times \frac{SKR + TTD}{2} \right)}{\sum_{i=1}^{N} \left(\frac{SKR + TTD}{2} \right)}$$

The TV-B model attempts to adjust the final index by a correction factor which reflects a loss of effectiveness due to unnecessary cost. This adjustment factor is calculated as:

$$\frac{\text{\# of Required Tasks}}{(\text{\# of Required Tasks}) + (\text{\# of Tasks Unique})}$$

This factor accounts for capabilities in a device that are not required. The adjustment factor is applied by multiplying it to the final index. This

adjustment is assumed not to be required when assessing theoretical device concepts, only existing devices.

Summary

TV-B is similar to TV-A in terms of many of the components which enter into the overall training device index. Two major subdivisions comprise the Honeywell approach. These include measures that assess the degree of correspondence between a device and operational equipment for which it was developed. The degree of correspondence assessment is similar to TV-A in that Task Commonality, Physical and Functional Similarity are determined. These values are later combined with an index of the amount of training required for a set of tasks; again, similar to TV-A.

A major distinction between TV-A and TV-B is that TV-B does not include an assessment of the training techniques incorporated in a device. That is, there is no measure of the appropriateness of the instructional features in relation to accepted learning principles. Another difference is that while TV-A adjusts the overall effectiveness index for failing to cover tasks on a device, TV-B additionally penalizes a device for including additional instructional features beyond those required. The rationale of this latter adjustment is the assumption that a decrease in training effectiveness results when unique or unnecessary skills are taught. The rationale continues into cost considerations as well. That is, additional training in non-required skills costs more, and therefore is undesirable. These assumptions and related adjustments may be suspect and unwarranted. Without an assessment of adherence to accepted instructional or learning guidelines there appears to be little basis for such a penalization. In fact, there may be instances where additional skills, beyond those required, may enhance overall transfer of training and this may go completely unrecognized by an evaluator.

The TV-B approach, however, does emphasize the relationship of effectiveness with cost considerations more than the TV-A model. This is particularly relevant when the objective is to assess the total long-term training cost in relation to effectiveness as the Guidebook indicates. Indeed rarely does device development proceed without cost considerations in terms of resources required for facilities, equipment, instructional material, personnel, students, supplies, etc.

Finally, TV-B, like TV-A, relies on a number of assumptions which include linearity and method of mathematical aggregation. These, along with other issues pertaining to reliability and validity, are major concerns and will be discussed further in a later section.

TRAINVICE-C (TV-C)

Overview

A revised version of the TV-A approach, referred to in this paper as TRAINVICE-C (TV-C) was developed to increase the practicality and flexibility of a device effectiveness model (Narva, 1979a; 1979b). TV-C attempted to provide a means for answering three questions about a training device: "what", "why", and "how".

The "what" question addresses what should be represented in a device. Two judgments are required in the answer. The first refers to the requirement for an activity to be incorporated into a device. The second refers to whether the device actually covers an activity.

The "why" question tries to uncover the reasons for including training activities on a device. The two stages of this issue include training criticality, or the level of proficiency required at the conclusion of training, and training difficulty, or how hard it is for a trainee to reach that proficiency level.

The "how" question pertains to the physical and functional characteristics of a training device. That is, TV-C assesses how well displays and controls (i.e., physical characteristics) follow accepted instructional or training guidelines, and the trainer's requirements. In addition, the "how" refers to the extent functions of displays and controls (i.e., functional characteristics) adhere to guidelines on instruction. Judgments are made for every skill or knowledge required on a training device, with values corresponding to these judgments substituted in a formula designed to reflect the percentage of maximum transfer which would be fostered by use of a particular training device.

Procedures

Coverage Requirements Analysis (CR)

The first analysis performed in TV-C is the Coverage Requirements Analysis (CR). The procedure consists of assigning a "1" or a "0" to each skill or knowledge (from the operational task analysis), depending on whether or not it should be covered by a device. In other words this analysis serves to determine which skills and knowledges warrant training. This screening process already existed in TV-A, as part of the Learning Deficit Analysis (a CS rating of "0")(Narva, 1979a, 1979b). TV-C simply highlights this issue for separate and initial attention. In either case, however, a high premium is placed on the judgment of a training analyst.

Coverage Analysis (C)

The Coverage Analysis (C) compares the skills and knowledges in the operational setting with those covered by a training device. Just as in the Task Communality for TV-A and Commonality Analyses for TV-B, a C value of "1" is assigned to each operational skill which is represented, a "0" when not represented in the training setting. The only difference between the TV-C approach to coverage and methods used in earlier versions of TRAINVICE is TV-C, ratings are made for each skill, whereas in the others the rating is made for each subtask.

Training Criticality Analysis (C_1)

All skills receiving a rating of "1" in both of the preceding analyses are then subjected to the Training Criticality Analysis (C_1). Each skill or knowledge is rated on the degree of proficiency which will be required after training (not mission criticality). The scale used to make this rating is almost identical to the Criterion Scale used in the Training Deficit Analysis for TV-A (See Appendix). The only difference is that the "0" value was dropped because a rating of zero proficiency has already been taken into account by the Coverage Requirement Analysis. The values for the C_1 variable range from "1" to "4".

Training Difficulty Analysis (D)

In the Training Difficulty Analysis (D) each skill receives a rating, from "1" (minimal or none) to "4" (substantial), on the degree of difficulty to learn that skill to required proficiency levels. Aside from the difficulty inherent in a skill itself, a rater must also consider the proficiency level of the trainees before training and that required after training. In essence, this analysis greatly simplifies the TV-A procedures for deriving the weighted learning deficit; especially the rank ordering of subtask difficulty.

Physical Characteristics Analysis (PC)

In considering the equipment on a training device, (i.e. the device characteristics), the TV-C approach is quite different from that in the earlier versions of TRAINVICE (i.e., physical and functional similarity). As alternatives to equipment similarity ratings, the TV-C physical and functional characteristics analyses represent attempts to have a training analyst assess more directly "how" a device will train skills. In this sense, the device characteristics analyses of TV-C resembles the training techniques analysis in TV-A.

The Physical Characteristics Analysis (PC) addresses the appropriateness of the physical equipment supporting the training of each skill. Each skill associated with the controls and displays is translated into a generic characteristic (e.g., Stimulus Capabilities: Visual Form - Visual Alphanumeric, etc.). The generic characteristics recommended are those contained in the ISD model (Braby, Henry, Parrish, and Swope, 1975). Each of the generic characteristics of the cue or response related to a display or control is rated on how well it follows available guidelines. The rating scale used ranges from "0" ("not adequate") to "3" ("outstanding"). The physical characteristics score, for each control or display, is the sum of the ratings given to each relevant generic characteristic. Similarly, the physical characteristics score for each skill is the sum of the scores given to each of its associated controls and displays.

In order to assist in making the physical characteristics ratings, TV-C refers a user to a series of learning guidelines (ISD). To use these guidelines, each skill must first be classified as belonging to one of ten behavioral categories (e.g. identifying symbols, detections, etc.). For each of the behavioral categories there is an associated set of learning guidelines. Narva (1979b) cautions about the lack of specificity of the ISD guidelines. These were originally intended to assist in the selection of instructional media. For this reason, the user must be selective in the application of the learning guidelines. Again, it must be emphasized that use of the guidelines does not directly generate physical characteristics ratings, it merely alerts the user to some of the general behavioral considerations associated with each of the behavioral categories to which a skill might belong.

Functional Characteristics Analysis (FC)

The Functional Characteristics Analysis (FC) attempts to assess how the physical characteristics of a training device are used. The first step in this analysis is to place each skill in one of the ten behavioral categories (as in the PC analysis). A user then refers to the set of ISD Learning Guidelines associated with each behavioral category and selects those appropriate to the specific skill under consideration. Ratings are given to a skill on how well each of the relevant guidelines are implemented or used in a training device ("0", not adequate; to "3", outstanding). The FC value given to each skill is the sum of the ratings made on each of the associated guidelines.

Index of Predicted Training Effectiveness

The calculation of the TV-C index of effectiveness was designed to represent the percentage of maximum transfer. the procedures to combine the components consist of a ratio in which the various values given to a device are combined in the numerator. The denominator is a combination of the maximum possible ratings which could have been given. The TV-C formula is:

$$\frac{(CR \times C \times C_i \times D \times (PC + FC))_i}{(CR \times c \times C_i \times D \times (PC_{max} + FC_{max}))_i}$$

where:

CR	Coverage Requirements Score
C	Coverage Score
C_i	Training Criticality Score
D	Training Difficulty Score
PC	Physical Characteristics Score
FC	Functional Characteristics Score
PC_{max}	Maximum Possible Physical Characteristics Score
FC_{max}	Maximum Possible Functional Characteristics Score

The form of the equation given in the first TV-C report (Narva, 1979a) shown above, was modified slightly in a second report (Narva, 1979b) to the following:

$$\frac{(CR \times C \times C_i \times D \times (PC + FC))_i}{(CR \times C \times 4 \times 4 \times (PC_{max} + FC_{max}))_i}$$

The value of "4" substituted for the criticality (C_i) and difficulty (D) variables in the denominator is simply the greatest value either of these variables could have. Both of the above equations yield indices which range between "0" and "1". A larger index value (i.e., closer to 1) presumably indicates a greater potential for transfer to the operational equipment.

Summary and Critique

Like the other versions of TRAINVICE, TV-C attempts to assess the training transfer potential of a training device by assigning values to a variety of judgments about a device.

Essentially, there are three major subdivisions within TV-C, an input, a training analysis, and a device characteristics analysis. The inputs include the operational and training requirements which are derived from a task analysis of each situation (i.e. operational and training). The training

analysis is an estimation of the required level of proficiency and difficulty to arrive at that level for each trainee. Device characteristics analyses include an evaluation of the physical and functional aspects of components incorporated into a device as these adhere to accepted instructional or learning priciples.

In two papers, ARI RM 79-6, 79-7, Narva (1979a, 1979b) outlined an extensive modification of the original TRAINVICE predictive model. The procedures described in these two papers are identical; only the calculation of the index was modified in the second paper (i.e., Narva, 1979b). The most striking difference between TV-C and the earlier models is the omission of an equipment similiarity (fidelity) analysis. The training techniques analysis, which had been dropped in TV-B was reintroduced in TV-C in the form of two separate analyses (physical and functional characteristics analyses). TV-C also contains a coverage requirement (or media selection) analysis, not included in TV-A, or TV-B. The procedures and rating scales used in the various preparatory analyses were almost completely changed in TV-C. Also, the level at which these analyses are performed is at the individual skill level, not subtask. Considerable changes were also made in the procedures used to calculate an overall index of effectiveness.

TV-C included a Training Criticality and Training Difficulty analysis as a weighting factor for required skills and knowledges. A skill or knowledge, therefore, which is required at a high level of proficiency, in addition to being difficult to learn is assumed to have more significance than one requiring a lower proficiency and which is easier to learn. Given two devices under evaluation, for example, one covers an important skill while the other does not. The evaluation model was originally intended to penalize a device in such a situation. TV-C fails to accomplish this. A "0" C value for a skill causes both the numerator and denominator to go to "0" for a skill not covered, although required. The result is as if that skill never existed. As will be discussed shortly, TV-D corrected this situation.

The terminology of Training Criticality Analysis is somewhat misleading. The word Criticality seems to suggest the notion of importance, either in the mission or training setting. As presented earlier, the C_i analysis addresses the required level of proficiency for trainees, and has nothing directly to do with criticality.

The criteria for a user to make judgments about each of the analyses appears to be too vague. The scale for the D analysis, for example, is:

1 = minimal or none

2 = some

3 = much

4 = substantial

These descriptions of the rating scale may reduce the reliability of the application. That is, because of a lack of specificity of definitions, judgments by different users may vary according to individual interpretations. This possibility exists whenever scales of this type are used, however, the more specific the criteria for assigning values are less likely

differences in interpretation can occur. This leads to a further restatement of the need to validate the methodology both in terms of construct and predictive validity. This issue will be discussed further in a latter section of this paper.

TRAINVICE-D (TV-D)

Overview

In a project to develop a user's guidebook for TV-C, further revisions were made to the evaluation model (see Swezey and Evans, 1980; Evans and Swezey, 1981). Despite their differences, TV-C and TV-D are almost identical in the variables or model components considered and in the procedures used to estimate these variables.

Two general uses of this model have been identified as predictive or prescriptive applications. A predictive application is used when existing training devices are available and a user wants to evaluate (or predict) their effectiveness. In its prescriptive mode, the model is purported to assist device developers in making design decisions in the early concept stages. Components are applied either separately or in combination. When components are combined, an overall index of device effectiveness is derived. The overall index or separate components analyses are only of value when two or more devices are under evaluation. In one sense this restriction is the result of the overall index having no intrinsic or absolute meaning. In another, each of the components can be used as a comparative assessment to identify deficiencies in a device under evaluation.

The components of this model are:

- Coverage (C)
- Training Proficiency (P)
- Learning Difficulty (D)
- Physical Characteristics (PC)
- Maximum Possible Physical Characteristics (PC_{max})
- Functional Characteristics (FC)
- Maximum Possible Functional Characteristics (FC_{max})

The formula for executing TV-D is:

$$\frac{\sum_{i=1}^{N} \left(\frac{PC + FC}{PC_{max} + FC_{max}} \right) (C \times P \times D)}{\sum_{i=1}^{N} (P \times D)}$$

A device evaluation actually begins by determining whether training is required on the skills and knowledges needed for performance on the operational equipment. Once the training skills or knowledges have been identified, a user then conducts analyses using the model components. These are briefly described below.

Procedures

Because TV-D is a direct derivative of TV-C, it addresses the same "What", "Why", "How" questions TV-C does. As will be discussed shortly, some changes have been made.

Coverage Requirements Analysis (CR)

Although not formally used in the overall index formula, the Coverage Requirements Analysis (CR) helps determine which skills or knowledges required in the operational setting should be represented on a training device. Working from a consolidated list of skills and knowledges, a user decides whether each skill or knowledge should be covered by a training device. If the decision is "yes", then a value of "1" is assigned to that skill; a "no" decision receives a "0".

Coverage Analysis (C)

If a skill or knowledge is required, a user must then decide if that skill or knowledge is actually represented. A Coverage Analysis (C) value of "1" indicates that it is, while "0" indicates it is not. If the analysis is conducted early in a device development phase then a required skill (i.e. CR = "1"), which was not originally covered in a device design (i.e. C = "0"), can be included. The effect of failing to cover a required skill is reflected in a lower overall index for a particicular device.

Training Proficiency Analysis (P)

This component assigns a value corresponding to the degree of proficiency which a trainee must attain for each skill or knowledge subsequent to training on a device. The Training Proficiency Analysis (P) is conducted on each skill or knowledge which received a CR value of "1", even if one device in a comparison failed to cover (i.e., C = "0") a particular skill or knowledge.

A four point (i.e., "1" to "4") rating scale is used to assign a P value, where "1" corresponds to a level requiring limited knowledge. When expert levels of knowledge are required, a P value of "4" is assigned. The P values are then summed across all skills and knowledges.

Learning Difficulty Analysis (D)

The Learning Difficulty Analysis (D) specifies the degree of learning difficulty associated with attaining a required skill or knowledge. Several

factors have been identified which enter into a user's decision in assigning a D value. These are the:

- o level of skill/knowledge proficiency to be attained by a trainee
- o entry-level capabilities of a trainee (i.e. pre-training on the skills or knowledges.
- o level of learning difficulty typically inherent in a skill or knowledge

In making D judgments, a user assigns a value ranging from a low of "1" to a high of "4". The higher a D value the more difficult a skill or knowledge is to learn. Like the P analysis, D values are assigned only to skills or knowledges which have been determined to be required (i.e. CR=1) and then summed.

Physical Characteristics Analysis (PC)

This is the first of two analyses which are referred to as Device Characteristics Analyses. In other words, attention is now focused on analyzing displays and controls on a device. The Physical Characteristics Analysis (PC) assesses how well the physical characteristics of a device support guidelines or principles of good instruction. A separate PC analysis is conducted for each device under consideration.

In conducting a PC analysis, a user must first determine the type of behavior that is required to accomplish a particular skill or knowledge. Each skill or knowledge is assigned to a behavioral category which coresponds to the type of performance required by a trainee. These behavioral descriptions were adapted from the U.S. Army Interservice Procedures for Instructional Systems Development (TRADOC Pam. 350-30, 1975). Next a user decides which instructional practices are applicable for developing the type of behavior associated with a skill or knowledge, which are listed under each behavioral category. These instructional practices or guidelines represent a standard against which each device will be evaluated. Because these guidelines correspond to skills or knowledges, they remain the same for each device under evaluation.

A user then identifies the Generic Stimulus and Response Characteristics for each display and control which correspond to particular skills and knowledges. That is, a user must identify the stimulus characteristics of displays and learner response modes. The list of possible stimulus characteristics (i.e. capabilities) and response modes are those presented by Braby, Henry, Parrish and Swope (1975). The PC analysis concludes by assigning a value or rating on how well each generic characteristic of a display or control supports the good instructional practices identified earlier. Values of the PC analysis range from "0", extremely deficient in implementing the guidelines, to "3", implementation is highly proficient, for each skill or knowledge. The total PC score then becomes the sum of the values assigned to each skill or knowledge.

Maximum Possible Physical Characteristics (PC_{max})

The maximum possible physical characteristics (PC_{max}) value for each skill or knowledge is simply three times the number of applicable generic stimulus and response characteristics.

Functional Characteristics Analysis (FC)

The second device characteristic analysis is the Functional Characteristics Analysis (FC). The FC analysis is similar to the PC Analysis in that it assesses how well the functional elements of a training device follow guidelines for good instructional practice. Skills and knowledges are, again, compared to the behavioral categories, and good instructional practices under each category. These instructional guidelines are now identified solely for functional and not physical characteristics. Again, these form a standard to which the functional worth of displays and controls are compared. In completing the FC Analysis, a user rates how well each display and control, corresponding to a skill or knowledge, implements the functional guidelines for good instructional practice. The scale used ranges from "0" extremely deficient implementation to "3" highly proficient implementation of the guidelines for each skill or knowledge. These values are summed for all skills and knowledges under consideration (i.e. CR = "1").

Maximum Possible Functional Characteristics (FC_{max})

Like the PC_{max}, the maximum possible functional characteristics score is three times the total number of applicable functional guidelines.

Index Calculation

The calcualtion of the final index is completed by simply substituting the values of each analysis discussed above and carrying out the operations in the formula:

$$\frac{\sum_{i=1}^{N} \left(\frac{PC + FC}{PC_{max} + FC_{max}}\right)\left(C \times P \times D\right)}{\sum_{i=1}^{N} \left(P \times D\right)}$$

The resulting index will be a number between "0" and "1". As this value approaches "1", the better training transfer capability of a device. The overall index, however, only has value when comparing two or more existing devices or device concepts.

Summary

While quite similar to previous approaches, some changes have been instituted. Training Proficiency Analysis (P) was formerly called Training

Criticiality (C_i) in TV-C. The term "Criticality" was considered misleading, perhaps suggesting importance of a skill, which was not the case.

The principle difference between the TV-D formula and that of TV-C is the removal of the coverage variable from the denominator. In this way the credit or penalty supposedly given to a device for covering or failing to cover a particular skill is weighted by the Training Proficiency and the Learning Difficulty scores. That is, an overall index of effectiveness would be enhanced more for covering skills that require a high degree of proficiency and are difficult to learn, than for covering relatively trivial skills. Similarly, when a skill is not covered, the degree to which an overall effectivness index is decreased is weighted by the proficiency and difficulty scores for a skill. In the TV-C formula, the credit given for coverage of a skill was weighted by criticality and difficulty, however, lack of coverage was not penalized at all. The presence of the coverage variable in both the numerator and denominator would cause both to equal zero for an uncovered skill. Thus, in TV-C, each skill not covered by a training device neither contributes to nor takes away from an overall effectiveness index.

Four of the rating scales used in the preparatory analyses for TV-C were modified in TV-D. These are the scales used in the Training Difficulty Analysis, Training Proficiency Analysis ("Criticality" in TV-C), and Physical and Functional Characteristics Analyses. In all cases, wording changes were made in the attempt to provide more guidance to the user than had been available in TV-C (see Appendix). There were no changes, however, in the numerical properties of the scales.

The Physical and Functional Characteristics Analyses contain additional changes in guidance given to users. The ten behavioral categories (from ISD) used in TV-C were given expanded definitions accompanied by examples. TV-D incorporated new learning guidelines, associated with each behavioral category, which were modifications of those already in the ISD. Moreover, each learning guidline was identified with a "P", "F", or "P/F"; to indicate whether a particular guideline was relevant to analyzing the physical characteristics, the functional characteristics, or both.

Several issues of concern have evolved regarding t application of TV-D. Actually these issues appear equally valid for t earlier models as well. The first is the manner in which values for the various components are aggregated into a single index. The components of TV-D appear to form a series of fractions, all based on separate criteria. These then become accumulated or summed in violation of basic rules for such addition. In other words, there is no attempt to find a common denominator.

A second concern is that different guidelines on "good instructional practices" are used for the PC and FC analyses. Further, the procedure for designating the PC and FC values is cumbersome, both of these issues seem to increase the possibility of poor reliability in assigning values.

By necessity, it seems that a long list of skills and knowledges are required to apply TV-D. Once these are identified, a series of additions and multiplications is required. Again, reliability seems to be vulnerable, if for no other reason then arithmetic errors. In addition, a user must begin a TV-D analysis with a consolidated list of skills and knowledges derived from

the list of all skills and knowledges required in the operational setting. Construction of the consolidated list requires a user to eliminate from consideration those skills and knowledges that are repeated on more than one task or subtask. A TV-D index, therefore, is desired on only a selected number of skills and knowledges, with no implication for a particular skill being repeated. Perhaps a logical argument can be made that if a skill or knowledge appears in more than one task, then that repetition should indicate some degree of importance. Yet, in executing TV-D, all skills begin as equal with only proficiency and difficulty as primary considerations or weighting factors.

Another issue of concern is the reliance on TRADOC Pam. 350-30 as providing "good instructional guidelines." These guidelines were developed for training programs in general and not for training devices. This application in a device effectiveness method is suspect.

III

Summary of the Models

Input

All four TRAINVICE models require task analytic and equipment information as input. The models vary somewhat in the detail of the task information required for input, as well as in the task taxonomic level at which variable values are estimated (e.g., task-by-task) (see Table 1). There are two types of equipment information required: physical (i.e. size, location, etc.) characteristics of the controls and displays and functional chracteristics (operation and use of the equipment). The four models are comparable in the amount of detail required in the physical information. There are, however, differences among the models in the level of resolution required in the functional information. The two models which involve equipment similarity analyses (TV-A and TV-B) require specification of the amount of information (in bits) transmitted between the human operator and the controls and displays. The two models without similarity analyses (TV-C and TV-D) may need more general accounts of the stimuli (or cues) supporting the behavior and the types of responses required.

TABLE 1

Models	Input Resolution	Level of Analysis
TV-A	Sub-task	Sub-task
TV-B	Task element; Skill	Task
TV-C	Skill	Skill
TV-D	Skill	Skill

Preparatory Analyses & Model Variables

The four TRAINVICE models involve several general types of preparatory analyses. Table 2 shows the commonalities among the models in terms of these analyses. Each kind of analysis produces an estimate of a value for a particular variable.

In the coverage and communality analyses, a "1" or "0" is used primarily to penalize for non-coverage of skills. Penalization issues are most relevant to each model's equations, and will be discussed below.

The class of variables in Table 2 called "Learning" variables concern: 1) the amount of increase required in the proficiency levels of trainees, and 2) the amount of difficulty inherent in training each task. In TV-A both of these are combined into a Weighted Learning Deficit score. In other words, an estimate of incoming trainee skill level is subtracted from a criterion proficiency level. This difference is then weighted by the ranked difficulty of training that particular skill. In TV-B, the difference in proficiency levels (Skills and Knowledge Requirements) is estimated in a similar way to TV-A.

Table 2

Summary of Classes of Variables and Variables Used by the Four TRAINVICE Models

Class of variable	TV-A		TV-B		TV-C		TV-D	
	Analysis name	Variable label	Analysis name	Variable label	Analysis name	Variable label	Analysis name	Variable label
Coverage requirement	--	--	--	--	Coverage requirement analysis	CR	Coverage requirement analysis	CR
Communality	Task communality	C	Task commonality	TC	Coverage analysis	C	Coverage analysis	C
Physical/ functional similarity	Physical similarity analysis	P	Physical similarity index	PS	--	--	--	--
	Functional similarity analysis	F	Functional similarity index	FS	--	--	--	--
	Similarity score	$S=\frac{(P)+(F)}{2}$						
Learning	Learning deficit analysis	D	Skills & knowledge requirements index	SKR	Training criticality analysis	Ci	Proficiency analysis	P
			Task training difficulty index	TTD	Training difficulty analysis	D	Learning difficulty analysis	D
Training technique	Training technique analysis	T	--	--	Physical characteristics analysis	PC PC_{max}	Physical characteristics analysis	PC PC_{max}
					Functional characteristics	FC FC_{max}	Functional characteristics	FC FC_{max}

Training Difficulty is estimated more simply by a rating scale, instead of a ranking procedure. The two values (proficiency requirements and training difficulty) are averaged in the TV-B final equation. In both TV-C and TV-D, the values of the proficiency and difficulty variables are estimated with rating scales and are kept separate throughout subsequent calculations.

The equipment similarity and training techniques variables are the only model components which are concerned with the features of a training device and how well they will support training. As can be seen in Table 2, the models vary widely in their emphasis, or lack of emphasis, on each of these variables.

The one model which addresses both variables is TV-A. Here, equipment similarity has two components: physical and functional similarity. Values are assigned to each and are averaged for an overall Similarity score. In order to derive the training techniques score, a user first categorizes each subtask according to Braby's, et al. (1975) task taxonomy. The task category then refers the user to a special set of learning principles for that category (after Willis and Peterson, 1961; and Micheli, 1972). The principles concern stimulus, response, and feedback aspects of equipment. A conservative estimate is made regarding the implementation of these principles by a device, which then generates a value for the Training Techniques variable.

In TV-B, training techniques are ignored, with an average of physical and functional similarity scores as the only predictor variable. The analysis used to generate the Similarlity score is almost identical to that in TV-A.

TV-C and TV-D abandon equipment similiarity as separate analyses. It is hard to disagree with this because there is little literature supporting the assumption of a general, monotonic relationship between equipment fidelity and training effectiveness (a minimal criterion for the selection of any predictor variable.) The traditional assumption of such a relationship has undoubtedly been based on approaches to transfer of training such as Osgood's (1949). The problem with such an assumption, in the context of training devices, is that it must lead to the conclusion that the best device for training is the operational equipment itself. Put differently, this approach assumes that the cues necessary to maintain skilled performance, on the operational equipment, are sufficient and in fact optimal to support learning.

The level of stimulation present on a training device however, may have different effects on various kinds of learners. Skilled performers, for example, have already learned to use to their advantage all the relevant cues available in the operational environment. To a novice, however, the stimulation presented by the operational environment may be, in large part, noise; (i.e., a source of distraction), therefore a hinderance to learning. Sometimes, it may be desireable, therefore, to reduce the number of cues available (i.e., lower fidelity) during initial training. While in other situations, it may be desireable to increase the amount of information presented in the training environment in order to augment feedback and knowledge of results. In yet other simulations, compressing the time frame of a task series may enhance training.

While presently there may be insufficient knowledge regarding relationships between fidelity and training effectiveness to warrant its use as a

predictor variable (without qualifications), equipment similarity cannot be ignored. In the solution adopted by TV-C and TV-D, fidelity is considered only in the context of fairly specific task domains, not as an end in itself. TV-C and TV-D adopted what is in essence an amplification of the Training Techniques Analysis in TV-A. The modified analysis (in TV-C and TV-D) directs a user to different sets of learning principles for different skills. Using these principles, a user assesses how well the physical and functional characteristics of a training device support training. The learning principles vary for each skill category. For some skills, the relevant principles include guidelines concerned with some aspect of equipment similarity. For other kinds of skills, fidelity is de-emphasized. Realistic and continuous feedback is recommended for tracking tasks, for example. Whereas "equipment realism can be at a minimum" for procedural tasks.

Model Output

The values determined for the preparatory analyses are combined in a specific computational formula for each model (in Table 3). Each formula is used to generate an overall index of training effectiveness which ranges between 0 and 1; the higher the index, the more effective a training device. All of the equations used by the models have been designed to predict training effectiveness, with overlap in the variables considered. The only mathematical property common to all of the formulae is the use of linear combinations. That is, the variables are combined in a simple multiplicative fashion.

TV-A is the only model whose formula was based on the Gagne, et al. (1948) savings measure of transfer of training. The index of effectiveness for a device is determined by the equipment similarity and training techniques scores, weighted by the learning deficit score. The weighting strategy employed was the "weighted mean". The general form taken by a weighted mean is: If each value x_i is associated with a weighting factor w_i, where $w_i \geq 0$, then $\sum_{i=1}^{n} w_i$ is the total weight, and:

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Note that the weights can not have negative values.

The equation used to generate the index in TV-B is not clearly related to any particular transfer of training measure. The TV-B index is determined by equipment similarity scores weighted by the required skills and knowledges and task training difficulty scores. Together, these two variables cover information similar to that in the learning deficit score of TV-A. The manner in which the weighting is accomplished in TV-B can only be considered a weighted mean when there are no "unique tasks" trained by a device.

The index of TV-C was developed to reflect the percentage of maximum transfer possible. The equation used to compute the index, therefore, is a

Table 3

Model Equations for the Calculation of Overall Effectiveness Indices

$$\underline{\text{TV-A}}\ \text{Index} = \frac{\sum_{i=1}^{N} C_i \times S_i \times D_i \times T_i}{\sum_{i=1}^{N} D_i}$$

N = number of subtasks

$$\underline{\text{TV-B}}\ \text{Index} = \frac{\sum_{i=1}^{N} \left(\frac{TC + PS + FS}{3}\right) \times \left(\frac{SKR + TTD}{2}\right)}{\sum_{i=1}^{N} \left(\frac{SKR + TTD}{2}\right)} \times \frac{RT}{RT + UT}$$

N = number of tasks

RT = number of required tasks

UT = number of "unique" tasks

$$\underline{\text{TV-C}}\ \text{Index} = \frac{\sum_{i=1}^{N} (CR \times C \times Ci \times D \times (PC + FC))_i}{\sum_{i=1}^{N} (CR \times C \times 4 \times 4 \times (PC_{max} + FC_{max}))_i}$$

N = number of skills

$$\underline{\text{TV-D}}\ \text{Index} = \frac{\sum_{i=1}^{N} \left(\frac{PC + FC}{PC_{max} + FC_{max}}\right) \times (C \times P \times D)}{\sum_{i=1}^{N} (P \times D)}$$

N = number of skills

N.B. See Table 2 for variable names.

ratio of the variable values estimated for a particular training device, divided by the maximum values which could be assigned to those variables (with the exception of the coverage variables).

The only part of the TV-D formula which retains the above percentage is the ratio of physical and functional characteristics scores, to their maximum values. The rest of the equation has been revised, primarily for reasons related to penalization of a device for non-coverage of particular skills.

In computing the overall figure of merit for a training device, a coverage penalty has been included in various ways in the four models. TV-A penalizes a device for not covering subtasks which require training. The penalization strategy used in TV-B lowers the index both for: 1) not covering tasks requiring training, and 2) covering tasks which do not require training (i.e., unique tasks). The implementation of this penalty in TV-B is present in almost all of this model's preparatory analyses, as well as being part of the final equation. The reason given for the penalization of unique tasks was that it would allow the TV-B index to reflect an unnecessary increase in cost, while lowering training effectiveness. The major problem with this rationale is the underlying assumption that all "extra" training features cost the same amount and generally lower effectiveness; i.e., the penalty is equal for all unique tasks. TV-B is the only model to use this penalization strategy.

In TV-C, there is no penalization for non-coverage. If a skill is not covered by a training device, zeroes are entered into the summations in both the numerator and denominator of the final equation. That is, nothing is contributed or taken away for skills not covered. The TV-D formula reintroduced the penalty for non-coverage. Moreover, the penalty for not covering a particular skill is proportional to the "importance" of that skill (i.e., adjusted by the proficiency and difficulty variables). In other words, the credit for coverage and the penalty for non-coverge are both weighted by the same variables.

Prescriptive Mode

In addition to its use in evaluating alternative training devices, an analytic model (such as TRAINVICE) is also needed to provide guidance in the specification of training device characteristics. That is, what is required is a prescriptive model as well as a predictive one. Whether or not both of these functions can be performed by one of the TRAINVICE models (or any other single model) remains to be seen. In all of the TRAINVICE publications, there is only one strategy recommended for the use of a predictive model in the prescriptive mode. This strategy is simply to perform the predictive procedures (ratings, etc.) with a device's design specifications as input. An index of the device's potential training effectiveness (if built) is then generated. If a prediction of poor transfer of training results, the device's design can then be changed in an attempt to improve its effectiveness. The new design can then be evaluated by generating a new prediction; and so on. In other words, the model does not directly specify the most desirable training device characteristics. Rather, the model is used to give feedback on the effectiveness of a proposed device; thus, providing indirect guidance in the design process.

Wheaton, Fingerman, Rose, and Leonard (1976) caution us about such an early (in the life cycle of a device) application of a predictive model. They state that an early application would rely almost exclusively on the Training Device Requirements document (TDR), and that the information in the TDR would be of insufficient breadth and quality to allow performance of the model's preparatory analyses. The only solution offered by Wheaton, Fingerman, Rose, and Leonard (1976) is the reformatting of the TDR.

In the absence of a major change in the TDR's scope and level of detail, the question will remain: Can an analytic model demonstrate an acceptable amount of predictive power when relying on rather unspecific task and equipment information? That is, can a predictive model work with low resolution imput? If the answer to this question is negative, then the other question which remains is: Can a truly (i.e., directly) prescriptive model be developed?

Separate Indices

The overall index of effectiveness, generated by each of the models, would clearly be of use when a choice must be made between two competing training devices. The single figure of merit for each device provides the decision maker with rather straightforward guidance; i.e., a "bottom line". The utility of an overall index would, however, be minimal when decisions must be made concerning: 1) training device design specifications and modifications (prescription); and 2) development of a program of instruction which will complement the strengths and compensate for the weaknesses of a training device (implementation). Either situation demands guidance which is task, or perhaps skill, specific. In other words, what is needed is a separate index of training effectiveness for each task (or skill). Whatever the form that a separate index eventually takes, its development will contribute not only to the task specific questions of design and implementation, but also to the construction of a valid overall index.

IV

Conclusions

This report reviewed the TRAINVICE models for predicting training device effectiveness. The models were presented as they were reported in the original documentation. It is hoped that we have remained faithful to the original authors' intents.

TRAINVICE appears to be a promising method for analytically assessing training device effectiveness during various stages of development. But progress in developing and refining the methodology has been slow. Army decision-makers need and can use a TRAINVICE approach now. Unfortunately, the research community is not ready to field this methodology.

To meet this demand, ARI is conducting programmatic research to validate and refine TRAINVICE methodology. As part of this research, a priori investigations of the mathematical sensitivity and distributional properties of the models are planned. The core of these sensitivity/distribution tests will be computer programs based on each of the TRAINVICE equations. The general procedure to be followed will be the generation of index values, given systematic variation of component variable values.

Validation efforts will consist of comparisons of model predictions and empirically obtained transfer of training data. Efforts are being made to identify a variety of training devices and simulators which have recently been (or will soon be) empirically evaluated. For each device, judgmental data will be collected on the variables considered by each of the analytic models. In this manner, an index of effectiveness can be generated using each model, and all indices can be compared to the same set of empirical data.

In addition to actual field studies, laboratory research will be conducted to test the predictive power of model variables more systematically. The experimental manipulation of these variables will consist of locating or constructing devices which will conform to extreme as well as moderate variable values. It is hoped that the examination of devices which are markedly different from each other, will permit the emergence of reliable effects. A major problem which plagued prior validation efforts was that the devices being compared were not significantly different in the amount of transfer predicted or obtained. The planned approach will help to avoid merely confirming a prediction of the null hypothesis.

An initial milestone is to develop a useable, although interim, version of a model that may be routinely applied to training devices as they progress through various stages in the acquisition cycle. While not expected to be perfect, an evaluation approach which systematically assesses a device, backed-up by guidance on its interpretation seems a possible reality in the foreseeable future.

Since the application of any of the models reviewed here is a fairly burdensome process, an associated milestone will be an automated (i.e.,

computer-based) implementation. The form the implementation is expected to take is an interactive program which will: 1) lead the user through the model procedures and guidance relevant to each judgment; 2) maintain records of all tasks, equipment, and judgmental rating information; and 3) perform all calculations and generate hard copy of predictive indices. This strategy should permit the user to focus almost all of his or her time and energy on making the judgments, which promise to be challenging in any analytic model.

The results of training device evaluations, both analytic and empirical, will ultimately be incorporated into a computer-based management information system. As the data base contained in such a system grows, it will permit training developers and researchers to track the history of individual training devices throughout their life cycles, from initial design to field utilization. Longitudinal training device data will, eventually, support the continuous validation and refinement of both predictive and prescriptive methods.

An investigation of current Army procedures followed in the writing of a Training Device Requirements document (TDR) will also be performed to support the development of prescriptive methods. As mentioned earlier, Wheaton, Fingerman, Rose, and Leonard (1976) identified the shortcomings of the TDR as the major limitation on an early, prescriptive application of analytic evaluation methods. Since the TDR investigation will address the ways in which information is generated and used during the acquisition of a training device, this effort is expected to enhance Army utilization of device evaluation data, and to improve the overall quality of these data.

In reviewing the TRAINVICE models, it became apparent that there is also a need for a thoroughgoing re-examination of the models' underlying assumptions about which characteristics of a training device will foster effective training. In particular, this investigation must concern the applicability of the various sets of Learning Guidelines to specific questions of device evaluation. The Learning Guidelines used in the TRAINVICE models were originally intended to aid in media selection decisions. It is still unknown, however, whether or not the same guidelines are of sufficient detail, or validity, to be of use in the evaluation of the transfer potential of a particular training device. A second problem which needs to be addressed is the assumption that each of the guidelines will promote transfer of training. In some cases guidance appears to be directed primarily toward enhancing the rate at which learning takes place, and in others, toward increasing skill retention. Although rate of learning, retention, and transfer are all considered measures of "good" training," they are not always similarly affected by the same variables. For example, a variable which increases rate of learning may have no effect on retention (Underwood, 1964).

Adequate definitions of each of the Learning Guidelines are needed. Such a definition would consist minimally of an identification of the manipulable parameters (i.e., independent variables) implied by each guideline, and the specific effects of those parameters on rate of learning, retention, and transfer of training. It is certain that the prediction of device effectiveness and the prescription of effective devices, will be greatly buttressed by the guidance which results from this effort. First, an extensive review of the research literature, both basic and applied, will be required to find

sources supporting each guideline and to identify areas in which new empirical research is needed. Once data have been collected the task of generating new guidelines, and of incorporating them into device evaluation procedures will remain. Clearly, the refinement of the Learning Guidelines must be considered a long-term goal.

To recapitulate briefly, our review of the TRAINVICE models has led us to the following general conclusions. Despite their various limitations, the TRAINVICE models are ambitious and promising methods for the analytic evaluation of training device effectiveness. The evident merits of these models warrant a programmatic series of validation and, eventually, implementation efforts. Any significant improvement in predictive or prescriptive methods will require a long-term re-examination of the principles underlying training device effectiveness.

The scope and amount of work outlined above is, admittedly, great. However, the potential utility of analytic evaluation methods and the persistent need for them are at least as great.

References

Boldovici, J.A., Boycan G.G., Fingerman, P.W., Wheaton, G.R. M60A1AOS Tank Gunnery Data Handbook. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences Technical Report TR-79-A7, March, 1979.

Braby, R., Henry, J.M. Parrish, W.F., Swope, W.M. A Technique for Choosing Cost-Effective Instructional Delivery Systems. Orlando, FL: Naval Training Equipment Center TAEG Report No. 16, April, 1975.

Caro, P.W. Equipment-Device Task Commonality Analysis and Transfer of Training. Human Resources Research Organization Technical Report TR-70-7, June, 1970.

Demaree, R.G. Development of Training Equipment Planning Information. USAF ASD Technical Report 61-544, October, 1961.

Evans, R.A. and Swezey, R.W. Development of a User's Guidebook for TRAINVICE II. Alexandria VA: US Army Research Institute for the Behavioral and Social Sciences, Research Note RN 81-28, 1981.

Folley, J.D., Jr. Development of an Improved Method of Task Analysis and Beginnings of a Theory of Training. Valencia, PA: Applied Sciences Associates, Inc. 1964.

Gagne, R.M., Foster, H., Crowley, M.E. The Measurement of Transfer of Training. Psychological Bulletin, 1948, 45, 97-130.

Jeantheau, G.G. Handbook for Training Systems Evaluation. Navy Training Device Center, NAVTRADEVCEN 66-C-0113-2, 1971.

Micheli, G.S. Analysis of the Transfer of Training Substitution, and Fidelity of Transfer Equipment. Orlando, FL: Naval Training Equipment Center TAEG Final Report, 1972.

Narva, M.A. Formative Utilization of a Model for the Prediction of the Effectiveness of Training Devices. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences, Research Memorandum 79-6, May, 1979a.

Narva, M.A. Development of a Systematic Methodology for the Application of Judgmental Data to the Assessment of Training Device Concepts. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences, Research Memorandum 79-7, May, 1979b.

Osgood, C.E. The Similarity Paradox in Human Learning: A Resolution. Psychological Review, 1949, 56, 132-143.

PM-TRADE. Training Device Requirements Documents Guide. Orlando, FL: Naval Training Equipment Center, 5 January, 1979.

Swezey, R.W., Evans, R.A. Guidebook for User's of TRAINVICE II. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences, May, 1980.

TRADOC Pamphlet 350-30: Interservice Procedures for Instructional Systems Development, 1975.

Underwood, J.J. Degree of Learning and the Measurement of Forgetting. Journal of Verbal Learning and Verbal Behavior, 1964, 3, 112-129.

U.S. Navy Training Device Center. Staff Study on Cost and Training Effectiveness of Proposed Training Systems. Orlando, FL: Naval Training Center TAEG Report 1, 1972.

Willis, M.P., Peterson, R.O. Deriving Training Device Implications from Learning Theory Principles. Volumn I: Guidelines for Training Device Design, Development and Use. Port Washington, NY: U.S. Naval Training Device Center, 1961.

Wheaton, G.R., Rose, A.M., Fingerman, P.W., Korotkin, A.L., Holding, D.H. Evaluation of the Effectiveness of Training Devices: II Evaluation Plan for the Preliminary Model. Second Interim Report. U.S. Army Research Institute. Washington, D.C.: American Institutes for Research, 1974.

Wheaton, G.R., Fingerman, P.W., Rose, A.M., Leonard, R.J., Jr. Evaluation of the Effectiveness of Training Devices; Elaboration and Application of the Predictive Model. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences, RM 76-16, July, 1976.

Wheaton, G.R., Fingerman, P.W., and Boycan, G.G. Development of a Model Tank Gunnery Test. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences, TR 78-A24, August, 1978.

Wheaton, G.R., Rose, A.M., Fingerman, P.W., Korotkin, A.L., Holding, D.H. Evaluation of the Effectiveness of Training Devices: Literature Review and Preliminary Model. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences, RM 76-6, April, 1976.

APPENDIX

Rating Scales Used by OTV

Class of variable	Analysis name	Variable label	Variable value	Value description
Coverage requirement	none	none	none	none
Communality	Task communality	C	1	The device could or does enable the trainee to practice the subtask in question.
			0	The particular subtask is not represented in the device.
Physical/functional similarity	Physical similarity analysis	(P)	3	Identical. The trainee would not notice a difference between the training device control or display and the operational control or display at the time of transfer. Note that they need not be absolutely identical, but there must be no "jnd" (just noticeable difference) for the trainee. Include for consideration the location, appearance, feel, and any other physical characteristics. Ignore the amount and quality of information transmitted.
			2	Similar. There would be a jnd for the trainee at the time of transfer, but he would be able to perform the task. There might be a decrement in performance at transfer, but any such decrement would be readily overcome.
			1	Dissimilar. There would be a large noticeable difference, quite apparent to the trainee, at transfer, and a large performance decrement, given that the trainee could perform at all. Specific instruction and practice would be required on the operational equipment after transfer to overcome the decrement.
			0	The control or display is not represented at all in the training device.

Rating Scales Used by OTV (Continued)

Class of variable	Analysis name	Variable label	Variable value	Value description
Physical/ functional similarity (cont'd)	Functional similarity analysis	(F)	3	Identical. $H_{ts} = H_{os}$ where H_{ts} is the amount of information in the training setting given a flow diagram of each subtask; H_{os} is the amount of information in the operational setting. The amount of information in a given setting is equal to the log (base 2) of the number of states in the stimulus or response functions under consideration.
			2	Similar. $H_{ts} \approx H_{os}$; they are within one $\log_2$ unit of each other.
			1	Dissimilar. $H_{ts} \neq H_{os}$; they are more than one $\log_2$ unit apart.
			0	Missing. $H_{os} > 0$ and $H_{ts} = 0$.
Learning	Learning deficit analysis	RS	4	Has a complete understanding of the subject or skill. Can do the task completely and accurately without supervision. Has received "skill" training.
			3	Understands the subject or skill to be performed. Has applied part of the knowledge or skill either on the actual job or a trainer. Has done the job enough times to make sure he can do it, although perhaps only with close supervision. Needs more practice under supervision. Has had "procedural" training.
			2	Has received a complete briefing on the subject or skill. Can use the knowledge or skill only if assisted in every step of the operation. Requires much more training and experience. Has received "familiarization" training only.

Rating Scales Used by OTV (Continued)

Class of variable	Analysis name	Variable label	Variable value	Value description
Learning (cont'd)	Learning deficit analysis (cont'd)		1	Has only limited knowledge or skill of this subject or skill. Has not actually used the information or skill. Cannot be expected to perform. Has had "orientation" only.
			0	No experience, training, familiarity, etc. with this skill or knowledge. Cannot perform a task requiring this skill or knowledge.
		CS	4	Should have a complete understanding of the subject, or be highly skilled. Is able to perform the task completely, accurately, and independently. Has had "skill" training.
			3	Should have understanding of the subject or skill to be performed. Has applied part of the knowledge or skill on the actual job or a trainer. Has done the job enough times to make sure he can do it, although perhaps only with close supervision. Needs more practice under supervision. Has had "procedural" training.
			2	Should have received a complete briefing on the subject or task. Is able to use the knowledge or skill only if assisted in every step of the operation. Requires much more training and experience to be able to perform the task independently. Has had "familiarization" training.
			1	Should have a limited knowledge of the subject or skill. Has not actually used the information. Is not expected to perform the task. Has completed "orientation" training.
			0	At the end of training, the trainee should have no experience or training.

Rating Scales Used by OTV (Continued)

Class of variable	Analysis name	Variable label	Variable value	Value description
		(RANK)	rank values	Ranks assigned to the subtasks in terms of estimated training time assuming that only the operational equipment would be available for training.
Training technique	Training technique analysis	Behavioral categories		The behavioral categories used are the same as the ones used in U.S. Naval Training Device Center: Staff study on cost and training effectiveness of proposed training systems. TAEG Report 1, U.S. Naval Training Center, Orlando, Florida, 1972. The categories are:
			1	Recalling facts and principles
			2	Recalling procedures
			3	Nonverbal identification
			4	Nonverbal detection
			5	Using principles, interpreting, inferring
			6	Making decisions
			7	Continuous movement
			8	Verbal detection and identification
			9	Positioning and serial movement
			10	Repetitive movement
			11	Written verbalization
			12	Oral verbalization
			13	Other verbalization, including signs
		Learning principles		A list of "training principles/techniques" based on a classification by Willis & Peterson (Deriving training device implications from theory principles. Volume I: Guidelines for training device design, development, and use. U.S. Naval Training Device Center, Port Washington, New York, 1961, AD 264 364.) The original TRAINVICE model classifies these principles into stimulus, response, and feedback categories.

Rating Scales Used by OTV (Continued)

Class of variable	Analysis name	Variable label	Variable value	Value description
Training technique (cont'd)	Training technique analysis (cont'd)	T	3	Optimal implementation of this technique; in complete accord with this principle.
			2	Good implementation of this technique; in excellent accord with this principle.
			1	Fair implementation of this technique; good accord with this principle.
			0	This principle or technique was inapplicable or irrelevant OR The device neither implemented this technique nor violated this principle.
			-1	Mild violation of this training principle; implementation of a mildly opposing technique.
			-2	Serious violation of this principle or technique.
			-3	Complete violation of this principle; implementation of strongly contraindicated technique.

Rating Scales Used by HTV

Class of variable	Analysis name	Variable label	Variable value	Value description
Coverage requirement	none	none	none	none
Communality	Task commonality index		1	Training device does allow practice of that operational task element.
			0	Particular task element is not represented in the training device, either because the task is truncated or simplified.
Similarity	Physical similarity index		3	Identical. The trainee would not notice a difference between the training device control or display and the operational control or display when he moves from the training to the job situation. Include for consideration the location, appearance, feel, and any other physical characteristics. Ignore the amount and quality of information transmitted.
			2	Similar. There would be a small noticeable difference for the trainee between the training device control or display and the operational control or display, but he would be able to perform the task. There might be a decrement in performance, but any such decrement would be small and readily overcome.
			1	Dissimilar. There would be a large noticeable difference quite apparent to the trainee, between the training device control or display and the operational control or display and a large performance decrement, given that the trainee could perform at all. Specific instruction and practice would be required on the operational equipment after practice on the training device to overcome the decrement.
			0	Missing. The control or display is not represented at all in the training device.

Rating Scales Used by HTV (Continued)

Class of variable	Analysis name	Variable label	Variable value	Value description
Similarity (cont'd)	Functional similarity index		3	Identical. The number of states in the training situation is the same as the number of states in the operational setting.
			2	Similar. The number of states in the training situation is at least half of the number of states in the operational setting.
			1	Dissimilar. The number of states in the training situation is less than half of the number of states in the operational setting.
			0	Missing. The control or display is not represented at all in the training device.
Learning	Skills & knowledge requirements index	(BT)	4	Has a complete understanding of the subject or skill. Can do the task completely and accurately without supervision. Has received "skill" training.
			3	Understands the subject or skill to be performed. Has applied part of the knowledge or skill either on the actual job or a trainer. Has done the job enough times to make sure he can do it, although perhaps only with close supervision. Has had "procedural" training.
			2	Has received a complete briefing on the subject or skill. Can use the knowledge or skill only if assisted in every step of the operation. Requires much more training and experience. Has received "familiarization" training only.
			1	Has only limited knowledge of this subject or skill. Has not actually used the information or skill. Cannot be expected to perform. Has had "orientation" only.

Rating Scales Used by HTV (Continued)

Class of variable	Analysis name	Variable label	Variable value	Value description
Learning (cont'd)	Skills & knowledge requirements index (cont'd)	(BT)	0	No experience, training, familiarity, etc., with this skill or knowledge. Cannot perform a task requiring this skill or knowledge.
	Skills & knowledge requirements index	(AT)	4	Should have a complete understanding of the subject or be highly skilled. Is able to perform the task completely, accurately, and independently. Has had "skill" training.
			3	Should have an understanding of the subject or skill to be performed. Has applied part of the knowledge or skill on the actual job or a trainer. Has done the job enough times to make sure he can do it although perhaps only with close supervision. Needs more practice under supervision. Has had "procedural" training.
			2	Should have received a complete briefing on the subject or task. Is able to use the knowledge or skill only if assisted in every step of the operation. Requires much more training and experience to be able to perform the task independently. Has had "familiarization" training.
			1	Should have limited knowledge of the subject or skill. Has not actually used the information. Is not expected to perform the task. Has completed "orientation" training.
			0	At the end of training, the trainee should have no experience or training.

Rating Scales Used by HTV (Continued)

Class of variable	Analysis name	Variable label	Variable value	Value description
Learning (cont'd)	Task training difficulty index	(D)	4	Requires as much time to train as the most time-consuming task element, considering all task elements for all tasks in the current analysis.
			3	Requires substantial training time, but less than above.
			2	Requires a moderate amount of training time relative to the most time-consuming task element.
			1	Requires only minimal training time relative to the most time-consuming task element.
			0	Requires no training time.
Training techniques	none	none	none	none

Rating Scales Used by NTV

Class of variable	Analysis name	Variable label	Variable value	Value description
Coverage requirement	Coverage requirements analysis	CR	1	Skill/knowledge "is judged to require its presence in the training situation."
			0	Skill/knowledge is not required in training setting.
Communality	Coverage analysis	C	1	Skill/knowledge in operational setting is also present in training setting.
			0	Skill/knowledge in operational setting is not present in training setting.
Physical/functional similarity	none	none	none	none
Learning	Training criticality analysis	Ci	4	Should have complete understanding of subject, or be highly skilled.
			3	Should have understanding of subject or skill to be performed.
			2	Should have received complete briefing on subject or skill.
			1	Should have limited knowledge of subject or skill.
	Training difficulty analysis	D		This rating is the degree of difficulty in attaining the preceding level of proficiency for a given skill/knowledge:
			4	Substantial
			3	Much
			2	Some
			1	Minimal or none

Rating Scales Used by NTV (Continued)

Class of variable	Analysis name	Variable label	Variable value	Value description
Training technique	Device characteristics analysis: Physical characteristics	Behavioral categories		The behavioral categories used are taken from the ISD categories (TRADOC Pam 350-30, Appendix A, pp. 34 through 105). However, only 10 of the 11 ISD categories are used; ISD category 11, "Attitude Learning," was dropped for the TRAINVICE application. The 10 behavioral categories used are:
			1	Rule learning and using
			2	Classifying-recognizing patterns
			3	Identifying symbols
			4	Detecting
			5	Making decisions
			6	Recalling bodies of knowledge
			7	Performing gross motor skills
			8	Steering and guiding--continuous movement
			9	Positioning movement and recalling procedures
			10	Voice communications
		Learning guidelines		The ISD list of guidelines under each of the preceding behavioral categories (see TRADOC Pam 350-30, Appendix A, pp. 34-105). These may be flagged by P, F, or P/F as was done with the SAI TRAINVICE.
		Cue		An undefined list of "cues" provided by the skill/knowledge (see p. 13 of Narva ARI Research Memorandum 79-6, 1979).
		Response		An undefined list of "responses" "subsumed under each skill" (see p. 13 of Research Memorandum 79-6, 1979).

Rating Scales Used by NTV (Continued)

Class of variable	Analysis name	Variable label	Variable value	Value description
Training technique (cont'd)	Device characteristics analysis: Physical characteristics (cont'd)	Generic characteristics		A taxonomy for classifying controls and displays by "stimulus capabilities." The taxonomy is from Braby, R., Henry, J., Parrish, W., and Swope, W. A technique for choosing cost-effective instructional delivery systems. TAEG Report No. 16, U.S. Naval Training Center, Orlando, Florida, 1975.
		Trainee response mode		A taxonomy for classifying trainee response mode for controls/displays of a given training device "subsystem." The taxonomy is due to Braby et al., 1975 (see Generic characteristics).
		PC		For each "Generic characteristic" and "Response mode," rate how well it implements set of "Learning guidelines":
			3	Outstanding implementation for requirements/guidelines.
			2	Good implementation for requirements/guidelines.
			1	Adequate implementation for requirements/guidelines.
			0	Not adequate for requirements/guidelines.
	Functional characteristics analysis	Behavioral categories		Same as under "Physical characteristics analysis."

Rating Scales Used by NTV (Continued)

Class of variable	Analysis name	Variable label	Variable value	Value description
Training technique (cont'd)	Functional characteristics analysis (cont'd)	Learning guidelines		Same as under "Physical characteristics analysis."
		Functional dynamic characteristics		Some type of undefined list for each skill/knowledge.
		FC	3	Outstanding implementation of guideline.
			2	Good implementation of guideline.
			1	Adequate implementation of guideline.
			0	Not adequate implementation of guideline.

Rating Scales Used by STV

Class of variable	Analysis name	Variable label	Variable value	Value description
Coverage requirement	Coverage requirement analysis	CR	1	Should be covered by the device; required in training (Sk/Kn).
			0	Not necessary; not required (skill/knowledge).
Communality	Coverage analysis	C	1	Skill or knowledge is covered by training device.
			0	Skill or knowledge is not covered by training device.
Physical/functional similarity	none	none	none	none
Learning	Proficiency analysis	P	4	Should have expert-level knowledge of subject and/or outstanding skill capability; errors in performance are rare; performance is excellent/superior.
			3	Should have adequate knowledge of subject or skill to assure reliable performance; errors in performance are infrequent to rare; performance can be characterized as smooth and experienced.
			2	Should have minimally competent knowledge of subject or skill for performing job or operating system; some errors in performance occur regularly, but basically the individual can sustain a minimally acceptable (or) "novice level" of performance.
			1	Should have limited knowledge of subject or skill; has at least been briefed on the subject or performed the skill once; however, system effectiveness would probably be seriously degraded by performance at this level of proficiency.

Rating Scales Used by STV (Continued)

Class of variable	Analysis name	Variable label	Variable value	Value description
Learning (cont'd)	Learning difficulty analysis	D	4	Highly difficult: Trainee requires extensive instruction, practice and/or study to accomplish the activity; requirements of learning at least border on expert performance standards.
			3	Difficult: Trainee can accomplish the activity following instruction, but only with consistent practice and/or study.
			2	Modestly difficult: Trainee can accomplish most of the activity subsequent to instruction with little practice or study; some of the activity does require minimal practice/study to sustain competent performance at the desired level of proficiency.
			1	Easy: Trainee can accomplish the activity once informed that it exists; virtually no practice or study is required.
Training technique	Physical characteristics analysis	Behavioral categories		The behavioral categories used are taken from the ISD categories (TRADOC Pam 350-30, Appendix A, pp. 34 through 105). However, only 10 of the 11 ISD categories are used; ISD category 11, "Attitude Learning," was dropped for the TRAINVICE application. The 10 behavioral categories used are:
			1	Rule learning and using
			2	Classifying-recognizing patterns
			3	Identifying symbols
			4	Detecting
			5	Making decisions
			6	Recalling bodies of knowledge
			7	Performing gross motor skills
			8	Steering and guiding--continuous movement
			9	Positioning movement and recalling procedures
			10	Voice communications

Rating Scales Used by STV (Continued)

Class of variable	Analysis name	Variable label	Variable value	Value description
Training technique (cont'd)	Physical characteristics analysis (cont'd)	Learning guidelines (P or P/F prefix)		A list of "Learning Guidelines for Good Instructional Practice" taken from ISD (TRADOC Pam 350-30, Appendix A, pp. 34-105). The guidelines are classified by the preceding 10 behavioral categories. In addition, a P (for Physical) or a P/F (for both Physical and Functional) has been placed by each guideline to indicate the appropriate type of technique given the analysis.
		Generic characteristics		A taxonomy for classifying controls and displays by "stimulus capabilities." The taxonomy is from Braby, R., Henry, J., Parrish, W., and Swope, W. A technique for choosing cost-effective instructional delivery systems. TAEG Report No. 16, U.S. Naval Training Center, Orlando, Florida, 1975.
		Trainee response mode		A taxonomy for classifying trainee response mode for controls/displays of a given training device "subsystem." The taxonomy is due to Braby et al., 1975 (see Generic characteristics).
		PC		For each Generic characteristic and Response mode, rate how well it implements set of Learning guidelines:
			3	Highly proficient implementation of guidelines; implementation is of exceptional quality (or) is identical with guidelines.
			2	Proficient implementation of guidelines; implementation is adequate to good instructional quality.
			1	Partially proficient implementation of guidelines; at least minimal implementation is achieved.
			0	Extremely deficient implementation of guidelines (or) no implementation at all.

Rating Scales Used by STV (Continued)

Class of variable	Analysis name	Variable label	Variable value	Value description
Training technique (cont'd)	Functional characteristics analysis (cont'd)	Behavioral categories		Same as under "Physical characteristics analysis" above.
		Learning guidelines (F or P/F prefix)		A list of "Learning Guidelines for Good Instructional Practice" taken from ISD (TRADOC Pam 350-30, Appendix A, pp. 34-105). The guidelines are classified by the preceding 10 behavioral categories. In addition, an F (for Functional) or a P/F (for both Physical and Functional) has been placed by each guideline to indicate the appropriate type of technique given the analysis.
		FC		For each control or display, rate how well its Functional characteristics implement above Learning guidelines.
			3	Highly proficient implementation of guidelines; implementation is of highly exceptional quality (or) identical with guidelines.
			2	Proficient implementation of guidelines; implementation is adequate to good in quality.
			1	Partially proficient implementation of guidelines; at least minimal implementation is achieved.
			0	Extremely deficient implementation of guidelines (or) no implementation at all.

55

011485