AFHRL-TR-81-44(II)

(12)

AIR FORCE HUMAN RESOURCES LABORATORY

AD A112570

DTIC FILE COPY

20000728010

OPERATIONAL TEST AND EVALUATION HANDBOOK FOR AIRCREW TRAINING DEVICES: OPERATIONAL EFFECTIVENESS EVALUATION

By

William V. Hagin
Stephen R. Osborne
Roik L. Hockenberger
James P. Smith
Seville Research Corporation
400 Plaza Building
Pensacola, Florida 32505

Thomas H. Gray

OPERATIONS TRAINING DIVISION
Williams Air Force Base, Arizona 85224

February 1982

Final Report

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235

NOTICE

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

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

This report has been reviewed and is approved for publication.

THOMAS H. GRAY
Contract Monitor

MILTON E. WOOD, Technical Director
Operations Training Division

RONALD W. TERRY, Colonel, USAF
Commander

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

REPORT NUMBER: AFHRL-TR-81-44(II)	2. GOVT ACCESSION NO.: AD-A112 570
3. RECIPIENT'S CATALOG NUMBER:	
TITLE (and Subtitle): OPERATIONAL TEST AND EVALUATION HANDBOOK FOR AIRCREW TRAINING DEVICES: OPERATIONAL EFFECTIVENESS EVALUATION	5. TYPE OF REPORT & PERIOD COVERED: Final
	6. PERFORMING ORG. REPORT NUMBER:
AUTHOR(s): William V. Hagin, Stephen R. Osborne, Roik L. Hockenberger, James P. Smith, Thomas H. Gray	8. CONTRACT OR GRANT NUMBER(s): F33615-78-C-0063
PERFORMING ORGANIZATION NAME AND ADDRESS: Seville Research Corporation, 400 Plaza Building, Pensacola, Florida 32505	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 62205F, 11231103
CONTROLLING OFFICE NAME AND ADDRESS: HQ Air Force Human Resources Laboratory (AFSC), Brooks Air Force Base, Texas 78235	12. REPORT DATE: February 1982
	13. NUMBER OF PAGES: 288
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Operations Training Division, Air Force Human Resources Laboratory, Williams Air Force Base, Arizona 85224	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)

aircrew training devices
human factors engineering
questionnaires
rating scales
system operability
study design
test and evaluation
transfer of training

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

The Handbook is comprised of three volumes and is intended to provide guidelines and procedures appropriate for Air Force Operational Test and Evaluation (OT&E) personnel to use in planning, conducting and reporting the results of simulator assessment efforts. Although of value to all test personnel, it is primarily for the typical novice test manager/director—a person who has subject matter expertise (e.g., a qualified pilot or operator), but who may have little or no previous OT&E experience. The Handbook provides detailed coverage on OT&E planning and management with special emphasis on measuring device operational effectiveness and suitability. In accord with its objectives, the Handbook was prepared to serve as a supplement to Air Force Manual

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

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Item 20 (*Continued*)

55-43, "Management of Operational Test and Evaluation," by providing those specific additional evaluation concepts and techniques necessary for aircrew training device test and evaluation.

PREFACE

This volume (Volume II. Operational Effectiveness Evaluation) is one part of a three-volume Handbook produced for the U.S. Air Force Human Resources Laboratory/Operations Training Division (AFHRL/OT). The Handbook is entitled, "Handbook for Operational Test and Evaluation (OT&E) of the Training Utility of Air Force Aircrew Training Devices." This effort has been accomplished by the Seville Research Corporation under Contract No. F33615-78-C-0063. Dr. Thomas H. Gray served as the Air Force Laboratory Contract Monitor (AFLCM) on the project. For Seville, Dr. William H. Hagin was Project Director, and Dr. Wallace W. Prophet was Program Manager.

The three volumes which comprise the total Handbook are intended to provide guidelines and procedures appropriate for use of Air Force ATD OT&E test team personnel in planning, conducting, and reporting the results of aircrew training device OT&E efforts. The three Handbook volumes are:

Volume I. Planning and Management

Volume II. Operational Effectiveness Evaluation

Volume III. Operational Suitability Evaluation

It is important that the reader understand that this Handbook was prepared to serve as a supplement to AFM 55-43, "Management of Operational Test and Evaluation" by providing those specific additional evaluation concepts and techniques necessary for ATD test and evaluation.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

LIST OF TABLES

LIST OF FIGURES

TABLE OF CONTENTS (Continued)

LIST OF FIGURES (Continued)

LIST OF FIGURES (Continued)

CHAPTER 1

INTRODUCTION

PURPOSE OF THE VOLUME

The U.S. Air Force has made a substantial commitment to the use of Aircrew Training Devices (ATDs)[1] for aircrew training. It is important that the Air Force has a procedure for assuring that these devices do train efficiently and effectively. That procedure is provided by the Operational Test and Evaluation (OT&E) process as outlined in AFR 80-14, "Test and Evaluation," in AFM 55-43, "Management of Operational Test and Evaluation," and in this Handbook, "Operational Test and Evaluation of Aircrew Training Devices."

Volume I of this three volume Handbook has provided a "road-mapping" of the total ATD OT&E process as it has been carried out within the Air Force by the Air Force Test and Evaluation Center (AFTEC) and the Major Commands (MAJCOMs). Volume I is basically an event and milestoning document telling what is to be done, when it occurs, and who does it. This volume of the Handbook, Volume II, is intended to provide the ATD OT&E test director (or any other users) with information and guidance concerning the actual test and evaluation of ATD training effectiveness. It will provide him first with an understanding of the relationship between the learning processes which the ATD exploits and the various evaluation design options open to him. It will also acquaint him with the many conditions and operational constraints that can influence his choice of an evaluation design. Proper concern for these conditions and constraints will be critical to the successful execution of any design he attempts to implement.

Perhaps the most important things this Handbook volume does are: (1) it gives the test director general directions concerning the selection of an evaluation approach best suited for the particular situation he faces; and (2) it provides detailed instructions for the application of the specific evaluation methodologies he may choose to utilize. These instructions range from questionnaire development procedures to models for the accomplishment of transfer of training (TOT) evaluations. By closely following the guidance provided and through

---

[1] As noted in Volume I, the term aircrew training device (ATD) has become generally accepted as including cockpit familiarization and procedures trainers, part task trainers, and mission trainers. In this Handbook, the term ATD is used to refer to all such training devices of sufficient cost and/or complexity to justify OT&E.

use of the techniques made available in this Handbook volume, the relatively inexperienced test director can markedly increase the likelihood that a competent ATD training effectiveness evaluation will result.

ORGANIZATION OF VOLUME II

This volume of the Handbook consists of seven chapters, including this Introduction. The next chapter, Chapter 2, "Aircrew Training," first provides the ATD test director with an overview of the principles of cue/response discrimination and generalization learning. That overview is intended to provide a necessary basic awareness of the learning principles upon which the effective use of any ATD depends. The discussion in Chapter 2 then addresses the nature of aircrew trainee learning tasks. In that discussion, a critical distinction is made between the learning task activities of beginning aircrew trainees and those of experienced, combat-ready crewmen. The differentiation made therein between the learning behaviors of these two categories of trainees has important implications for subsequent ATD utilization and effectiveness evaluation procedures. Two key points are made: (1) A given ATD may not be equally effective for initial skills acquisition training and for skills maintenance; and (2) the evaluation methods suitable for use in one training situation (e.g., undergraduate training) may not necessarily be the most appropriate for use in another situation (e.g., continuation training).

The third chapter proceeds to define ATD training effectiveness in terms which are considered particularly relevant to ATD OT&E. Chapter 3 also reviews a number of the effectiveness evaluation methodologies that have been used in the past to evaluate training devices and programs. Much of the discussion in this chapter is being provided as background information that will be helpful when the time comes for the test director to commit to a particular evaluation approach.

Chapter 4 surfaces a number of factors that can influence the test directors's final decision regarding that evaluation methodology best suited for his particular evaluation situation. Among the many factors which can have a substantial impact on his implementation of an analytic versus a demonstration approach to the evaluation, three of the more critical are discussed in detail: management's commitments, criterion measurement, and user attitudes. The treatment given attitude effects is particularly important because of their possible effects during the OT&E.

The chapter concludes with a specific algorithm for making the choice between an "analytic" type evaluation which depends on the use of subject matter experts or a demonstration type of evaluation which is keyed to actual training.

Once he has decided which of these two evaluation approaches to use--analytic or demonstration--the test director will need detailed guidance concerning how to use specific evaluation tools and techniques. That level of information is provided in the next four chapters of this volume.

Chapter 5 first tells the ATD OT&E test director what he needs to know about questionnaires. It describes their uses and limitations, and it provides specific instructional guidance concerning the construction of questionnaires. It should be noted that Chapter 5 does not make a strong case for the use of questionnaires, but it does stress the importance of being as rigorous as possible whenever their use is required.

Chapter 5 next provides information regarding rating scales. It describes their uses and limitations; it tells how to construct rating scales; and it discusses the analysis of rating results. It should be noted that both questionnaires and rating scales are most likely to be useful during in-plant test and evaluation and analytic types of I/QOT&E.

Chapter 6 is devoted to transfer of training (TOT) evaluation methods which depend, not on subject matter or other user evaluative judgments obtained by means of questionnaires and/or rating scales, but upon the conduct and evaluation of actual hands-on training activities. Where feasible to implement, the TOT evaluation methodology should ordinarily be preferred over either questionnaires or rating scales. Since this method of ATD effectiveness evaluation is also the most difficult to implement--even by reasonably sophisticated training technologists--great care has been taken to make this chapter as explicitly instructional as possible.

As was noted earlier in the Handbook, ATD effectiveness is a function of both trainee station characteristics and the instructional features of the instructor/operator station (IOS). The final chapter of this Handbook volume, Chapter 7, deals with the procedures for evaluating the efficiency and potential effectiveness of the IOS. This has been done since the utility and efficiency of the IOS is necessarily evaluated independently of that of the trainee station, even though there obviously is a necessary interaction between the trainee and instructor stations during actual ATD utilization.

CHAPTER 2

AIRCREW TRAINING

INTRODUCTION

The ATD OT&E test director will normally be a person who is a subject matter expert in the operation of the airborne equipment to which the ATD is related or of equipment similar to it, but who has had little, if any, test and evaluation experience. As such, he will have progressed through the various phases of aircrew training that were prerequisite to reaching his current level of proficiency. The names given these phases--undergraduate, transition, and continuation training--convey an immediate understanding from a management perspective of the intended purposes of such training. Undergraduate training (or initial crew training, as the case might be) develops initial job skills, transition training deals with application of these basic operator skills to different operational hardware, and continuation training addresses hardware-specific skills maintenance.

For the pilot, that training will have consisted of undergraduate flying training, transition crew training on one or more operational aircraft, and continuation training conducted at regular intervals to maintain his peak level of proficiency. Other aircrewmen progress through a basically similar multiphased program. For some aircrew positions (e.g., the navigator), there are the same three phases: undergraduate, combat crew, and continuation training. For others (e.g., the boom operator), there may be only two phases: initial combat crew and continuation training.

As a consequence of his own training experiences, the typical ATD OT&E test director will have reasonable insights concerning the content of that training, the sequencing of instruction, and even the relative efficacy of device training, insofar as ATDs were involved. More than likely, it was in recognition of such competencies as a subject matter expert and instructor that he was designated to be the ATD OT&E test director. It is less likely, however, that the ATD OT&E test director-to-be will be expert regarding the ways in which the underlying learning technology relates to these three types of aircrew training.

It is important that test directors be knowledgeable concerning the fundamental differences in the learning tasks faced by aircrew trainees during the above described various phases of Air Force aircrew training, i.e., initial, transition, and continuation training. Test directors must also recognize that these differences in the nature of learning tasks should be reflected in both the design and

utilization of the ATDs supporting these separate phases of aircrew training. It is also important that they appreciate the fact that these differences in learning behaviors will influence selection of an evaluation design most appropriate for the particular ATD OT&E with which they are concerned.

It is, therefore, the purpose of this Handbook chapter, first to familiarize the ATD OT&E test director with the nature of initial aircrew learning, and then to show how the learning behaviors of both the transitioning aircrewman and the highly skilled, combat-ready aircrewman differ from that initial learning activity. The discussion is not intended to make test directors expert instructional technologists with respect to aircrew training, but it is intended to convey a general understanding of the learning processes whereby aircrew skills are developed and honed, whether training is conducted in the air or on the ground. It should also convey an understanding of the differences in learning behavior between initial learning activity and later skill maintenance and retention efforts.

INITIAL AIRCREW LEARNING

Information concerning the outside world and an individual's "place" in that world comes in the form of stimuli. Stimuli are physical objects, events, or energy that can activate a sense receptor. Some stimuli may come from sources external to a perceiver (e.g., reflected light stimulating the eye, sound reaching the ear, pressures touching the skin surfaces, etc.), while other stimuli may be internal in origin (e.g., kinesthetic sensations arising from body movement, positioning information from joints, etc.).

The aircrew trainee is constantly receiving such stimuli from a number of sources. He sees the horizon; he reads the instruments; he feels the aircraft movements; and he hears the sounds of flight. As he practices his inflight tasks, these stimuli take on specific meaning in relationship to those tasks. For example, when flying under turbulent conditions, g-forces on the body stimulate internal receptors in the muscles, and instrument fluctuations stimulate the eye. At first these stimuli have little meaning since the trainee does not know what information to extract from these stimuli. As he gains experience, however, he learns to derive pertinent information from those stimuli so that the proper responses can be made.

Cue and Response Discrimination Learning

The process whereby a beginning trainee learns to pick out the relevant cues and to select the correct responses is called discrimination. The first step in this discrimination learning process involves simply recognizing that a given stimulus or response is different from another stimulus or response. Learning that the

disappearance of the horizon at the top of a canopy means something different from its disappearance below the line of the instrument panel is one example of this first step. Learning that feeling light in the seat has a different meaning from feeling heavy is another example.

The second step in the process of discrimination learning occurs when there is a response-related meaning which can be attached to the different "positions" of the horizon line in its perceived relationship to the canopy or to the light-heavy feeling of the seat. For the "horizon" example stated above, to maintain level flight, the corrective movement of the stick would be backward for the first condition, but forward for the second. This same corrective stick movement would also apply to the seat pressure feeling: backward for the light feeling; forward for the heavy sensation. In other words, the stimuli, "position of the horizon line relative to the canopy," and "light-heavy seat pressure" now have become specific cues in terms of stick response.

A third, and final, step in this process is required before skilled aircrew performance results. That step is a repetitive refinement in the cue/response, discrimination/action loop. The trainee may have a clear understanding of the cue-response relationships and yet be unable to perform the required behaviors. As every aircrew instructor knows, this process requires extensive practice before increasingly smooth and precise aircraft control is acquired.

This simple definition of the discrimination learning process should not be interpreted to suggest that discriminations are simple processes or that they can be easily learned. The more complex the skill involved in aircrew performance, the larger the number of moment-to-moment, even instant-to-instant, discriminations that must be made. Also, as task complexity increases, discriminations may depend upon very subtle differences in patterns of numerous stimuli. The difference between the performance of a novice and an expert is that the expert has learned to derive more detailed information from the stimulus patterns to serve as more refined cues. He can discriminate subtle differences that a novice cannot. Further, he also can translate the subtle information provided by these refined cues into subtler (fine grain) control inputs. Thus, initial acquisition of skilled performance by the beginner involves learning how to discriminate the finer, more subtle cues and cue patterns that are seen by the expert, and then being able to make those more precise responses to them as the expert does.

Cue and Response Generalization Learning

The next level of learning required of the novice is that he be able to use his newfound skills in more than one specific situation. Application of skills previously learned in one situation to a

situation different from that in which they were first learned involves a process called generalization. The importance of generalizations is so obvious that the need to consider them in training often is overlooked. Without generalization, people would have to learn anew to cope with every situation they encounter.

Generalization occurs to the extent that any given new situation is interpreted by the learner as being similar to a previously experienced situation. This similarity assessment is based on the information conveyed by the stimuli present in the two situations. For example, procedures learned in a low fidelity cockpit mock-up can be generalized to (i.e., performed in) a high fidelity simulator or an actual aircraft. Although the two may differ as to actual stimuli presented, the "cueing" information content present in the mock-up and the corresponding responses to be made are sufficiently similar to those of the more sophisticated device or aircraft that the proper responses can be made. It is this kind of generalization that underlies the training effectiveness of such simple devices.

Workload Management Learning

As the beginner's cue-response discrimination and generalization skills approach those of the expert, he becomes increasingly capable of dealing with more complex tasks and of handling larger numbers of simultaneous task activities. For example, at first the novice pilot has his hands full keeping the wings level and holding a constant attitude. Soon, however, he not only can do these two things well, but he can do so under a variety of inflight situations. Next he finds himself increasingly involved in performing numbers of complex tasks on a time-shared basis. He not only is flying the aircraft skillfully, but he is also manipulating a number of other systems at the same time--he flies, operates radios, navigates, etc. Finally, he learns how to prioritize his task activities, and when to attend to that which is at the moment most critical. He has now become adept at managing his workload. Having learned all the necessary cue/response discriminations and generalizations necessary for competent airborne performance and having mastered the required workload management skills, he can be considered a qualified pilot.

Factors Affecting Initial Learning

There are many, many factors which are known to affect the above described initial learning behaviors, e.g., syllabus design, instructional strategies, practice effects, motivation, etc. Obviously, not all of these factors can be addressed in this Handbook. There are two, however, which are especially important to effective ATD utilization and subsequent training transfer: practice and motivation. ATDs are particularly capable of optimizing practice, and their effectiveness is very much a function of user motivation. These factors are discussed below.

Practice. The practice provided by a training program entails structured repetitions of a task activity accompanied by information (feedback) concerning the appropriateness of performance. It is through practice of similar tasks that cue and response discriminations are learned and through practice of dissimilar tasks (with common components) that generalizations of cues and response are learned. Practice also improves workload management capabilities. Basically, "practice makes perfect."

The degree of original learning that results from practice is determined in part by the conditions under which practice occurs. There are numerous conditions of practice that are important, although the usefulness of a particular condition depends somewhat upon the type of task to be learned.

Task size: One condition of practice concerns the size of the units practiced during the training session. When "whole" procedures are used, the learner practices the task as a single unit, whereas when "part" procedures are used, the task is divided into components that are learned and practiced separately. In general, if components of a task are highly organized (inter-related), an increase in task complexity leads to whole methods being more efficient than part methods; if a task has "low" component organization, an increase in task complexity leads to part methods being more efficient.

Trial spacing: Another aspect of practice concerns the spacing of practice trials, the length of training sessions, and the intervals between training sessions. For example, some data suggest that spaced practice for motor skills typically is more effective in acquisition and leads to better retention than "massed" or highly concentrated practice. In general, spaced or distributed practice seems to be superior for learning lengthy or difficult material, whereas massed practice seems to be more effective for learning short, relatively simple material.

Feedback: No matter how long a task is practiced, performance is not likely to improve unless the learner finds out which aspects of his performance are correct and which are incorrect. Knowledge of results, or feedback, refers in general to those conditions in a performance situation which inform the learner about the accuracy of his performance and progress. The information may be qualitative or quantitative. It can be provided by the instructor or training system, or it can be provided by the task itself. One goal of training is to teach students to recognize feedback that occurs naturally as a result of their actions. As this occurs, students acquire the ability to tell for themselves when discriminations and responses are appropriate by noting their outcomes. In this manner, task-intrinsic feedback replaces the artificial feedback (e.g., instructor provided) of the training situation and allows development of skilled performance in the operational task situation.

Although feedback is critical to the success of practice, it is not sufficient to say that feedback enhances learning. Merely informing a student that he was wrong is not as effective as telling him why he was wrong and how he can avoid making similar mistakes in the future. In fact, merely informing a student of an incorrect response may only frustrate him. Moreover, the type, amount, and timing of feedback should be tailored to the task being learned and to the achieved or momentary performance level of the student.

Guidance: Guidance involves directing the activities of the learner. Typically, guidance during practice involves the prompting of a correct response, i.e., "guiding" performance along lines that minimize errors. Guidance may entail directing the learner's attention to stimuli that are relevant to the task being learned, explaining how to interpret the informational content of stimuli, or explaining and demonstrating how to respond. Guidance also is involved when a learner is told what not to do. It is through guidance and prompting that we ensure that students practice the things that lead to desired learning.

Overlearning: Another aspect of practice that has implications for both initial skill acquisition and subsequent transfer is overlearning. Overlearning involves continued practice beyond the attainment of established criteria or standards of proficiency. Because the criteria for success during training often are set arbitrarily, the intent of overlearning is to ensure thorough learning of the task. Overlearning may be especially important when the task is not likely to be practiced often in the operational setting or if a significant time interval separates final practice of the task in the training environment from initial performance of the task in the operational environment. Overlearning also may be necessary to maintain achieved levels of performance during periods of emergency and stress, such as may be encountered during combat.

Motivation. Motivation involves behavior that is active, purposive, and goal-directed. Clearly the motivational level of a student affects his performance. The motivated student is an active participant in the learning process and typically works harder as his motivational level increases.

Although the initial aircrew learner usually approaches his training program with positive motivation, the manner in which that training is conducted can have a major influence on the student's motivation and, consequently, on his performance. Properly applied feedback, for example, can increase the motivation of a student. On the other hand, inappropriate or nonprescriptive feedback may be very frustrating to the student and decrease his motivation.

It is important to point out that the motivational level of a student may be as strongly influenced by the conditions of training as

it is by the student's desire to finish training successfully. Motivation not only affects the performance of the student within the ATD training environment, but also may influence his later performance in the transfer setting.

TRANSITION AIRCREW LEARNING

The transitioning aircrewman has already mastered the basics of his job, i.e., the pilot can fly, and the navigator knows how to navigate. The transition learning task involves learning how to perform these jobs in new aircraft and with new subsystems. Discrimination, generalization, and workload management learning are still the basic learning tasks. New cues must be recognized and associated with old responses. Sometimes different responses are required even though the new cues are largely the same as the old. This is particularly the case in some fighter aircraft where cue responses which are perfectly proper and safe in one aircraft can have disastrous consequences in another. As a result, the transition learner may have to "unlearn" some of his previous cue/response relationships in addition to learning the new ones.

It should be noted that the transitioning aircrewman faces many of the same learning tasks as does the beginner. He must learn new cue/response discriminations and generalizations. He may have more systems and systems of greater complexity to manage. As a consequence, most, if not all, of the factors previously identified as influencing the beginner's learning behaviors also apply to the transition learner. The transitioning aircrewman needs practice, feedback, guidance, etc. His motivation to learn is essential, and his need for overlearning is comparable.

From the perspective of ATD OT&E, the learning behaviors of the initial learner and of the transitioning learner are much the same. However, in the case of a highly experienced combat-ready aircrewman who is transitioning into a new piece of operational equipment, for example, the F-4 pilot "ace" learning to fly the F-15, somewhat different practice feedback and guidance are required. In such cases, the learning behaviors will more closely resemble that of the "continuation" learner.

CONTINUATION AIRCREW LEARNING

Continuation training refers to the periodic practice provided the mission capable airman. The highly skilled combat-ready airman continues to exercise both discrimination and generalization learning as he refines his job skills. As previously pointed out, he has learned which stimulus patterns provide the most efficient and effective cues. Not only can he discriminate fine-grain differences that

the novice trainee cannot, he even may use very different stimulus patterns from those the novice employs. He also has developed finely tuned response repertoires. By and large, he does whatever he has to do to optimize his weapons system's performance, and he does it at the proper times. He excels at work load management and copes exceedingly well with stress induced by job, task, and environmental factors.

His major "learning" task is thus one of doing whatever is required for maintenance of that achieved high skill level. The learner behaviors involved in skill maintenance are markedly different from the learning acquisition behaviors of the novice aircrewman. The novice, for example, needs frequent practice on those job tasks involving motor manipulative skills, or they are quickly extinguished. The highly skilled aircrewman, however, may need relatively less frequent practice on motor skills but may still require regular practice on procedures and systems management activities. Contrary to the novice who may need frequent guidance from an instructor so that he may correct errors, the experienced operator usually has his own internalized performance evaluation models against which to judge his needs for further practice.

IMPLICATIONS FOR ATD EVALUATION

The preceding discussion of the three basic categories of aircrew trainee learning tasks and of the factors that can influence such learning was intended to "set the stage" for a subsequent understanding of how ATDs function in support of that aircrew training. The discussion of ATD effectiveness considerations which follows in Chapter 3 will, therefore, relate to this present conceptualization of aircrew training. The discussion in Chapter 3 will first develop an operational definition of ATD effectiveness and identify some of the more significant factors which can operate to influence--positively or negatively--the ultimate effectiveness of ATDs. The relevance to ATD effectiveness evaluation procedures to the distinctions made in this chapter between initial skill acquisition learning and later skill maintenance behavior will be discussed. The effectiveness of an ATD used in support of initial training can differ markedly from its effectiveness when used to provide skill maintenance training. As a consequence, the test director must be sensitive to those different roles when planning and/or conducting ATD evaluations. Ultimately, these factors relate to the basic training objectives and goals to which the ATD procurement is a response and, consequently, to the test and evaluation objectives which the ATD OT&E seeks to address.

CHAPTER 3

ATD EFFECTIVENESS: DEFINITION AND EVALUATION METHODS

INTRODUCTION

ATDs work for very simple reasons. They can provide the cues for, and allow the responses appropriate to, the learning of an almost limitless number of aircrew tasks. In so doing, they can provide the novice or the transitioning learner an opportunity to practice the often difficult cue/response discrimination/generalization learning tasks described in Chapter 2 in an environment which is relatively free of stress. They also allow for the manipulation and sequencing of ATD learning task activity to optimize efficiency and effectiveness. Subsequent transfer of that learning from the ATD to the aircraft, therefore, can markedly enhance initial aircrew training efficiency and effectiveness.

ATDs also provide the combat ready aircrewman with a convenient and economical means of rehearsing his skills. They are particularly valuable to the highly proficient aircrewman in the rehearsal of complex procedures and in the practice of contingency behaviors. Many of these contingencies, such as in-flight engine failure, cannot be practiced in the airplane itself without risking lives and equipment. The ATD, therefore, becomes the only means whereby the aircrewman can achieve realistic practice in coping with such events.

Role of the Instructor

ATDs also work because they provide the instructor with greater flexibility in his management of the learning process than he has using operational equipment for training. They have a distinct advantage over aircraft in this regard. For example, aircraft must be carefully controlled during flight for safety and other reasons, and many flight behaviors must be cut off because of their possible adverse consequences. In contrast, the instructor can utilize the ATD in whatever ways he might wish to optimize the learning process. Thus, the instructor can allow flight behavior sequences to proceed to their full consequences if he wishes.

Using ATDs, the instructor may proceed directly to higher level tasks if he so desires. Furthermore, the instructor can provide substantially more practice or feedback via features such as "FREEZE" and "RESET." With an ATD, moreover, the sequence of individual skills can be taught according to instructional efficiency rather than being driven by safety requirements or sequencing limitations imposed by the aircraft during flight. Thus, discriminations underlying skills can

be taught in ATDs at times, and in ways, that promote more efficient development of these skills than is possible in the aircraft.

Importance of Training Program Design

Exploiting the training potential of ATDs requires operational training programs which incorporate sound principles and practices of instructional technology. Increasingly, aircrew training programs are incorporating these principles through Air Force-wide application of Instructional Systems Development (ISD) technology. Operational evaluation of ATD training effectiveness must always be considered within the context of the ATD-supported training program developed through the ISD process. A given device may be effective in one program, but be of little value in another simply because of the way that the training is conducted. That is, training effectiveness and transfer effects do not exist in the abstract or general sense. Instead, they are specific to tasks, situations, programs, and methods of ATD utilization; and, as suggested in the preceding chapter, they are particularly sensitive to the type of learning tasks involved.

The implications of those facts for ATD OT&E are not trivial. As will be pointed out in greater detail in those subsequent chapters dealing with specific OT&E evaluation methods and results reporting, the OT&E of an ATD must be planned within the context of its intended operational training program utilization, and the results interpreted within that same context.

UNDERSTANDING ATD EFFECTIVENESS

Chapter 2 provided a general understanding of the aircrew trainee learning process and pointed out some of the syllabus, training environment, and management practices characteristics which can affect the learning process. With this background, ATD training effectiveness and efficiency can now be addressed more specifically from the viewpoint of definition and measurement. The following discussion first considers the question of what constitutes ATD effectiveness and efficiency, and then proceeds to review in greater depth several of the various means by which ATD effectiveness and efficiency can be determined.

An ATD has no intrinsic training value. Whatever its design, from the simplest part-task device to the most complex weapons system trainer, it achieves effectiveness only when used. Therefore, an ATD's realized effectiveness will be a function of its design characteristics, the training environment in which it is employed, and the manner in which it is utilized. An ATD's designed-in capability may establish an upper limit regarding what it can potentially contribute to effective training, but the manner of its use determines the extent

to which that potential for training will be realized. Thus, an evaluation of a device's training effectiveness should consider the circumstance of its use as well as its designed-in potential.

The training effectiveness of an ATD is also directly related to the training objectives that have been specified and the criteria which have been established for meeting those objectives. Effectiveness is an index of the extent to which ATD training achieves, or supports the achievement of, defined training objectives. A given ATD and its associated training program may be highly effective with respect to one set of objectives and totally ineffective with regard to another set. Effectiveness, thus, is a function of the intrinsic capability of the device, the manner of its employment, and the training objectives which have been established (including their associated criteria).

ATD Effectiveness for Skill Acquisition/Maintenance

Chapter 2 stressed the fundamental differences between the learning tasks of the initial and transitioning trainee as opposed to the skill maintenance behaviors of the mission-ready aircrewman. The effectiveness of ATDs--and the operational test and evaluation of that effectiveness--will not necessarily be the same for these two categories of learning behaviors which are defined as follows: (1) Skill Acquisition, the learning behavior characteristic of the novice and transitioning aircrewman; and (2) Skill Maintenance, the learning behavior of the accomplished aircrewman. Major differences are likely to exist because the manner of ATD employment and the target training objectives are usually different for these two categories.

Effectiveness for skill acquisition. The effectiveness of ATD training during skill acquisition training has been commonly expressed in terms of training transfer. That is, ATD training for skill acquisition is considered to have been effective if that training facilitated (transferred positively to) subsequent trainee performance in an aircraft. Transfer effects may also be negative because training in an ATD may actually interfere with subsequent performance in the real world. Thus transfer of training (TOT) effects may be either positive or negative, or they may be zero, i.e., no effect at all.

As will be seen in Chapter 6 of this volume, the actual measurement of TOT effects is a sophisticated procedure in which ATD effectiveness is reflected by measures of task skills as performed in the aircraft or as reductions in the training time later required in the aircraft to reach specified skill levels following device training.

Effectiveness for skill maintenance. Expressing ATD effectiveness in terms of skills maintenance may also be considered as a type of transfer of training evaluation, since it presumes that use of the

ATD will maintain a given skill at a higher level than would occur were the device not used at all. The subsequent "transfer" of those ATD-maintained skills to the operational aircraft represents an extremely high-value outcome of ATD training. This type of transfer effect is of principal interest to the MAJCOM users of ATDS who are faced with the maintenance of high levels of combat/mission ready skills. Their evaluation concern is over the extent to which ATD training interspersed between operational aircraft training flights sustains mission-ready skills.

ATD effectiveness in skills maintenance is usually reflected in terms of aircrew performance measured before and after the intervening ATD training. If, after extended practice in a given ATD, an aircrewman performs as well, or nearly so, as he did during his preceding aircraft flights, then it is presumed that his "training behaviors" in the ATD transferred positively, and the ATD has proved to be effective. As with traditional TOT, actual measurement of such effects is a complex process, and one which depends heavily on good criterion measures of operator performance. (In fact, to date, few such objective evaluations of the use of ATDs for skills maintenance have been accomplished.)

It should be noted also that there are skills for which ATD training appears to be effective, but which (for safety or other reasons) cannot be performed in the operational aircraft. An example of this type of situation is the typical use of ATDs to teach certain in-flight emergency procedures. In such circumstances, the training effectiveness of an ATD during a test would be measured solely in terms of improvement in performance in the device following practice.

DEFINITIONS OF ATD EFFECTIVENESS

Consistent with the above discussion, ATD training effectiveness has been defined formally within the Air Force [1]. That definition states:

> ATD effectiveness is the satisfaction of some portion of overall aircrew training requirements. These requirements are skills needed for mission accomplishment and are expressed in terms of performance, skills surrounding the performance, and appropriate standards (levels of performance).

This definition is relevant whether the training program involves initial skill acquisition or skill maintenance because the emphasis is

on the satisfaction of training requirements and not on the specific content of those requirements.

Thus, a device with its associated curriculum (instructional program) is generally considered to be effective if it can produce and/or can maintain specified portions of overall Air Force aircrew skill requirements. The intent of this Air Force definition is to assure that ATD effectiveness is expressed in terms of specific behaviors, the conditions under which the behaviors are to be manifested, and those criterion levels which define competent performance. The effectiveness index of a particular ATD thus becomes an expression of the degree to which training activity on that device achieves a given training objective or of the proportion of multiple objectives which can be satisfied. The following examples are provided to illustrate ways in which this definition might be interpreted and applied:

1. Training on Device A reduces the effort required to learn to perform a steep turn in the airplane, with altitude deviation less than $\pm$ 50 feet, from ten trials to five trials; 50% fewer trials are required so the device may be viewed as 50% effective for initial training on that maneuver to the criterion stated.

2. If that same device is used to practice an already learned-to-criterion-level steep turn such that first trial performance in the airplane (after an extended lapse of time) is within criterion, and if such performance would not be within criterion without the device practice, the ATD may be considered 100% effective for maintenance of that skill.

3. If the same device is used to train a family or set of skills such as basic instruments, including the four maneuvers, instrument takeoff, climbs, unusual position recoveries, and letdowns, but is found to allow development of criterion performance for only three of these four basic tasks, it may be considered 75% effective for that particular set of tasks.

Alternative Definitions

There are other definitions of training effectiveness in the literature, but they are not substantively different one from the other or from the Air Force definition provided above. Each stresses (as does the Air Force definition cited) the importance of having specified objectives and established criteria to provide a basis for determining whether the training objectives have been met. Each of these definitions also emphasizes the importance to the meaning of measured training effectiveness of knowing the situation or environment within which effectiveness has been determined. Jeantheau [2] says, for

example, that "'Training effectiveness' simply refers to the outcomes of training: If the training is effective, the performance of the trainee meets the objectives that were sought, at the desired levels." Thus, the test and evaluation of the training effectiveness of an ATD can have no clear meaning apart from the specification of that which was tested (i.e., the tasks, skills, maneuvers) and the manner in which it was measured.

ATD Effectiveness vs. Efficiency

Obviously, ATD training must be effective to be of any value. It should be equally obvious that ATD training, albeit effective, must also be in some measure efficient relative to the utilization of training time and resources, and in terms of training costs. It is generally the case that ATD acquisition and operating costs are much less than those of the operational equipment. For example, Orlansky and String [3] report the median ratio of ATD-to-aircraft operating costs to be 0.12 for some 33 different aircraft/simulator training systems. This means that about eight ATD hours can be made available for the cost of one aircraft hour.

It is important that the test director also understands that ATD training effectiveness, particularly for initial skill acquisition, tends to diminish as a function of prior practice [4]. This means that the later hours of practice in an ATD may be less efficient in terms of added learning (i.e., as shown by transfer performance) than were the first few hours of device training. ATD efficiency is also influenced by a number of other variables. Some of these variables, such as ATD design and training syllabi, have already been identified. It is important that the test director have an appreciation of the nature of the effects that these and other variables can have on ATD training efficiency.

FACTORS INFLUENCING ATD TRAINING EFFECTIVENESS

The list of variables that can influence the effectiveness of ATD training is long. It includes, for example, such things as the device's design characteristics; the structure of the training program; the attitudes and entry skill levels of students; institutional bias; device fidelity; and a host of other variables. As will be seen, some of these variables have greater impact on ATD effectiveness under certain conditions than they do under others.

Design Characteristics

Training devices are designed and built to perform very specific training functions. Obviously, these designed-in capabilities present

some upper limit of the device's training potential. Unless the device provides the cues which are essential to eliciting the appropriate task-specific responses, the device ordinarily cannot be used to train that task. The cues provided (and their underlying stimuli) do not necessarily have to be identical with those available in the aircraft itself, but they must have functional equivalence.

Instructional Environment

It has become generally recognized that the manner in which a training device is used may be of equal or greater importance in determining its ultimate effectiveness than are many of the specific design characteristics of the device itself. Unless appropriate attention has been paid to the instructional environment within which the device is being used, its full training potential may not be realized. There are two principal aspects of the instructional environment which must be addressed carefully when examining the effectiveness of any given ATD: (1) the syllabus; and (2) instructors' qualifications and participation.

Syllabus effects. The principal effect of the syllabus is to define the basic instructional content of the overall training program which the ATD supports, and of the ATD training program itself. The syllabus also defines the order in which training objectives will be met, the manner in which they will be taught, and the amount of training time or resources devoted to each. Of particular pertinence to ATD OT&E is the fact that the syllabus is the principal means for controlling the manner of device use within the training environment and the relationship of that use to the larger training system. Thus, the syllabus exerts a major influence on both device effectiveness and efficiency.

The syllabus may require that all device training be completed before the trainee proceeds to the aircraft, or it may provide for a sequencing of device training interspersed with periods of training on the actual hardware. There are numerous examples of the successful utilization of ATDs employing either of these strategies. For example, airlines trainees typically complete all simulator training prior to going to the airplane. Doing so is consistent with the goal of ultimately accomplishing all upgrade and requalification training in simulators. The military, on the other hand, customarily intersperses or "blocks" ATD and aircraft training within training stages. For example, in Navy Jet UPT, all basic instrument training in the ATD precedes inflight basic instrument practice, but students then return to the ATD for an additional block of instrument training on airways navigation before that stage of training is given in the airplane [5].

Although there do not appear to be any hard and fast rules governing the syllabus sequencing of ATD and aircraft training sessions, it appears preferable to provide a block of ATD training in an amount sufficient for meaningful learning to occur in the device before transition to the airplane. As a general rule, that which can be learned in the ATD should be learned there to criterion proficiency before going to the aircraft. This philosophy has been applied in several research evaluations of complex simulators with apparent success [6,7].

Instructor effects. The role played by the instructor and his immediate managers is crucial to ATD utilization. Instructor attitude effects have already been mentioned, but there are other instructor effects of concern. For example, without special training covering the desired procedures for effective ATD utilization and indoctrination regarding the intended outcome of the planned ATD training evaluation, the instructors and their managers may employ the same instructional strategies and procedures in the device that they do in the airplane. The effects of such an instructional approach can be especially detrimental when device features that have no aircraft counterpart are involved. For example, the instructor coming to an ATD for the first time may know little or nothing about the effective instructional use of ATD features such as FREEZE and RESET which allow for real-time diagnosis of performance.

Instructors also may not fully exploit the multiple trial opportunities the ATD presents, e.g., approaches, etc., because they cannot give similar training in a single inflight training period. They may even express concern that doing anything in the simulator other than that which could be done in the airplane gives the simulator an "unfair" advantage, even though the possession of such advantages is one of the principal reasons that simulators are procured. In any event, the more effective ATD training programs are generally those in which instructors have been taught specifically how to use the ATD instructionally and how to exploit its instructional support features. This fact must be taken into account in any ATD OT&E effort.

Trainee Characteristics

As noted, trainee characteristics can be a factor in determining the realized training effectiveness of a given device. There is an interaction between trainee characteristics and the specific tasks to be trained. For these reasons, any effective training program and/or its evaluation must recognize the importance of the learner population characteristics and the relationships between these characteristics and the particular tasks being trained.

There also is a need to be aware of the possible relationships among such factors and the instructional strategies being employed. For example, ATD instructional strategies appropriate for novice

pilots will probably be totally inappropriate for the use of the ATD in maintaining the high skill levels of combat ready aircrewmen.

Attitudes Effects

People have attitudes about many things. ATDs are no exception. The prevailing attitudes about ATDs held by aircrew trainees and their instructors are based on a number of factors, ranging from their experience with modern ATD technology, or lack thereof, to concerns that excessive pressures to use ATDs in lieu of aircraft for training cost reductions may degrade training effectiveness. Unduly positive or negative attitudes can be a problem during ATD OT&E, since either attitude may have a tendency to bias the results of the evaluation. The test director, therefore, should be familiar with the nature of such attitude effects and with their implications with reference to his choice of an evaluation approach. A section of Chapter 4, "Factors Influencing Selection of an Evaluation Approach," specifically addresses the management of attitude effects during ATD OT&E.

APPROACHES TO ATD OPERATIONAL EFFECTIVENESS EVALUATION

A competent evaluation of ATD operational effectiveness will depend on the collection of valid data. Such data can be obtained in two general ways. First, opinion and/or judgment data regarding the device's training potential or perceived effectiveness can be collected from training experts or experienced crew members. The family of methodologies employed to collect data of this type comprise what is referred to here as an analytical approach. In a second way, performance data which are independent of expert opinion can be collected while the device is actually being us d for training purposes. Data based on trainee performance in the applied training environment are obtained from a family of methodologies referred to here as the demonstration approach. The analytical approach, therefore, is opinion/judgment based while the more objective demonstration approach is based on performance data.

Data collected from both analytical and demonstration approaches can be used during the evaluation of an ATD. There are different points in the development of the device, however, when the collection of each type of data becomes feasible and appropriate. Analytical data such as that provided by rating scales can provide useful estimates of ATD training effectiveness during in-plant evaluations and other early phase points, while demonstration data such as that provided by TOT methodologies can provide verification and validation of these earlier estimates through the demonstration of training effectiveness in the operational environment.

The literature contains many examples of both analytic and demonstration study designs, each of which has been used many times in evaluating the effectiveness of training programs and devices. A comprehensive survey of this literature by Caro [8] identified a number of analytic and demonstration evaluation approaches (models) which have been used with varying degrees of success in Air Force ATD training effectiveness evaluations. Brief descriptions of a number of these evaluation models are described below for the purpose of acquainting the test director with what has been used in the past. The appropriateness of these evaluation approaches for ATD OT&E will be addressed later in this volume, while specific guidance on how to conduct rating scale and TOT evaluations will be discussed in Chapters 5 and 6, respectively.

Analytic Approach Models

When data based on actual trainee performance are not readily available, other types of data which reflect ATD training effectiveness must be obtained. Several analytic models have been employed under such circumstances to generate evaluative data related to the effectiveness of the ATD itself, to the manner of its use, or both. Two such models will be discussed: the ATD fidelity model and the opinion survey model.

The ATD fidelity model. The ATD fidelity model yields data which describe the device in terms of how close a physical correspondence there is between it and the operational vehicle. Use of this model is based upon the assumption that an ATD which is very similar in appearance and handling to operational vehicle (high fidelity) will aid in the achievement of higher transfer of training than will an ATD which is not as similar (low fidelity) to the operational vehicle [9].

The ATD fidelity model is most often used when it is not practical to collect performance data and when other types of data are not available. While the model has wide appeal among operational personnel, there are limitations to use of the fidelity model as an unequivocal indicator of device effectiveness. Data describing device fidelity can be used as a partial basis for predicting training effectiveness, but their use for determining actual device effectiveness is inappropriate. Bryan and Regan [10], for example, have noted that a simulator can be a very faithful copy of operational equipment and be either effective or ineffective with respect to a particular training requirement. Likewise an ATD may be relatively low in fidelity, yet be effective. In fact, some well designed training equipments deviate intentionally from the operational device in order to enhance learning. Fidelity, per se, therefore is not an unquestioned indicator of effectiveness.

The ATD fidelity model has the limitations described above because it almost always ignores the manner in which a device will be used and the objectives of device training, two considerations which must underlie any operational determination of ATD training effectiveness. Therefore, its use should be restricted to those situations where other types of data cannot be obtained or where it has been determined that data regarding similarities between the ATD and the operational vehicle will be useful in and of themselves. Despite this caution, it should be noted that this method is often the only feasible evaluation alternative available to the test director, especially during an in-plant IOT&E.

The opinion survey model. There are instances in which attempts have been made to determine the effectiveness of ATD training solely through use of opinion data with no use of operational training or performance testing data. While this method is not generally recommended, it is sometimes used. For example, it has been used on some occasions when it was necessary to make procurement decisions based on the predicted training effectiveness of a newly developed, but untested, ATD, or between ATDs under development. In such instances, analysts may be forced to evaluate the probable effectiveness of an ATD by asking operators, instructors, training specialists, and even students, for their opinions concerning the perceived training value of the device or certain of its features, or the probable impact upon subsequent operational performance of training in the various devices. Unfortunately, such data may lead to erroneous conclusions because such opinions are often expressed without regard to how the device is used or what the objectives of device training are. Meister, Sullivan, Thompson, and Finley [11] have shown that estimates of ATD training effectiveness based upon instructor opinions varied widely among the different instructors expressing such opinions. The opinion survey model is, therefore, too unreliable for OT&E applications.

Demonstration Approach Models

Use of the demonstration approach usually involves some form of the basic Transfer of Training (TOT) model. TOT methods are generally considered to be the most appropriate means for demonstrating whether ATD training will improve the trainee's subsequent operational aircraft performance, because they embody the basic concept underlying the operational use of training devices, i.e., transfer of training itself (see discussion pages 25 and 26).

In its simplest form, the TOT model requires two groups of trainees: a demonstration group that receives device training prior to further training or performance testing in the aircraft; and a control group that receives training only in the aircraft. More complex transfer models may involve more than one ATD demonstration group in order to evaluate the differential advantages of alternative

ATD utilization scenarios, e.g., different amounts of ATD training, motion vs. no motion, etc.

These two groups must be comparable, of course, in terms of relevant prior training and experience. Care must always be taken to ensure that the control "treatment" itself does not influence that group's subsequent performance in the criterion situation. Such an influence could be facilitative, e.g., a period of rest for the control group while the demonstration group engages in fatiguing or stressful training; or debilitating, e.g., a period of fatiguing or stressful activity such as operational missions or extended duty required only of the control group because of their availability for additional assignments. Particular care should be taken to ensure that members of both groups are prevented from engaging in flying or related operational activities likely to influence their performance on criterion tasks and thus invalidate demonstration and control group comparisons.

This basic transfer design permits demonstration and control group differences in performance in the aircraft to be attributed to the influence of ATD training received by the demonstration group(s). The transfer design is particularly advantageous, in that it is sensitive to both positive and negative transfer effects. Several variations of the transfer model of interest in the present context are described in following paragraphs. More specific guidance on how to conduct TOT evaluations is presented in Chapter 6.

The self-control transfer model. This variation of the transfer model is of possible interest for a situation in which a device might be employed at an intermediate stage of training, i.e., when operational training is interrupted for a period of training in the device. In such a situation, the students in the demonstration group could serve as their own controls, and their performance data obtained in the operational aircraft immediately following simulator training could be compared to similar inflight data obtained on them immediately prior to their simulator training. The difference in these two sets of inflight performance data, then, could be attributed to the intervening simulator training program.

The pre-existing control transfer model. There are instances in which a concurrently trained control group may not be necessary. For example, when ATD training is added to an existing training program, or when a new ATD-supported training program replaces an old one, student performance data from the existing or older program can be compared with similar data from the new program to determine the latter's effectiveness. For such a comparison to be valid, the pre-existing data must have been gathered under conditions which would have been applicable to a control group trained concurrently with the experimental group. A disadvantage of the pre-existing control transfer model

is that differences in performance between the two groups may be the result of changes which have occurred in the trainee population during the time between the selection of demonstration and control students.

When a control group cannot be employed and suitable control data do not exist, simulator training effectiveness can be hypothesized if students can perform a particular task in the operational vehicle following its learning in the simulator without an opportunity to learn that task in the operational vehicle. Data gathered in this manner can be suspect, since improvements in performance may not be solely due to ATD training. Nevertheless, such data can carry considerable weight, particularly when a task critical to flight safety is involved and a plausible case can be made that the underlying skills probably are attributable, at least in part, to the device-supported training programs.

The ATD-to-ATD transfer model. Many studies of the effectiveness of ATDs involve transfer of training from one device to another rather than transfer to operational equipment. For example, if instrument skills learned to proficiency in Device A can be shown to facilitate instrument task performance in Device B, some measure of training effectiveness for those skills can be inferred for Device A. Should it be known that Device B produces positive transfer of instrument skills to the aircraft, it would seem likely that Device A might also produce instrument skills that would transfer to the aircraft. Of course, this is an assumption, which should be verified, if possible, using other TOT methods.

There is one situation, however, in which the device-to-device transfer model is clearly appropriate. This situation exists when Device B is actually the criterion vehicle. For example, the effectiveness of training in a part-task training device can be determined by measurement of subsequent performance in a full-mission simulator if the objective of such part-task training is to reduce the use of the more complex device. In this situation, it would be presumed that performance in the simulator would involve intermediate training objectives, with the final objectives relating to subsequent performance in an operational vehicle.

The backward transfer model. Another simulator transfer evaluation design is known as the backward or inverse transfer of training model. In a backward transfer study, an operator who already has demonstrated mastery of relevant training objectives in the operational vehicle is "transferred" to the simulator, where he is required to perform tasks corresponding to those he had mastered operationally. If he can perform such tasks to criterion levels without some amount of practice in the simulator, backward transfer is said to have occurred. This fact is taken as evidence that transfer in the simulator-to-vehicle sequence, although of unknown quantity, likely will be positive.

The backward transfer design should be used with caution to evaluate ATD effectiveness for use with novice and/or transition trainees for at least three reasons: (1) positive results assume that a suitable training program has been developed for use of the simulator; (2) experienced personnel already proficient at operational tasks are likely to have experience and skills not possessed by recent training program graduates and may, therefore, be able to transfer to the device because of these more general skills rather than because they possess the skills needed to operate a particular vehicle or perform a particular mission; and (3) the simulator may be suitably designed for the eliciting of a particular set of behaviors by skilled performers, but may lack the cues necessary to elicit these behaviors from beginners.

The backward transfer model may prove to be a useful tool for evaluating the effectiveness of an ATD for maintaining the skills of mission-ready pilots. While backward transfer data should not be the sole justification for adopting a particular simulator for these purposes, such data will provide an important step in the recommendation process. However, note that while positive transfer evidence would be reassuring, negative results could be misleading. It is possible that some tasks are performed in the aircraft by experienced personnel in response to cues not present in the simulator. Therefore, these personnel might be unable to perform such tasks in the simulator without training in it. The same simulator might, however, provide other (or surrogate) cues which these trainees can learn to use to perform those same tasks in the simulator for subsequent transfer back to the aircraft. An example of how this model can be employed is reported by Adams and McAbee [12].

The uncontrolled transfer model. There are circumstances in which a separate control group cannot be employed, the self-control or the pre-existing control transfer models are inappropriate, and suitable control data do not exist. Such circumstances might be dictated by any number of considerations: political, administrative, or safety. For example, it might be unacceptable to "penalize" members of one group by requiring that they undergo a different and possibly inferior no-ATD training program. In some instances, a control group simply may not be feasible. The effectiveness of lunar landing simulators could not be determined, for example, by employing a no-simulator-training control group of astronauts. An example of how to apply the uncontrolled TOT model is documented by Thorpe and his colleagues [13].

The ATD performance improvement model. The ATD performance improvement model is considered an example of a demonstration model, but it is not a transfer model, per se. Instead, transfer to the aircraft is presumed to occur if improvement occurs in the performance of trainees in ATD as a result of training they receive in that device.

If such improvement does not occur, there would be little expectation that subsequent operational performance in the aircraft would be improved as a result of simulator training. Because of this dependency relationship, improvement in performance in the simulator often is cited as evidence that simulator training is effective. This typically is done when circumstances preclude the employment of a transfer model.

Clearly, there are circumstances in which the ATD Performance Improvement Model can provide the best available estimate of whether a simulator training program is effective. It must be noted, however, that this model yields only indirect evidence of simulator effectiveness. Performance improvement (learning) in the simulator is a necessary condition for transfer to the operational equipment to occur, but its existence does not prove conclusively that improved performance in the simulator will definitely result in improved operational performance. An example of how this methodology can be employed is found in Burger and Brictson [14].

REFERENCES

1. Msg. Hq. USAF, XOODD, Subject: Aircrew Training Device (ATD) Effectiveness, 8 May 1978.

2. Jeantheau. G. G. Handbook for training systems evaluation (Tech. Rep. NAVTRADEVCEN 66-C-0113-2). Orlando, FL: Naval Training Device Center, January 1971.

3. Orlansky, J., & String, J. Cost-effectiveness of flight simulators for military training. Volume I: Use and effectiveness of flight simulators (IDA Paper P-1275). Arlington, VA: Institute for Defense Analyses, August 1977.

4. Flexman, R. E., Roscoe, S. N., Williams, A. E., Jr., & Williges, B. H. Studies in pilot training: The anatomy of transfer. Aviation Research Monographs, 1972, 2(1).

5. Chief of Naval Air Training. Curriculum, advanced jet (TA-4J) (CNATRA Instruction 1542.20B). NAS Corpus Christi, TX: Author, September 1976.

6. Caro, P. W., Isley, R. N., & Jolley, O. B. Mission suitability testing of an aircraft simulator (HumRRO Tech. Rep. 75-12). Alexandria, VA: Human Resources Research Organization, June 1975.

7. Woodruff, R. R., Smith, J. F., Fuller, J. R., & Weyer, D. C. Full mission simulation in undergraduate pilot training: An exploratory study (AFHRL-TR-76-84). Brooks AFB, TX: Air Force Human Resources Laboratory, December 1976. (ADA039 267)

3. Caro, P. W. Some factors influencing Air Force simulator training effectiveness (HumRRO Tech. Rep. 77-2). Alexandria, VA: Human Resources Research Organization, March 1977.

9. Osgood, C. E. Method and theory in experimental psychology. New NY: Oxford University Press, 1953.

10. Bryan, G. L., & Regan, J. J. Training system design. In H. P. Van Cott & R. G. Kinkade (Eds.), Human-engineering guide to equipment design (Rev. ed.). Washington, DC: Government Printing Office, 1972.

11. Meister, D., Sullivan, D. J., Thompson, E. A., & Finley, D. L. Training effectiveness evaluation of Naval training devices. Part II: A study of Device 2F55A (S-2E trainer) effectiveness (Tech. Rep. NAVTRADEVCEN 69-C-0322-2). Orlando, FL: Naval Training Device Center, January 1971.

12. Adams, J. A., & McAbee, W. H. A program for a functional evaluation of the GAM-83 Melpar trainer (Rep. No. APGC-TN61-41k). Eglin AFB, FL: USAF Air Proving Grounds, October 1961.

13. Thorpe, J. A., Varney, N. C., McFadden, R. W., LeMaster, W. D., & Short, L. H. Training effectiveness of three types of visual systems for KC-135 flight simulators (AFHRL-TR-78-16). Brooks AFB, TX: Air Force Human Resources Laboratory, June 1978. (AD-A060 253)

14. Burger, W. J., & Brictson, C. A. A7E transfer of training effectiveness: Device 2C15A CPT and Device 2F84B OFT/WST (NAVTRAEQUIPCEN 74-C-0092-2). Orlando, FL: Naval Training Equipment Center, August 1976.

CHAPTER 4

EVALUATION APPROACH SELECTION AND IMPLEMENTATION FACTORS

INTRODUCTION

The various analytic and demonstration methods identified in the preceding chapter represent a range of potentially useful approaches for ATD OT&E efforts. From these the test planner must choose an evaluation model that will provide valid data relevant to his testing objectives, and one that he can implement in the operational testing environment. Although selection of the design most appropriate for the type of OT&E involved and its implementation in the OT&E testing environment of concern are more straightforward than it might at first appear, there are a number of factors which can markedly influence both the design selection and design implementation processes. It is the purpose of this Handbook chapter to identify the more critical of these factors and to provide guidance to the test director regarding their effects and management.

FACTORS INFLUENCING APPROACH SELECTION

There are many factors which the ATD test director must consider when deciding upon a particular demonstration methodology. The most critical of these factors are discussed in detail below. However, the test director should keep in mind that other factors not discussed below may also influence the final decision.

Evaluation Location

Choosing between an analytic or a demonstration approach is relatively easy when planning for an in-plant ATD IOT&E/QOT&E because the in-plant environment usually is not suitable for the conduct of controlled training activities involving students. As a result, only an analytic evaluation model is usually feasible for use in that environment. Selection of the most appropriate evaluation design is, however, somewhat more complex after the device has been installed at the user's training facility. While analytic evaluation models can be useful in that setting, use of an analytic method during on-site FOT&E is normally less desirable than is the use of a demonstration method in which actual trainee performance data are collected. As noted earlier, one of the TOT models is generally more appropriate for this situation.

Calenda Time Availability

Proper advance planning should assure that sufficient calendar time is available in which to implement the preferred design. It is, however, extremely important that the test director identify as early as possible during the planning phase the calendar times required for each evaluation design being considered. Should calendar time be critical (as is often the case with ATD acquisition and test), the test director may find it necessary to select a briefer, albeit less preferable, evaluation approach than would otherwise be the case.

Management's Commitment

The importance of firm resource and support commitments from the MAJCOMs and other organizational management activities involved in an ATD OT&E cannot be overstressed. There is no question that the strong general commitment of top management is critical to the conduct of competent ATD IOT&E/FOT&Es. Problems arise, however, when mid-level management and the test team become aware of precisely what will be required to achieve a worthwhile ATD training effectiveness evaluation. These problems are compounded when implementation of a particular ATD evaluation approach is perceived to represent a departure from established training practices or to require substantial extra effort at the working level. This is especially true if the impending evaluation is perceived to be potentially disruptive to the accomplishment of the unit's basic mission.

The resolution of such conflicts is obviously a management responsibility. The OT&E test director should be sensitive to such potential problems and should seek to resolve them in advance of the actual OT&E to the maximum extent possible. Unfortunately, such support problems often do not surface (or are not faced) until the OT&E is well underway. In such cases, the result is likely to be a desire to compromise the OT&E design in an attempt to "keep the test going." The tendency is to accommodate to most contingencies believing that "some test is better than none."

Although such an intent is understandable--even commendable--the test director must clearly understand from the outset the risks to test integrity that can result from going from a preplanned, desired level of support to a "bare bones" level. He must also be certain that the resource managers involved in the test are equally aware of these risks before the test starts. All concerned, the test director and the resource managers, have an equal responsibility to recognize the support costs of the various demonstration evaluation designs and be certain that those costs can be endured during the test as a condition of commitment to any one such evaluation design.

Final commitment should be viewed by all parties involved as the equivalent to a contract. For this reason, in those following chapters of the Handbook which address specific evaluation designs and techniques, there are clear criteria for application of a given design to, or exclusion from, each type of evaluation activity, whether it be in-plant, or at the user's facility. In addition, there are relatively simple statements of the minimum support required if the design is to be considered at all, and of the support desired for a good test. Such information should be of help to the test director in resisting external pressures to attempt IOT&E/FOT&E when the device is clearly not ready for test or, once the test has started, to continue testing should the specified support conditions be significantly compromised.

New vs. Old ATD Design Features

The design complexity or novelty of the ATD will influence which evaluation approach is ultimately selected. A cockpit familiarization and procedures trainer, for example, poses a different evaluation problem than does a weapon system trainer (WST) with visual system, especially if that visual system represents a "new" application of visual simulation technology. Not only are these two devices likely to be different in basic crew station design complexity, but they may be quite different in terms of instructor station capabilities as well.

The principal evaluation concerns regarding ATD design features are ATD complexity and the extent to which previously untried simulation technology has been included on the ATD. The test director should first compare the design features of the ATD to be tested with those of ATDs in operational use. If he finds that the device to be tested differs little from those in operational use--as would likely be the case with a cockpit procedures trainer or an instrument flight trainer--he would be reasonably safe in inferring that the two devices should have comparable training effectiveness capabilities. In such an instance, he would hardly be justified in planning for a time-consuming and probably costly transfer-of-training evaluation. Rather, he would be wise to choose a comparatively quick and economical analytic approach such as a rating scale evaluation.

Should the device be an operational instrument trainer, but with the recently added complexity of a new visual system (e.g., a high-resolution day visual), a demonstration approach utilizing one of the TOT methods might be appropriate. In this instance, the test director should be particularly wary of the possible degraded performance of the basic simulator resulting from improper integration of the visual add-on. Such concerns could warrant a "fidelity" evaluation as well as TOT. In addition, if the device is to be used in support of

mission-capable pilot skill maintenance, the test director might find a reverse transfer approach of value.

Finally, if the device represents a major advance in ATD technology applications for which no (or very little) "hard" effectiveness data are available, the test director would be well advised to plan as extensive an OT&E as resources would permit. A series of separate evaluations could be carried out. These might include fidelity evaluations in-plant, instructor rating assessments during IOT&E, and a full-blown FOT&E TOT evaluation.

Intended Uses of the Device

Care must be taken in the process of "finalizing" the OT&E evaluation approach to assure that the results will clearly apply to the trainee population of interest. For example, if the device is to be used for both transition and continuation training, it must be evaluated for both applications. A transfer of training evaluation using CCTS subjects may not reveal the ATD's real potential, or lack thereof, for support of mission-readiness maintenance. Not only must the results be gathered on appropriate subjects to be applicable, the type of evaluation that is appropriate to those subjects must be utilized. In this regard, some of the earlier discussions of initial, transition, and continuation training are pertinent.

Criterion Measurement Availability

Consideration of any of the demonstration evaluation models immediately raises the issue of the feasibility and/or availability of adequate criterion performance measures. It is important to note that adequate criterion measures must be available for measuring performance in both the ATD and in the operational equipment or aircraft.

Discriminating and detailed quantitative criteria of success are particularly critical to the conduct of TOT evaluations where measures must be sensitive to any meaningful improvements in performance that may be attributable to the use of the ATD. Also, if such measures are to be useful, they must be free from user bias, and relevant to both the training and testing objectives. Finally, any measures selected or developed must be feasible of implementation within the environment of ATD IOT&E/FOT&E.

The development of useful performance measures meeting the above listed requirements involves a technical expertise which may not be available to the typical ATD OT&E team. As a result, later sections of this Handbook provide further guidance concerning the criterion measures needed for TOT evaluations and identify the nature of the problems to be addressed. Also, distinctions are made between subjective and objective evaluation and measurement, the measurement options

available, and the effort required for measure development. From this information the ATD OT&E team can judge whether the efforts required exceed on-board capability and whether expert consultative support should be accessed.

Measurement selection and evaluation credibility. ATD transfer effectiveness evaluations are particularly sensitive to the kinds of measures employed. As a result, the confidence which may be placed in the results obtained with any particular evaluation design will be a function of the objectivity and reliability of the data that are obtained during the test. Many of the candidate demonstration evaluation designs described earlier require measurement capabilities that are sensitive to and/or indicate the extent to which the prescribed specific training objectives have been met. Thus, one of the major considerations in selecting any such evaluation model and achieving credible results is the measurement capability that will prevail--or that can be developed and implemented for the duration of the evaluation.

It should be noted that TOT evaluations conducted in operational training contexts have proved to be markedly vulnerable to problems of acquiring valid performance data. One of the principal limitations often noted is the difficulty in developing appropriate, objective measures of performance, for both the device and the operational vehicle. In fact, relatively few transfer studies have been reported in which comments do not appear in the "Discussion and Conclusions" section of the report concerning the inadequacies of the performance measures used.

What to measure. There are two fundamental questions pertinent to the measurement of trainee performance which must be considered in planning an ATD training transfer effectiveness evaluation. The first deals with selecting aspects of performance to be measured.

Numerous measures of trainee performance can be used as dependent variables in an ATD training effectiveness study, as long as they are objective, reliable, and relevant to the objectives of the training being conducted. Since most current ATD programs have been developed through the ISD process, the procedures established by this process should provide a ready source of information about what to measure and how to measure it. This range of potentially useful measures includes, for example, mission success indices; error scores; time measures reflecting time to criteria, time on target, duration of exposure to enemy surveillance, etc.; indices of communications content, frequency, and duration; and measures of training effort and efficiency, such as the number of trials to reach criterion performance in the aircraft, the amount of training time saved, or the savings in overall training costs resulting from the ATD training program.

The basic requirement for selecting any performance parameter to measure is its relevance to both the objectives of simulator training and to the objectives of the evaluation. For example, an objective of both training and OT&E might involve the landing of the aircraft. A simple pass/fail measure may be acceptable for training management purposes, i.e., the landing met pre-established criteria or it failed to do so. However, more specific and detailed objectives and measures of performance would be preferable in OT&E. For example, separate measures might be desired which indicate whether device-trained students tended to be more or less accurate in their touchdown point, whether long or short and how much, etc. Such indices would be of particular concern in ATD applications such as Navy carrier landing training.

The specificity of the parameters measured should be determined by the specificity of the evaluation objectives. Since many present-day aircrew training programs are based upon quite specific, systematically derived criterion-referenced behavioral objectives, it is to be expected that a well-conducted ATD training effectiveness study would yield multiple specific measures of trainee performance keyed to previously described training objectives. However, the OT&E planner must be concerned with the relevance of these training objectives to the decisions which may be made as a result of the evaluation. The training relevance of the performance measures used can usually be assured by keying the parameters selected to the approved training objectives, but determining the relevance of these measures to test objectives may be more difficult. In any event, the OT&E planner must guard against the error of selecting measurement content simply because there already exists a measure of that content that can be used.

An extremely important aspect of this concern over what to measure has to do with the handling and analysis of the resulting data. The OT&E test director must be certain of the team's capability to manage data volume and to ensure the appropriateness of the statistical tests used to evaluate that data. The volume of data and the analyses employed are likely to require electronic data processing support. If so, the test director must ensure that such support will be available.

How to measure. Determination of how to go about measurement is obviously driven by consideration of what is to be measured. It is also driven by factors such as feasibility, safety, and acceptability. It is also obvious that the objectivity of performance measurement is a matter of the methods and procedures to be used in acquiring and recording the performance data of concern.

The advantages of employing an automated measurement capability in terms of their objectivity are often cited, but despite such

advantages, automatic performance data recording devices cannot always be used because of feasibility, safety, or cost effectiveness criteria, or simply because they are not available. As an alternative, manual data recording techniques typically require the use of checklists or other forms upon which trained observers record operator performance on specified parameters while the performance takes place. If manual techniques are to be used, the test planner must be sure that such measures exist, or that he has the expertise required for their development available to him.

Support Resources Availability

The resources required to support many of the evaluation approaches--particularly the demonstration-type models--can be a crucial factor in the approach selection process. If he plans an analytic approach, the test director must be sure that there is adequate time available at the ATD for the evaluators to accomplish the necessary flight scenarios and rating procedures. If he anticipates using one of the demonstration models, he must also be assured of access to sufficient aircraft time for trainees to demonstrate transfer effects. Perhaps the most crucial concern for the test director planning to utilize a demonstration model is the availability of instructors. Often instructors are already in short supply and just cannot absorb the added workload. Students may be equally hard-pressed to participate. Student flow is often programmed to meet inflexible completion dates. As a result, the variations in their progression that may be required by an evaluation design may not be feasible to implement.

Subject Population

The population of subjects available is of critical importance. Not only must the subject pool be representative of the trainees who will be using the device (as previously noted), but there must be enough subjects available to support the evaluation design being entertained. There is general agreement among training device evaluators that an N of 20 or more subjects per evaluation group is appropriate. It is also generally accepted that an N of less than 10 is not sufficient for demonstration-type evaluations. (See Chapter 6 for more specific guidance.)

Appropriate Curriculum

The curriculum and syllabus determine the manner in which the device is to be used. Since it has become almost axiomatic that, "It's not the device per se, but how you use it that counts," it follows that any useful demonstration of an ATD's effectiveness should be based on realistic curriculum and syllabus. The test director should, therefore, include the using command ISD team in his early planning activities.

SELECTING AN APPROACH

The discussion in this chapter has surfaced a number of considerations that have an impact on the selection of an evaluation approach for a particular ATD OT&E. These highly critical factors have been addressed as they relate generally to the task of ATD evaluation approach selection and implementation. The discussion of these factors was intended to assure that the ATD OT&E test director recognizes their importance to a successful OT&E.

Being acquainted with the importance of these factors to ATD OT&E is not sufficient in itself. The test director also needs a procedure that will help him select that evaluation approach best suited to his particular situation. Such a procedure is provided by the algorithm diagrammed in Figure 4-1. Starting with "Evaluation Location," the test director proceeds systematically to consider each of these factors in turn. The final choice between a demonstration or an evaluation approach will always rest with the test director. His is the final judgment.

It must be emphasized that use of this selection algorithm does not necessarily assure the test director that either an analytic or a demonstration approach will work. In some situations, he should be prepared to recommend that no test be conducted until these areas of concern are under reasonable control. With such assurance, of course, the test director still is faced with the need to determine what particular evaluation strategy to pursue and what specific test methods to employ.

IMPLEMENTATION OF EVALUATION

Regardless of the evaluation approach finally decided upon, the test director must also be concerned that little or no change occurs as concerns these factors during the actual conduct of the evaluation. Experience has shown that the most likely factor to be of concern is user attitudes toward the particular ATD of interest and toward the test per se. For this reason, the remainder of this chapter will address user attitude effects--their source and management.

USER EFFECTS ON EVALUATION APPROACH

User attitudes toward aircrew training devices will influence how effectively they are utilized and how accurately they are evaluated. For example, negative attitudes toward ATDs may thwart effective instruction because less capable personnel may be assigned as ATD instructors or because of inadequate level of maintenance support being provided for the devices. Also, unfavorable attitudes held by

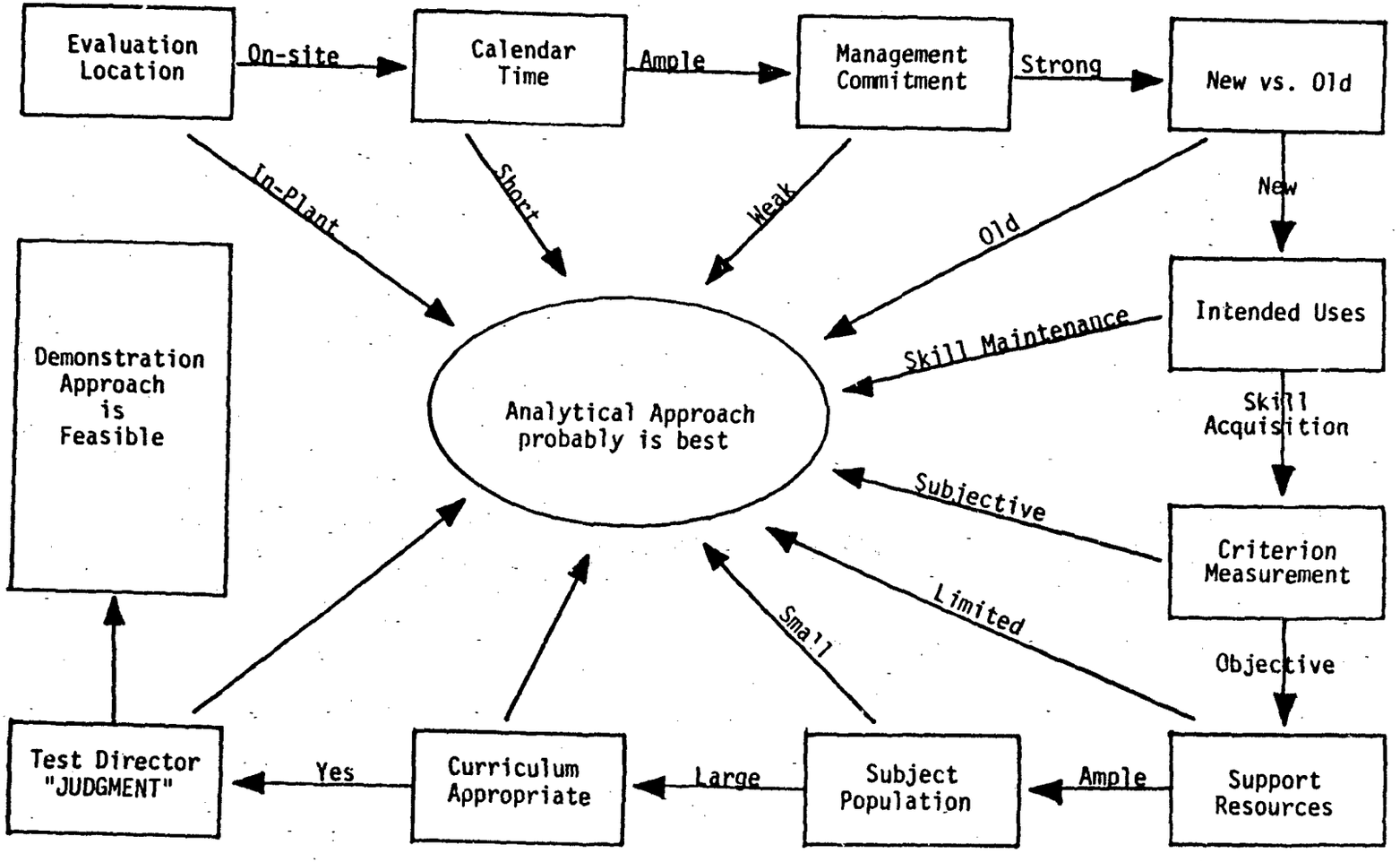


Figure 4-1. "Evaluation Approach" selection algorithm.

instructors and students can impair the effectiveness of ATD training by compromising the delivery and acceptance of that training.

Attitudes toward ATDs can be influenced by (1) the basic design of the ATD, (2) the manner of introduction of the device to the using organization, and (3) the manner in which ATD training is conducted and managed. The test director should be aware of these sources of user attitudes towards ATDs and their use and should strive where possible to promote positive attitudes towards the device to be evaluated. Indeed, one of his test activities should be to assess user attitudes towards the ATD to be evaluated as an initial step towards the conduct of a competent and valid evaluation.

ATD Design Factors

Three aspects of ATD design can have a major impact on user attitudes. These are: (1) the physical and functional (dynamic) correspondence of the device to the actual operational aircraft, i.e., the fidelity of the device; (2) the design and ease of operation of the instructor/operator station; and (3) the perceived extent of user participation in the process of designing the device.

ATD fidelity. Physical fidelity appears to be a major determinant of personnel attitudes toward ATDs and ATD training. If the ATD does not react in a fashion identical to the operational equipment, it may be considered nonacceptable regardless of whether it was intended to be a high fidelity ATD or not. The importance of physical fidelity is due primarily to the dominance of the aircraft, or actual crew station, as a model for ATD design and its use as a standard of comparison for ATD evaluation. Training devices that incorporate a high degree of physical and operating fidelity typically are regarded as more "impressive" than devices of lower physical fidelity. Similarly, high fidelity devices are viewed as providing the "realism" necessary to impose realistic training stress on the students, an important factor in the eyes of many aircrew members. Consequently, devices that do not replicate the actual aircraft may be seen as providing deficient training and, thus, may produce negative attitudes. It does not follow, though, that low fidelity devices, if properly designed and used, need necessarily produce negative attitudes.

Instructor/operator station design. The instructor/operator station (IOS) provides the interface between the ATD instructor and the device and, to a considerable extent, between the instructor and the student. The efficiency of the instructional process is heavily dependent upon how suitably designed these interfaces are because this design will have an impact on the nature of the attitudes instructors hold toward the ATD.

Common deficiencies in IOS design that have been identified as affecting instructor performance and attitudes include the following: inadequate human engineering of displays and controls (e.g., displays and controls that are difficult to see or reach, lack of labels); cumbersome input tasks for interacting with the computer (e.g., lengthy keyboard entries for commonly used functions or problem set-ups); cockpit repeater instruments arranged differently from those in the aircraft (thus presumably conflicting with highly developed scan patterns); instructional features that were inappropriately designed with respect to training requirements (e.g., record/playback feature controls that did not allow instructors to reconstruct needed portions of tactical intercepts); and poorly organized CRT displays.

In summary, a poorly designed IOS degrades the credibility of the ATD, increases instructor workload, and often frustrates instructors who must use the device. All of these factors may contribute to unfavorable attitudes toward the device and will have an impact on any evaluation process.

ATD design process. User attitudes toward new equipment are influenced by perceptions of the adequacy with which the design process took into consideration user needs. Users tend to feel more positive toward a device, or training program, when they are confident that the designers and the procurement agency have taken the viewpoints of the user into consideration. In contrast, negative feelings are more likely toward a device or training program that is designed in isolation from users and figuratively "dumped" on them. The perception that the using community has been excluded from the design and acquisition process may lead to a "not invented here" attitude that, in turn, hampers effective integration of the device into the training program and the effective evaluation of the device.

Introduction of the ATD into the Training Community

The procedures employed during the introduction of an ATD into training often have an influence on initial attitudes toward that training. These initial attitudes also may affect how the ATD is used during the remainder of its life cycle and how effectively it is evaluated. Three factors appear particularly important in this regard. They include the activities conducted in preparation for the introduction of the ATD, the roles of managerial and instructional personnel, and the manner in which the device is introduced into the training community.

Preparation for ATD introduction. It is a common practice that the introduction of a major ATD may be accompanied by a variety of changes in the conduct of both ATD and non-ATD training. For example, the introduction of a device into some units has been viewed as an appropriate time to adopt an "ISD'd" training program or to introduce

proficiency advancement techniques. Such changes themselves may foster unfavorable (or favorable) attitudes. Unfortunately, the concurrent introduction of an ATD into a training program may lead training personnel to blame (or praise) the ATD for unrelated changes in other portions of aircrew training. This, in turn, can generate undue attitudes towards ATD training which may have an adverse impact on an evaluation of the device.

Personnel roles during ATD introduction. Personal familiarization and experience with new equipment before it is formally used on the job tends to promote favorable user attitudes toward the equipment. Early personal exposure provides the opportunity to: (1) overcome negative attitudes based on unfamiliarity; (2) compare one's own performance using the new equipment with that using older equipment; and (3) exchange attitudes and ideas about the equipment with peers and colleagues.

Members of initial ATD instructor cadres who have participated in device development and testing activities (e.g., in-process reviews, factory acceptance tests, and OT&E) often attribute the formation of highly favorable attitudes toward the device to their initial "hands-on" experience and personal involvement with the development and testing of the ATD. Such highly positive or negative attitudes will affect evaluation outcomes and should be assessed prior to and monitored during the evaluation process in an attempt to neutralize their impact.

Management of ATD introduction. First impressions of the effectiveness of a device can affect its subsequent acceptance and use. Hence, it is important that the device function according to expectations before aircrew training activities with the device are allowed to begin. Negative attitudes have resulted in cases where ATDs were used for training before they were ready, where the training program was only partially prepared, or where the instructors were not fully trained. Such unfavorable attitudes often have persisted even after the initial problems were solved and will affect evaluation results if not identified and eliminated to the maximum extent practical.

Conduct and Management of ATD Training

The third, and perhaps most important, factor that affects attitudes toward ATDs is the manner in which ATD training is conducted and managed. Four aspects of the conduct and management of ATD training are of concern: (1) the attitudes of instructors; (2) the content or structure of ATD training, i.e., course syllabi and training scenarios; (3) the management of student learning; and (4) the management of ATD training resources, e.g., ATD scheduling and maintenance. Negative attitudes generated by any of the above can affect the conduct of a valid evaluation process and, therefore, the test director should be familiar with them.

Instructor attitudes. Instructors' attitudes towards ATDs have been identified as a most important variable in determining student perceptions of ATD training. Students, particularly those inexperienced with the ATDs (e.g., UPT, UNT, or transition training students), will tend to adopt the behavior and attitudes of others, especially those they view as particularly competent or whom they hold in esteem. If their instructors do not value ATD training and do not exhibit favorable attitudes toward that training, students likely will not either.

Negative attitudes on the part of instructors may also affect the quality of ATD training. An instructor with a negative set is not likely to be motivated to use the ATD as best he can, or perhaps even in the manner intended.

Course syllabi and training scenarios. Another major factor related to attitudes toward ATD utilization is the presence or absence of adequate training program syllabi, scenarios, and other training courseware. In general, more favorable attitudes result if training syllabi and scenarios are perceived as realistic. Unrealistic programming of problems, e.g., illogically correlated malfunctions, presenting system failures at an unrealistic pace, or the use of training procedures and tactics that are no longer practiced in the operational aircraft, may have adverse effects on trainer acceptance. In short, attitudes toward an ATD may be influenced by the perceived credibility and relevance of the training provided.

Management of student learning. The student's attitude toward ATD training also will depend on his daily experience with that device and the manner in which ATD training is managed. In this regard, it is important that the student: (1) understand the significance of each ATD task and the need for its mastery; (2) make progress toward mastery of relevant skills during each practice session; (3) perceive that progress is being made; and (4) feel that his efforts are respected and valued by his instructors.

The quality of instruction, including performance evaluation, can be a major problem in trainer acceptance. Poor quality training is often due to the assignment as ATD instructors of operational personnel who are not fully knowledgeable concerning the trainer's capabilities, the learning process, and/or the ATD instructor's role.

Management of ATD training resources. Managerial practices related to training, resource scheduling, and maintenance have previously been identified as factors influencing attitudes toward ATDs. The reliability of modern ATDs allows them to support 20 or more hours of training per day. Frequently, ATD training is conducted 16 hours per day, six days per week. However, scheduling ATD training during the instructor's normal off-duty hours, especially without giving

instructors adequate recognition for their efforts or providing them compensating time off, can create strong resentment toward ATDs and ATD training. In undergraduate training programs (e.g., UPT or UNT), students are protected by regulations from such schedules; however, this is not the case for continuation training where a student might be scheduled for ATD training late at night and still be expected to report for work the following day at the usual time. Obviously, training programs that generate such increases in workload are going to be a source of resentment, especially if the training provided by that program is not highly valued.

In continuation training, ATD training typically is specified as a recurring requirement that must be satisfied a fixed number of times per training interval, usually quarterly. With few exceptions, all operational aircrew members must undergo the same number of ATD training sessions regardless of their experience or skill level. The frustration and resentment that experienced aircrew members may feel toward such failure to recognize their individual training needs or individual skill levels may be directed toward the training devices themselves. This, in turn, might result in attitudes that could affect an ATD OT&E effort involving such personnel.

For many of the older training devices, certain important modifications made to the operational equipment systems have not been made on the trainers. Such lack of due attention to the management of ATD training resources produces negative attitudes and may serve to reinforce existing perceptions that ATD training is held to be of little utility. These attitudes may then extend undeservedly to new ATDs that, in fact, are capable of providing meaningful training for the acquisition and maintenance of highly developed skills, or to ATDs that are designed to provide full acquisition of basic skills.

Inadequate ATD maintenance leading to lack of device availability, disruptions during training, and training with degraded systems is an additional factor that may produce negative attitudes towards ATDs. Maintenance-related interruptions of training sessions, whether during an OT&E or not, are frustrating to instructors and students alike. It should be noted that long delays and difficulties in obtaining replacement parts also have a negative impact on the attitudes of ATD maintenance personnel, thereby further complicating the problem.

MANAGEMENT OF ATTITUDES DURING OT&E

Attitude effects must be dealt with regardless of the evaluation design selected for use during a particular ATD OT&E. Whether an analytic or a demonstration evaluation design has been selected, attitude effects must be adequately accounted for. Attitude effects may be

compensated for during an ATD OT&E through formal and/or informal procedures. Formal methods are those which provide "control" for such effects within the experimental design, by means of the data analysis procedures utilized, or both. Informal methods include guidelines and suggestions intended to minimize the potential for an undue influence of attitudes--positive or negative--during the ongoing OT&E.

From a technical perspective, formal attitude effects management methods are preferred to the informal methods. Formal methods are more precise and the results from their application are more readily recognized. However, they are difficult to apply in most ATD situations. Informal methods, on the other hand, while more feasible to use, have a principal weakness in that there is no way of knowing whether their application was successful.

Formal Methods

Formal attitude effects management methods are not simple to use. To begin with, their use depends upon being able, in some fashion or other, to determine in quantifiable terms just what the prevailing attitudes are. Such a determination usually requires that an adequate attitude measuring instrument is available. Unfortunately, such measurement devices generally are not readily available to the ATD OT&E test director. This means that a test director who elected to employ formal attitude management procedures would have to develop his own attitude measurement scale.

The development of valid attitude measuring instruments is much more technically involved than most people recognize. The construction of a "good" attitude measuring instrument requires a technical expertise beyond that usually available to the ATD OT&E test director. Furthermore, the construction and validation of such a scale can be time-consuming and usually requires subject resources not readily available during ATD OT&E. Less formal methods, therefore, offer a more cost-effective and practical means of minimizing test bias during ATD OT&E due to attitude effects.

Informal Methods

By the time the test director is ready to finalize his test, he should have developed an awareness of the general attitudes held toward ATDs, and perhaps the device of specific concern during the forthcoming test, by the students, instructors, and evaluators who will be involved. If he suspects that attitude bias may be a problem, he should follow one or both of the following procedures.

Control by elimination. With this method, those persons possessing an undue extreme attitude toward ATDs are excluded from participation in the experiment. This typically is an inappropriate

method of controlling student attitude effects, but it can be a usable method for selecting/screening instructors and/or performance evaluators. Persons whom the test director judges should be excluded may, for whatever reasons, be allowed to continue their participation even though their data may not be used. It is usually best, however, that they not be allowed to do so. Their presence, their behaviors, and verbalizations can often be a contaminating influence on the other participants and on test results.

It can also happen that undesirable attitude effects are not identified until during the conduct of the test itself. There are obvious reasons why the test director should be reluctant to eliminate participants midstream in test, but in some situations he may have no other choice. He is particularly obliged to address any attitude behaviors which become disruptive.

Control by distribution. In this method, bias due to attitude effects is distributed equally between the different demonstration and control groups. For example, if an instructor with apparent extreme attitudes toward ATDs necessarily had to take part in the evaluation, it might be arranged that he instruct equal numbers of students from each group involved. Similarly, a biased evaluator could be required to evaluate the performance of an equal number of students from each group.

Although the above two techniques for controlling attitude effects have been discussed separately, they can be combined in a number of ways to achieve the desired control. For example, those instructors and evaluators with extremely favorable or unfavorable attitudes toward ATDs might be eliminated, student attitudes might be matched between groups, and any remaining bias due to attitudes might be distributed by having instructors and evaluators instruct/evaluate an equal number of students from each group.

COMMENT ON EVALUATION METHODS

The discussion in Chapters 1-4 of this volume has sought to establish a background acquaintance with many of the factors of concern in planning and executing an ATD OT&E effort. The general discussion of aircrew training, evaluation designs, measurement, attitudes, and similar factors will serve to assist in developing an appropriate understanding of such factors by the test director and, thereby, assist him with his planning for the OT&E. However, he needs more specific information in certain of the details, techniques, and procedures involved. Therefore, the next three chapters deal with much more specific topics.

Because of the pervasive importance of questionnaires and rating scales as evaluation methods, these topics are treated in some detail in Chapter 5. For similar reasons, transfer of training methods are treated in Chapter 6. Chapter 7 discusses evaluation of the instructor/operator station of the ATD and some of the particular evaluation problems it presents. Finally, there is an appendix that presents specific procedures relating to selected statistical analysis methods that are appropriate to ATD OT&E. Thus, while Chapters 1-4 of this volume have dealt more with the "what" and "why" of ATD OT&E, Chapters 5-7 deal much more with the "how" of ATD OT&E.

CHAPTER 5

RATING SCALES AND QUESTIONNAIRES

INTRODUCTION

Rating scales, properly developed, can be useful evaluation tools, particularly when it is not feasible to observe actual training and the evaluation must depend upon the judgments of subject matter experts. In addition, they provide a method for obtaining estimates of a device's training effectiveness prior to its introduction into the operational environment. Rating scales, in fact, have been used extensively as data collection instruments during in-plant ATD IOT&Es, and it is clear that they will continue to play an important role in future training program and training device evaluations. The rating scale method, as an analytical technique for assessing the operational effectiveness of an ATD, has many advantages, but also some limitations. It can, however, be extremely effective and reliable provided that it is used properly, and provided that its limitations are recognized and dealt with.

Advantages of Rating Method

A rating scale method has several advantages over a demonstration approach such as TOT.

First, in some cases rating techniques for evaluating an ATD may be the only methods that can be applied. For example, I/QOT&Es, conducted in the contractor's facilities, typically are not amenable to evaluations employing a demonstration approach. To attempt a TOT evaluation in this environment, for example, would be time consuming, expensive, and, in most cases, unworkable.

Second, the rating scale method typically is easier to implement and more flexible than the various demonstration methods. It does not require the establishment of separate control groups and the associated tasks of matching subject characteristics and equating conditions among ATD and control groups.

Third, the rating scale method can be implemented with minimum disruption of normal training operations. This is an important consideration when an ATD is to be evaluated in an operational setting (e.g., Phase II FOT&E).

Fourth, use of the rating scale method may be more appropriate than experimental techniques under some circumstances. For example,

it would be difficult to assess device fidelity or perceived instructor workload with demonstration methods. Similarly, it would be difficult to assess the training capability of an ATD for tasks that cannot be taught in the aircraft (e.g., certain emergency procedures) using such methods as TOT.

Fifth, the rating scale method also allows an ATD evaluation to be conducted at a level of specificity that cannot be achieved easily with demonstration methods. For example, the training capability of ATD instructional features (singularly, or in sets) can be estimated easily with the rating scale method, whereas determining their training capability would be extremely difficult using demonstration methods.

A properly conducted rating scale evaluation is capable of providing quantitatively meaningful data concerning the expected training capability of an ATD, and/or its potential value to the Air Force as a training resource. Not only is the rating method an acceptable alternative for evaluating an ATD, but in some cases it can be the preferred alternative, especially for those cases where the required resources (e.g., subjects) are not available or where the level of control required of a rigorous experimental evaluation cannot be achieved. A well conducted rating method evaluation is always preferred over a poorly managed experimental evaluation.

Limitations of Rating Method

The rating scale method, as a technique for OT&E, also has certain limitations.

First, rating data typically represent estimates rather than demonstrated training results. Rating data, therefore, ultimately must be verified or confirmed. Estimates derived from the rating method can be verified (validated) experimentally or by operationally demonstrated effectiveness. For example, estimates of the training capability of an ATD derived during an IOT&E can be checked against transfer-of-training data collected subsequently during FOT&E or by later observations of actual operational training application.

The second, and perhaps the greatest, weakness of the rating method is the ease with which it can be applied improperly. Anyone can develop a rating scale and collect data with it: Whether or not the collected data are meaningful will depend upon how well the rating scale was constructed, the manner in which data were collected, and how the collected data were analyzed and interpreted. The successful use of rating scales in evaluating an ATD, therefore, requires well constructed rating scales, evaluators who are well trained, careful management of the rating process, and appropriate statistical analysis of the obtained data.

Chapter Organization

This chapter is subdivided into three major sections: A, B, and C. Section A provides a discussion of basic rating scale concepts, including kinds of scales, types of scaling methods, and techniques for developing rating scales. Also discussed are the kinds of variables that can influence ratings. This section concludes with a discussion of the importance of proper management of the rating process. Section B of this chapter describes specific rating methods for assessing the fidelity and training capability of an ATD. Sample rating scales are also made available, and guidelines for analyzing and interpreting the collected data are provided. Section C provides guidance on how to construct questionnaires for use during the ATD evaluation process. The proper role of questionnaire use during the evaluation process is also discussed.

A RATING SCALE CONCEPTS

INTRODUCTION

As pointed out in the introduction of this chapter, a properly conducted, well done rating scale-based evaluation can be an important component of ATD OT&E. Therefore, it is important that the test director have a clear understanding of the concepts upon which good rating procedures depend and that he be able to distinguish effective rating scales from those that are not well constructed. This section of the chapter is intended to provide that understanding and ability. It first provides a basic discussion of measurement scale concepts as an essential prerequisite to an understanding of rating scale construction and use. The section then addresses the actual construction of rating scales, including a discussion of factors that influence rating scale design. It concludes with a discussion of approaches to counteracting factors which can bias ratings.

TYPES OF MEASUREMENT SCALES

It is necessary to know what types of measurement scales exist in order to understand the properties of the data obtained with those measurements. There are basically four kinds of measurement scales: nominal, ordinal, interval, and ratio. It is important to note that, as one moves up the hierarchy from nominal to ratio, each scale will have all of the qualities of the preceding scale, plus one or more qualities that the preceding scales do not have.

Nominal Scale

A nominal scale simply names the categories (nominal means named only). Actually, the nominal scale isn't a true scale at all; it simply names categories, without suggesting any numerical or order relationship among those categories.

The nominal scale is very common. For example, people may be classified as either males or females, smokers or nonsmokers, pilots or nonpilots. Data from a nominal scale can be analyzed statistically, but only certain analysis techniques can be used with nominal data. As a result, sensitivity of the analysis to training or ATD effects will be less than for higher level measuring scale data and their appropriate analysis techniques.

Numbers can be assigned to categories on a nominal scale, but such numbers do not imply any kind of order. Sometimes the assignment of numbers to categories can be misleading, because people are accustomed to associating numbers with order or quantity. For example,

aircrew members might be assigned a number on the basis of their crew position: pilots = 1; copilots = 2; and navigators = 3. However, these numbers have nothing to do with order or quantity.

Ordinal Scale

The ordinal scale has an advantage over the nominal scale, because it describes quantitative order; i.e., an ordinal scale uses numbers to rank items from least to greatest. For example, runners in a race can be ranked according to the order in which they finished; first, second, third, fourth, and so on. Thus, the ordinal scale indicates order, but it provides no more than that kind of information. Nothing can be said, for example, about the magnitude of the difference in time between those runners who finished first, second, or third.

Interval Scale

With an interval scale, the magnitude of the interval between items is known, in addition to the order of the items, because the units of measurement in an interval scale are equal. Thus, statements can be made about the size of the difference between any two measures. However, the interval scale does not have an absolute zero point. As a consequence, statements cannot be made about the ratio of one quantity to another (e.g., Pilot A's proficiency is two times as great as that of Pilot B).

Standard measures of temperature (e.g., Fahrenheit, Celsius) are interval scales of measurement. For example, if it was 40° on Monday, 30° on Tuesday, and 20° on Wednesday, the successive difference in temperature between each of these three days is the same, i.e., 10°. However, it was not twice as hot on Monday as it was on Wednesday even though 40 is two times 20. Moreover, a temperature of zero does not mean that there is no temperature.

Determining whether a given scale is an interval scale or an ordinal scale may be difficult at times ("experts" sometimes disagree), because a scale can vary in the degree to which it produces interval level data. Nonetheless, interval scales allow use of the arithmetic mean (commonly referred to as "the average") as a measure of central tendency and, consequently, allow use of more powerful statistical tests than would otherwise be possible. If there is a serious question whether or not data are from an ordinal or an interval scale, the more conservative approach is to assume the data to be from an ordinal scale.

Ratio Scale

In ratio scales, not only are the intervals of measurement equal, but there is also an absolute zero point. This makes it possible to

make statements about the ratios between any two values on a ratio scale. An example of a ratio scale is a standard measure of distance (e.g., inches) in which the distance between 1 and 2 inches is the same as the distance between 2 and 3 inches; 2 inches is twice as long as 1 inch; and zero inches indicates zero distance.

RATING SCALE CONSTRUCTION

There are a number of factors to be considered in constructing a rating scale, including (1) the number of categories to use; (2) whether to use an even or odd number of categories; (3) whether all categories should be labeled, or just the endpoints of the scale; (4) if categories are labeled, how should those labels be selected; and (5) if category labels are to be described, how should those descriptions be written.

Number of Categories

A rating scale obviously must have at least two categories. However, if the scale has that few steps, much of the ability of the raters to make finer discriminations may be lost. On the other hand, raters will find it difficult to use a scale with too many steps, especially if the number of distinctions exceeded the raters' powers of discrimination. The optimum number of scale points also will depend upon the willingness of the raters to make the effort to use the discriminative powers they have. Most rating scales contain between five and nine categories. This number allows the rater enough choices without his being overwhelmed. Seven-point scales probably are the most common. They seem to be consistent with subjects' introspections on the number of discriminations they can make. Researchers have found that five or seven scale points are an optimum number for most purposes, and that fewer divisions irritate respondents; also, larger numbers were found to produce unsatisfactory response distributions. In terms of the reliability between different raters, five or seven steps also appears to be optimum. Fewer steps may be used, if the object being rated is rather obscure, if the raters are untrained and only moderately interested, or if a number of ratings of different aspects of the thing rated are to be combined.

The choice of the number of categories should be driven by the purpose, or decision making level, of the rating scale and the desired sensitivity of the ratings to changes in the object being rated. It would be wasteful to collect data with a seven- or nine-point scale if those data were going to be reduced to support a binary decision. For example, in order to make a preliminary determination of whether or not an ATD provides the necessary cues and response opportunities to train a particular task, it may be easier, and just as effective, to rate each task or subtask in question as "trainable" or "untrainable"

with the device, rather than to collect ratings on a five- or seven-point scale and then immediately reduce those ratings to the two categories. On the other hand, if the device is being rated to estimate potential transfer of training ratios, then finer grain distinctions become important. In cases of this sort, the data should be collected on a rating scale with a greater number of steps, preferably five or seven.

Odd vs. Even Number of Categories

Another factor to consider is whether to use an even number or an odd number of categories. Scales with an odd number of divisions are preferred in order to provide respondents with a neutral position on the scale. Scales with an even number of divisions sometimes are used, however, if it is deemed advisable to force respondents to choose one pole or other to improve the discriminative power of the instrument. As a general rule, it is recommended for most OT&E purposes that an odd number of categories, with a neutral point, be used.

Category Labels

A major decision that must be made concerns how to label the categories. Category labels are particularly important, because they not only tell the rater what he is to be thinking while rating, but they can affect the measurement quality of the data. Rating scales should be constructed so as to generate at least interval data (see preceding discussion on type of measurement). Labeling only the end points of a scale or placing numbers over each scale category (see Figure 5-1 [a] and [b]) is generally considered to provide interval data.

Combining the labeling approaches from Figure 5-1 (a) and (b), as is shown in Figure 5-1 (c), would also be acceptable from the perspective of providing interval level data. Adding additional verbal labels as is shown in Figure 5-1 (d) may be acceptable, but so doing risks possible degradation of the true interval quality of the data. These verbal labels, "agree strongly," "agree," "neutral," etc., may not represent equal intervals. For example, one might question whether the difference between "agree" and "strongly agree" in scale (d) is the same as the difference between "agree" and "neutral." If subjects do not consider the differences to be equal, the rating scale will lack the essential property that makes it an interval scale, which, in turn, will impose a limitation on the type of statistical operations that may be performed on the data.

The examples provided in Figure 5-1 are relatively simple, in that the scales range from strong agreement to strong disagreement. Choosing appropriate labels that maintain an interval scale property becomes more difficult if the labels used are more descriptive and/or

when the scale must span a more difficult concept. For this reason, it is recommended for ATD OT&E purposes that the labeling of rating scales generally be confined to an approach similar to that illustrated in Figure 5-1 (c). This is true whether five- or seven-point scales are used.

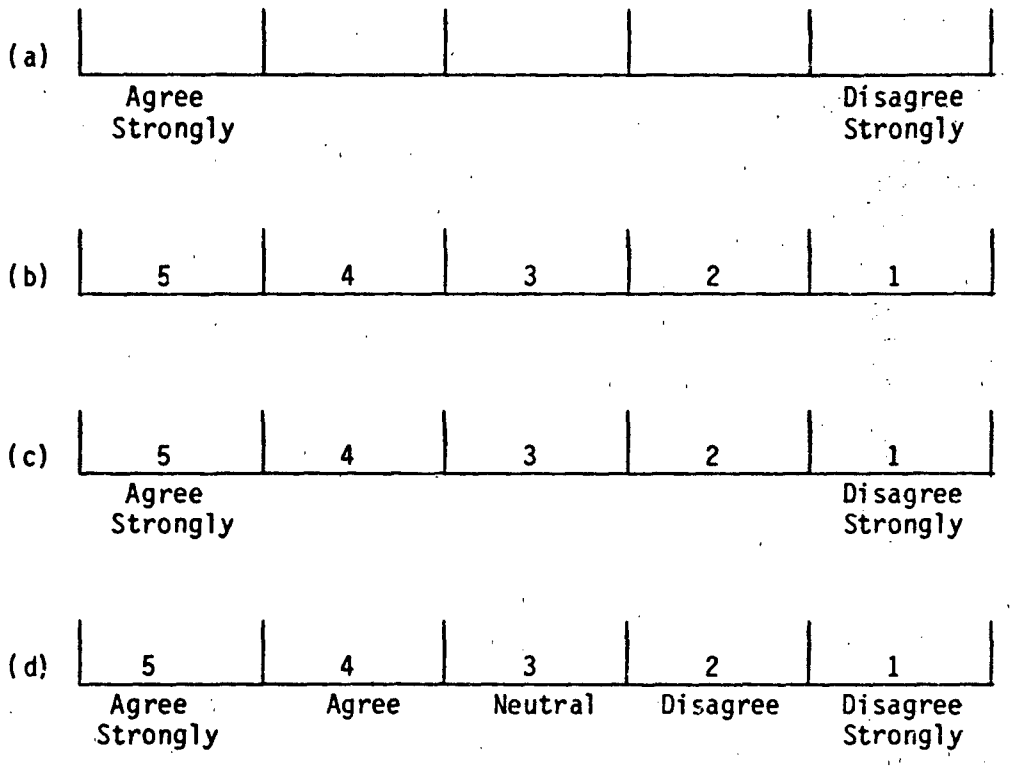


Figure 5-1. Examples of rating scales.

Category Descriptions

The way in which evaluators interpret the verbal descriptions provided for every category being rated will be a crucial factor in the accuracy and validity of their ratings. The following guidance provides some suggestions for constructing/selecting scale labels or descriptions, some simple tests that can be used to assess the adequacy of the scale, and some solutions to common problems:

1. Category descriptions should be as short as possible without omitting any important information. In many cases raters will commit category descriptions to memory, especially if they have many ratings to complete. Lengthy descriptions are difficult to memorize; moreover different raters are likely to attend to different aspects of category descriptions if they are too long. Lengthy descriptions may contain multiple criteria and/or dimensions on which to base the rating. Consequently, it is impossible to determine which, or how many, of the criteria the rater is using to make his evaluation.

   Consider the following example from a rating scale that has been used to evaluate the instructional features of an ATD:

Description	Rating
POSITIVE training value; information required to monitor/evaluate the sequence of events was readily observed/recorded. Aircrew deviations, substandard performance, incorrect techniques/procedures were readily detected as they occurred. Instructor interface is uncomplicated/ expeditious, allowing anticipation of reactions.	5

   There are at least three dimensions contained in this scale point description: (1) information requirements, (2) real-time detection of aircrew deviations, and (3) instructor interface relative to the anticipation of crew reactions. Each of these dimensions should be addressed separately.

2. Category descriptions should be written in simple language that is easily understood by the people who are to use the scale; the use of jargon, unfamiliar terms, or ambiguous concepts should be avoided. The following example also was taken from a rating scale that has been used to evaluate the training capability of an ATD:

Description	Rating
POSITIVE training capability; activity or system simulation offers realism equivalent to actual aircraft operation.	5

The term "realism" may be ambiguous, i.e., it may have multiple meanings. For example, realism may refer to the physical correspondence of the device to the aircraft; it may refer to the functional correspondence (e.g., visual cues) of the device to the aircraft; or it may refer to the potential dangers or elements of an actual airborne environment. More importantly, the meaning of realism may vary between raters; hence, they may use different criteria on which to base their evaluations. Ambiguous terms should be avoided or defined in a manner that renders them unambiguous.

3. Category descriptions should be distinct in meaning from other category descriptions on the same scale. Otherwise, raters may be confused over which scale point to use. For example, it may be difficult to distinguish among the following three descriptions taken from a training capability rating scale.

Description	Rating
Training capability is nearly equal to that experienced in the aircraft.	5
Training capability is less than that which would be experienced in the aircraft.	4
Minimal training capability that must be complemented by training in the aircraft.	3

Distinguishing between "training capability is nearly equal to that experienced in the aircraft" and "training capability is less than that experienced in the aircraft" may be difficult for the respondent. Also, a rating of 3 indicates that complementary aircraft training will be required; however, that might likely appear to the respondent to be the case for ratings of 4 or 5, thus making the scale discriminations difficult for him.

4. Successive category descriptions on the same scale should be constructed so that they represent equal intervals of subjective judgment. As an absolute minimum, raters should be able to rank order the category labels/descriptions without knowledge of their corresponding scale numbers. For example, if

we constructed a five-point scale with the labels, "Good," "Fair," "Poor," "Excellent," "Very Good," it would be relatively easy to arrange these labels in order of increasing "goodness."

However, arranging the following scale labels, taken from a training capability rating scale, in an ascending or descending series (rank order) is not easily accomplished:

ACCEPTABLE GOOD POOR POSITIVE NEGATIVE

The chief problem here is that the terms "Acceptable," "Good," and "Positive" are similar and are difficult to discriminate from one another, as are the terms "Poor" and Negative." Every scale should be subjected to such an ordinality test.

Note that if it is not possible to rank order category descriptions in the manner discussed above, the scale probably fails to meet the requirements of an ordinal level scale. The failure of category descriptions to meet ordinality requirements may be because they are written in an ambiguous manner, or because they represent a dimension (aspect, factor) different from the dimension represented by the other category labels that constitute the scale.

5. All the category labels/descriptions of a given scale must represent the same dimension of whatever is being rated. For example, if we were rating a concept along the dimension "good-bad," then all of the scale descriptors must belong to, or be representative of, the "good-bad" dimension. Consider the following scale that embodies the dimension of height:

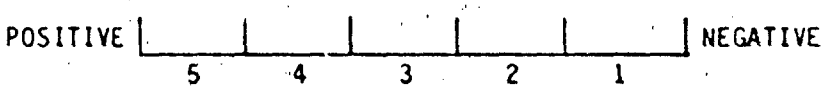
Label	Rating scale
Very Tall	5
Tall	4
Average	3
Small	2
Very Short	1

In this scale, the label "Small" may refer to height, weight, or both. Hence, it does not unequivocally represent part of the dimension of height and should not be used as a label for this scale. This problem is easily remedied by replacing the

word "Small" with the word "Short." When all of the labels of a scale are representative of the same dimension and only one dimension is addressed, the scale is said to be unidimensional. Unidimensionality is a fundamental requirement of a properly constructed scale. It also is one that when violated decreases the likelihood of the scale providing even ordinal level data, and increases the likelihood that the scale will be ambiguous and confuse the raters who must use it.

Possible solutions. The types of problems with labels/descriptions that have been discussed are fairly common, not only in ATD evaluation efforts, but in other kinds of program evaluations as well. For example, consider the scales depicted in Figures 5-2 and 5-3. Many of the problems discussed are represented therein. However, most of these problems have relatively simple solutions. It is not necessary to provide a description, or even a label, for every scale point. In fact, whether or not intermediate scale steps are labeled appears to have little effect on the way raters distribute their ratings along a scale, except in cases where scale descriptions are confusing or ambiguous. Therefore, many of the problems of scale descriptions discussed above may be eliminated by dropping the intermediate labels, keeping labels brief, and striving for unidimensionality. The five-point numerical rating scale shown in Figure 5-2, for example, could be rewritten and displayed graphically as follows:

Description	Rating
POSITIVE: Activity or system simulator offers training capability equivalent to actual aircraft operation	5
NEGATIVE: Activity or system simulator offers training capability that is totally unacceptable	1



Description	Rating
POSITIVE training capability; activity or system simulation offers realism equivalent to actual aircraft operation	5
GOOD training capability; insignificant but perceptible departures from realism; detection of unrealistic cues requires close scrutiny	4
ACCEPTABLE training capability; deviations from realism attract the attention of the operator, but are not significant in the overall event scenario	3
POOR training capability; deviations or unrealistic cues are distracting and readily apparent without close operator scrutiny	2
NEGATIVE training capability or unsafe; totally unrealistic cue or use may result in injury to personnel or damage to equipment	1
Not tested; outstanding TD/SR	0

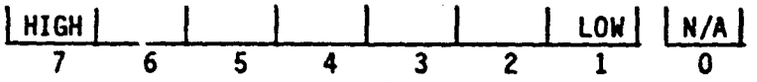
Figure 5-2. A training capability rating scale designed to assess the training capability of an ATD.

Description	Rating
Fidelity that will provide a level of training identical to the aircraft	7
Fidelity that will provide training similar, but not identical, to the aircraft	6
Fidelity is such that minor improvements will enhance training potential	5
Fidelity is such that minor improvements are desired to improve training potential	4
Fidelity is such that minimum training potential exists; minor improvements are recommended	3
Fidelity is such that the OFT provides very little training potential; improvements are required	2
Fidelity is such that very little training potential can be realized from the OFT; negative training may occur; major improvements are required	1
Not available for evaluation	0

Figure 5-3. A fidelity rating scale designed to assess the fidelity of an ATD.

Similarly, the scale in Figure 5-3 could also be improved by deleting the intermediate scale descriptions as shown below:

Description	Rating
HIGH: Fidelity that will provide a level of training identical to that provided by the aircraft	7
LOW: Fidelity is such that negative training may occur; major improvements are required	1



FACTORS INFLUENCING RATING SCALE RESULTS

In designing a scale, it is also important to know which factors can unduly influence the results and which factors will be irrelevant. These factors can be classified into two major categories, those dealing with scale characteristics and those dealing with rater characteristics.

Scale Characteristics

Serial position. The serial position of an item, i.e., its location within the sequence of items, is important. It has been shown, for example, that the first and last items in a list tend to receive higher ratings than other items. This potential problem can be controlled for by constructing several forms of the rating questionnaire, each of which presents the items in a different position. Any bias induced by serial position effects will be equalized when the data for all raters are averaged. Thus, in a ATD evaluation, it may be necessary to vary systematically the order in which tasks are rated or evaluated.

Anchor labels. The choice of anchoring labels also is important. Some aspects of labeling have already been discussed, but there are others of concern. For example, raters tend to avoid using extreme labels. Hence, the use of extreme labels as anchors, or scale endpoints, may cause raters to use only the middle categories of a scale, thus functionally reducing the number of steps in the scale. This can

impose a serious restriction if there are few scale steps to begin with (e.g., five or fewer). The four-point scale below, designed to assess instructional feature use, might be, for all practical purposes, a two-point scale:

1 = I would never use this feature.

2 = I would occasionally use this feature.

3 = I would frequently use this feature.

4 = I would always use this feature.

It is equally important not to select anchors that are too neutral. Otherwise, the majority of responses are likely to fall along the endpoints of the scale.

The net effect of choosing anchors that are too extreme or too neutral is functionally to reduce the number of categories in the scale and, hence, reduce its power to discriminate differences in that which is rated.

Finally, the familiarity of the terms can influence judgment. The reliability of the scales will be reduced if unfamiliar anchoring terms are used, or if unfamiliar concepts are to be scaled.

Format. The actual format of the scale is relatively unimportant. It doesn't appear to make a difference whether one uses horizontal scales (left to right) or vertical scales (up and down). It also doesn't seem to make any difference which anchor term is on the left and which is on the right. However, for vertical scales, it is usually recommended that the highest ranking be placed at the top of the scale.

Numerical scales, consisting of a sequence of numbers with precise definitions, allow subjects to assign a number to whatever is being rated in accordance with these definitions or descriptions. For example:

Excellent = 5

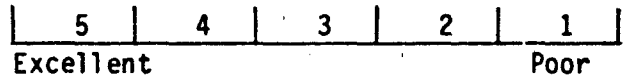
Above Average = 4

Average = 3

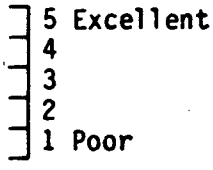
Below Average = 2

Poor = 1

Although numerical scales of this type have been used successfully, it is frequently recommended that they be combined with a graphic scale format. One possibility using a horizontal scale is shown below:



Another possibility using a vertical scale format might be as follows:



There are numerous acceptable variations to the two examples shown. However, whenever possible, it is advantageous to display the entire scale, to label the endpoints, and to denote each scale point with a number. Displaying the entire scale may increase the likelihood that the rater will consider the entire range of ratings for each rating he makes. Providing numbers for each scale step may help to reinforce the instructions to treat each scale step as an interval of equal psychological judgment. The examples above have the further advantage of allowing ratings to be made quickly by simply placing a checkmark at the appropriate location on the line. This is especially convenient if many ratings must be made.

Rater Characteristics

Response style. The way that people use rating scales, or their response style, is an important aspect of rating scale use. Some people show extreme response style--i.e., they rate items as either very good or very bad, and they do not seem to have neutral opinions on anything. This rater style is most likely to be evidenced whenever raters are exceptionally familiar with, or involved in, the topical area being rated. Other people have a neutral response style--i.e., they tend to place their ratings close to the center of the scale. This latter tendency is especially likely to occur when the raters are unfamiliar with the subject matter object or concepts that they are rating.

Leniency effects. Some raters also tend to use the "pleasant" end of a scale almost entirely, whereas others tend to use the less

favorable end more often. In general, there is a positivity bias in raters; people tend to use the more pleasant end of the rating scale, especially if they are rating another person's performance.

Halo effects. Halo effects occur when raters judge more than one characteristic of an object or person and the rating on one characteristic may be influenced by the ratings on the other characteristics. One result of the halo effect is to bias the rating of any characteristic in the direction of the general impression, or attitude, toward the object rated. To the extent this occurs, the ratings of some characteristics will be less valid. For example, there is some evidence that the rating of the training effectiveness of an ATD for a given task may be strongly influenced by ratings or impressions of its overall performance characteristics, its handling qualities, and/or its visual system. The potential influence of halo effects can be reduced by having evaluators rate only one characteristic at a time with strict instructions to attend only to the characteristic under consideration.

MANAGEMENT OF THE RATING PROCESS AND RATER TRAINING

The conditions under which a rating-method evaluation of an ATD is conducted can affect both the quality and the integrity of the obtained ratings. It is important that each rater provide an independent evaluation, i.e., one that is not influenced significantly by the ratings or opinions of other raters, and one that is not influenced significantly by the conditions under which the evaluation is conducted. Although factors that bias the outcome of a rating method evaluation cannot be eliminated completely, they can be minimized by careful supervision of the rating process and proper training of the evaluators.

Managing the Rating Process

The following guidelines and suggestions are provided for managing a rating method evaluation of an ATD:

- If possible, it is desirable to have raters evaluate only one characteristic or aspect (e.g., visual cues, motion cues, performance characteristics, etc.) of the ATD at a time. Raters also should be instructed to ignore (to the extent possible) the influence of other characteristics. For example, visual cues should not be given a poor rating just because the ATD handles poorly.

- Ratings should be assigned as soon after the task (e.g., examining visual cues) is completed as possible. Ideally, ratings should be assigned as the task is observed/performed,

although in most cases this is impractical. In some cases it may be necessary that an entire mission scenario be completed before ratings are assigned in order to maintain mission continuity. However, too long a delay may compromise the validity of subsequent ratings.

- Task length also should be kept as short as possible/practical. Long tasks can be divided into subtasks. Ratings can then be assigned after each subtask is completed, except in cases when the entire task must be judged as a whole.

- Raters should be instructed not to discuss their impression of the ATD with other evaluators. Otherwise, their ratings may bias (or be biased by) the ratings of those evaluators.

- Ratings should represent the independent judgment of the individual rater, and consultation/discussion with other evaluators prior to assigning a rating should be prohibited.

- Assigning ratings can be a tedious task. Therefore, it is important to schedule frequent rest periods for evaluators. In general, a five-minute break every 15-20 minutes is sufficient.

- The order in which tasks are evaluated should vary for different groups of evaluators, and over the course of the evaluation, to control for order effects.

- The evaluation should be supervised to ensure the consistency of the rating process. Periodic "spot-checks" should be made.

- The rating scales should be pretested by the evaluators who are going to use them. This will provide an opportunity to amend ambiguous or confusing items.

- The procedures for recording and collecting rating data should be worked out and tested prior to the evaluation. If possible, this should be accomplished in conjunction with the pretesting of the rating scales themselves.

Training of Raters

OT&E evaluators typically are selected because they are expert operators of the operational equipment, not because they are experts at assessing the training capability or training effectiveness of an ATD. On the contrary, many of the evaluation concepts they will be asked to employ will be totally unfamiliar to them. They also may be expected to make judgments concerning issues that extend beyond their

area of expertise (e.g., the training capability of an instructional feature, or the extent to which the fidelity provided by an ATD affects its training capability). In addition, evaluators may differ in the amount of their previous aircraft and ATD experience as well as in the favorableness of their attitudes towards ATDs and ATD training. All of these factors can influence an evaluator's frame of reference for making judgments and significantly influence the reliability and validity of his evaluations. The following guidelines suggest some areas of training that appear to be especially important relative to the evaluation of an ATD:

- The raters should be briefed on the basic philosophy of using an ATD as a training resource. It should be stressed that an ATD is a training device designed to provide the cues and response opportunities required to train a given task; it is not a substitute or "ground-bound" aircraft. Raters should be further instructed to base their judgments only on those cue/response opportunities actually required to train a task and to ignore differences in the ATD and aircraft that are irrelevant to training the task being evaluated. The point here is that evaluators should view the ATD as a training device and not as a substitute airplane.

- Raters should be thoroughly familiar with the concepts to be used in the evaluation (e.g., physical fidelity, functional fidelity, training capability, etc.). Prior to the actual evaluation, it is a good idea to have all of the evaluators write definitions of the concepts that are going to be used in the evaluation. Unless all the definitions are consistent, further rater training in this area is required.

- Raters should be aware of the tasks/subtasks for which the ATD was designed to train and of those tasks for which the ATD was not intended. In short, the ATD should be evaluated according to its intended purposes, and all the evaluators should be aware of what those purposes are.

- Raters should be instructed on the proficiency level the ATD was designed to train. Different criteria should be used to evaluate the training capability of an ATD designed to train basic flight skills (e.g., undergraduate pilot training) and an ATD designed to maintain highly proficient performance (continuation training).

- To the extent possible, evaluators should be provided objective standards on which to base their evaluations.

- Raters must be convinced that what they are doing is important and that their efforts are appreciated. The overall

quality of evaluation is closely related to the motivation of the evaluators. In addition, raters should be briefed on the results and what effects their evaluation will have on future system performance.

SECTION B: SPECIFIC ATD RATING SCALE EVALUATIONS

Given the background of the basic rating scale methodology, as discussed in Section A, the test director must still apply those basics to the specific ATD evaluation problems and requirements for which he is responsible. Although most ATD evaluations will involve a variety of types of data--not just rating scale data--the use of the rating scale can be (and usually is) a significant aspect of ATD and OT&E. Therefore, the discussion in this section will treat the application of rating scales in terms of their initial planning and their application to two of the three major parts of a typical ATD OT&E, i.e., (1) fidelity assessment; and (2) training capability assessment. The third major aspect of the ATD OT&E, assessment of the instructor/operator station (IOS) is treated separately in Chapter 7.

A number of activities must be completed prior to the conduct of an operational effectiveness evaluation. For example, one very important activity is the training of evaluators as discussed above. Among these, there are several preparatory activities of particular importance to rating scale development. Two of these activities are defining evaluation objectives and selecting which tasks/activities to evaluate. The following subsections address these activities. Following that, the issues of ATD fidelity assessment and training capability assessment are treated in detail.

DEFINING OBJECTIVES

Evaluation objectives should be as specific as possible and stated in a manner that both suggests what should be measured and the format of the data to be collected. This will make it easy to assess later whether or not, or to what degree, the OT&E test objectives are met.

Test objectives should be based on the Statement of Intended Operational Employment (SIOE) for the ATD, any identified Critical Questions, and on information developed through frequent interactions with representatives of the Using Command concerning their anticipated training requirements and expectations for the device. The SIOE typically will contain information both about the categories of aircrew tasks and the various types of trainees the device is intended to train. For example, the following information is based on the SIOE for an Operational Flight Trainer (OFT):

> Standardization evaluation flight checks will be accomplished in the OFT; checks will concentrate on emergency and instructional procedural evaluation.

- The OFT will be used to achieve proficiency in all normal and emergency procedures prior to the student's first aircraft flight.

- The OFT will be used by students, qualified pilots, and IPs for refresher training on emergency procedures and for normal procedural refresher training whenever aircraft currency lapses; this procedural training will be in both visual and instrument conditions.

- The OFT will be used to achieve proficiency in all phases of instrument flight, including approaches and departures under adverse weather conditions down to and beyond minimums (simulated emergency recoveries). All normal and contingency instrument procedures and maneuvers will be trained (limited circling-approach training due to forward field-of-view-only restriction). Turbulence and variable wind conditions will be included, as will various instrument communication, and navigation equipment degradations and failures.

- The OFT will be used for visual navigation training.

- The OFT will be used to train air refueling rendezvous and basic refueling procedures. This includes emergency and degraded air refueling system operations (receiver and tanker).

- The OFT will be used for conversion, operational, instructor, and transition course pilot training.

- The OFT will be used to support continuation training for staff pilots and instructor pilots and to support local checkouts.

- The OFT will be used to train maintenance personnel in engine run procedures and functional check flight pilots in test flight profile procedures.

- The OFT shall have the capability (within the scope of the limited Night-Only Visual System) for tactical training in a simulated combat environment for many types of real world operational missions.

- The OFT will be used to train systems operation, air-to-air intercepts and air-to-ground weapons delivery, including limited visual weapons delivery on a bombing/storage range.

- All modes of operation of the fire control radar, head-up display, inertial navigation system, and stores management system will be trained.

- The OFT will be used to train normal and backup procedures for selecting, arming, monitoring, releasing and jettisoning all aircraft compatible ordnance, dispensers, and carriage racks.

The overriding objective of an ATD OT&E is to assess the extent to which an ATD satisfies the training it was designed to support. It should be pointed out that other training capabilities of the device may be found to exist during the evaluation process and that these capabilities should be duly noted and documented.

Lists such as the preceding can be developed to identify task areas (e.g., visual navigation), tactical tasks (e.g., air-to-ground weapons delivery), airc aft systems, (e.g., fire control radar), and the different categories of trainees (transition, continuation, etc.) for which the device was designed to train. Such listings can be used to develop the Operational Effectiveness objectives for the OT&E. The general objectives in Operational Effectiveness, then, can be stated as:

1. Determine those tasks, percentage of tasks, or task segments that can be trained in this ATD for each distinct population (category) of student; and

2. For those tasks or task segments for which training <u>can</u> be provided, determine/estimate the <u>degree</u> of ATD training that can be provided.

These general objectives may then be made more specific by identifying those specific task areas or system operations for which the ATD was designed to train. From the examples above, those might be:

- Emergency procedures
- Normal procedures
- Instrument flight maneuver
- Visual navigation training
- Air refueling rendezvous and basic refueling procedures
- Air-to-surface maneuvers
- Air-to-air intercepts

Still more specific objectives can be generated by listing the tasks to be trained under each task area. For example,

AIR-TO-SURFACE

- Air-to-surface combat
- Attack maneuvers
- Ordnance delivery

A further refinement might then be,

ATTACK MANEUVERS

- Pop-up attack
- Loft/LADD type attack
- Level/laydown attack

Figure 5-4 shows diagrammatically the process of making general objectives successively more specific. The process continues down to the task level, or to the subtask level if a given task has meaningfully discrete segments that can be evaluated individually. By mapping objectives in this manner, it is possible to see graphically how the various subobjectives addressed may, in turn, be combined to provide input toward assessing more general objectives. In addition, stating general objectives in terms of determining the number, or percentage, of tasks that can be trained, and/or the extent to which they can be trained, suggests a format for collecting and displaying evaluation data. It also provides a straightforward approach toward determining how well general objectives are satisfied. For example, if an ATD was intended to support the training of all instrument flight maneuvers normally trained in the aircraft, then a measure of the number of maneuvers that can, in fact, be trained in the ATD provides a meaningful measure of the extent to which that general objective was satisfied. Stating evaluation objectives in this manner may make specification of evaluation criteria (Threshold, Standard, and Goal) easier in the sense that it may key the data in a form that makes use of such criteria meaningful.

It may not always be possible to evaluate every possible task/subtask that an ATD was intended to train in the time allotted to conduct an OT&E, especially for sophisticated ATDs that potentially can train virtually all tasks that are normally trained in the aircraft. Therefore, an important pretest activity is to identify which tasks/subtasks will be evaluated. The first step in the selection of tasks is to list all those tasks that the ATD was intended to train. (This should have been accomplished as part of identifying specific test objectives.) Next, the specific tasks/subtasks to be evaluated can be

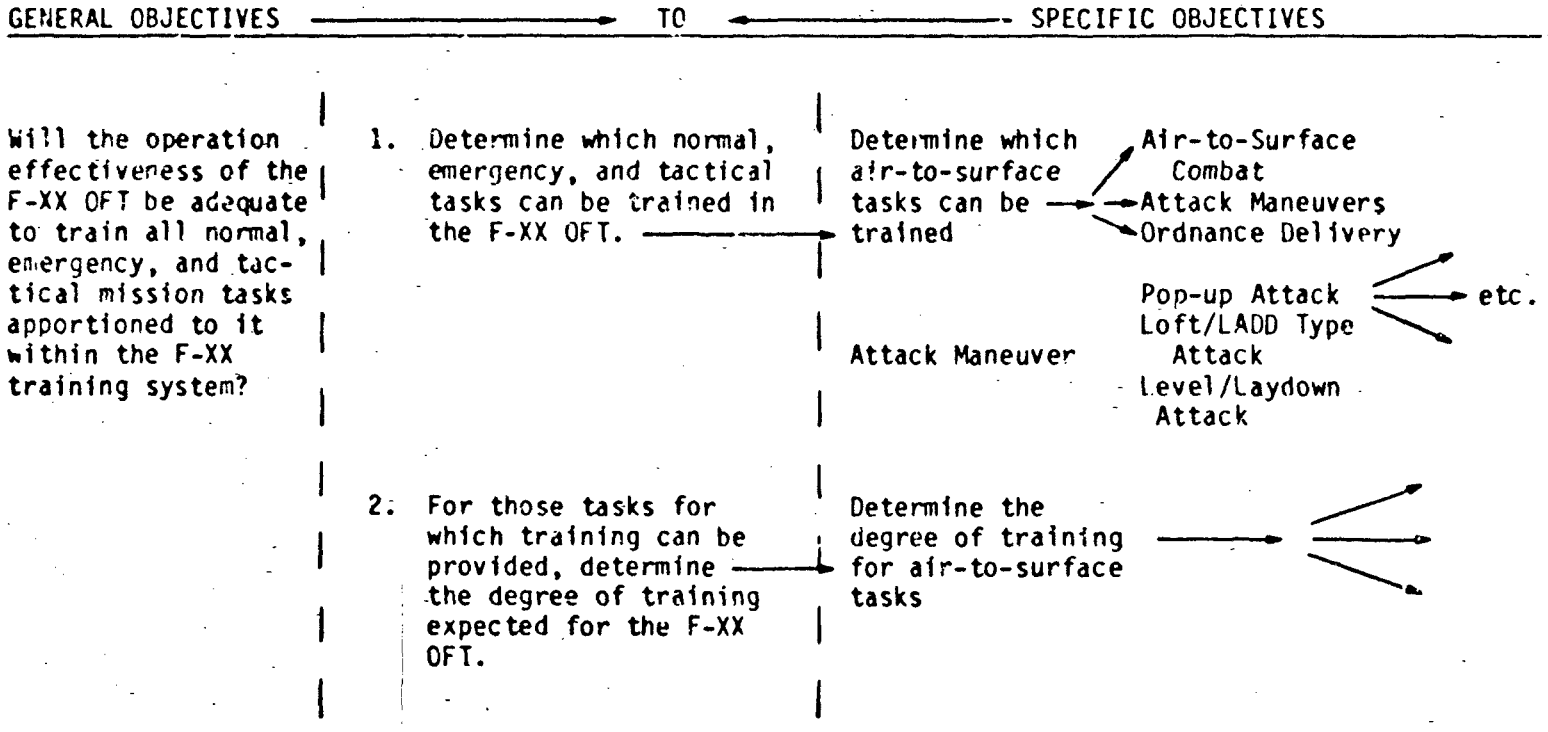


Figure 5-4. Deriving specific objectives from general objectives.

selected. The following guidelines should be observed in selecting tasks to be included in the evaluation:

- The tasks selected must allow sufficient data to be collected to allow specific evaluation objectives to be met.
- The set of tasks selected must be representative of the overall group of tasks the ATD was designed to train. For example, if an ATD is intended to train takeoffs, landings, and air-to-surface weapon delivery, then the selected set of tasks must be representative of those task areas. In addition, the set of selected tasks for a given task area (landings) should be a representative sample of all the individual landing tasks for which the ATD will be used. They should vary in complexity and they should be as independent as possible in terms of the required cues to train the task.
- "High value" tasks should be given priority. This would include tasks that are important for the aircrew member to master, but that cannot be trained safely in the aircraft (e.g., certain emergency procedures); they may be tasks that are especially expensive to train in the aircraft (e.g., tasks that require multiple aircraft such as formation or air-to-air combat); they may represent tasks that are extremely difficult to train and/or ones that contribute significantly to the elimination of aircrew members from training; or they may represent tasks that cannot be trained in a normal training environment, particularly combat skills tasks (e.g., tactical training in simulated threat engagements).
- The selection of tasks should be coordinated with the ASD Test Director if the test is a combined DT&E/QOT&E activity. Some of the task areas selected for evaluation may be covered as part of Acceptance Test Procedures (ATPs).

FIDELITY ASSESSMENT

There are two categories of fidelity of concern to ATD OT&E. The first, physical fidelity is the degree to which physical aspects of an ATD replicate those of the aircraft crew station. Illustrative elements of physical fidelity include control and display location, and their size, shape, and feel. Physical fidelity also can refer to either static and dynamic features such as the cockpit itself (static) or a "g-seat" (motion). The second, psychological fidelity, is the degree to which a simulated event or object contains the same information (cues) as that provided by the event or object in the actual crew station.

There is considerable difference of opinion over which type of fidelity is the more important, or how much fidelity, physical or psychological, is required for an ATD to be effective. Unfortunately, there is no simple resolution. In some cases, physical fidelity is important; in other cases, it is not. On the other hand, a high degree of psychological fidelity is almost always a critical prerequisite to an effective training program because there are few instances where the information available in the operational device is more than necessary to train the task to proficiency in an ATD. However, discussions of the importance of fidelity are meaningful only in the context of the training of a particular task, i.e., considering what is being trained and to what level of proficiency. For example, simple photographs of cockpit panels and instruments mounted on a cardboard background may be sufficient to effectively train simple procedures, but they probably would not be sufficient to train dynamic instrument flight maneuvers.

The purpose of conducting an assessment of the fidelity of an ATD is to determine whether or not the device provides the cues (e.g., visual, motion, aural, etc.) required to learn a particular flying task, and whether or not the device allows the operator to make the types of responses necessary for learning to occur. In short, a fidelity evaluation is conducted to determine if the cues and response opportunities provided by the ATD are sufficient to support training, and to identify and correct any deficiencies in that regard.

Fidelity and training capability evaluations are often conducted during the reliability demonstration portion of the operational suitability evaluation. However, there is not, in most cases, sufficient time during reliability demonstration to conduct a thorough fidelity evaluation and a training capability assessment. If at all possible, the fidelity evaluation should be completed before the reliability demonstration is begun. In that way many of the identified fidelity deficiences can be corrected before the training capability assessment begins and the fidelity assessment can better support the training capability assessment.

A fidelity evaluation involves assessing both the physical fidelity of the simulated crew station or system and the psychological fidelity of the cues required to train a specific set of tasks. Insofar as possible, in evaluating ATD fidelity, the information required to train a task should be compared to the information provided by the ATD for that task.

Physical Fidelity Assessment of Crew Station

The first area that should be evaluated is the actual configuration of the simulated crew station while physical fidelity

assessment is normally accomplished by ASD. The extent of this evaluation during OT&E will depend, in part, upon how far fidelity evaluation has progressed during acceptance testing and what, if any, discrepancies are detected during later evaluations. In some cases, the fidelity evaluation may have been completed in toto during acceptance testing and need not be again addressed during OT&E. Close coordination with the SimSPO acceptance test director is encouraged in order to make this determination.

The initial step in conducting a physical fidelity assessment of a crew station is to make a line drawing(s) of the simulated crew station and identify and number all the systems and panels. Figure 5-5 shows an example of the forward cockpit of an A-10 ATD. Each instrument or panel should be listed including switches, knobs, displays, etc. From this listing a checklist can be prepared to guide the crew station evaluation.

Evaluators should be instructed to check the ATD crew station against the actual crew station with regard to:

- Presence of all crew station instruments, components, parts, and systems. (Components that are not present on the ATD should be noted.)
- The placement of cockpit components.
- Color, lighting, illumination, and appearance.

It is recommended that physical fidelity items be rated according to the following scale:

SUFFICIENT = The physical appearance of the ATD crew station is sufficiently similar to the aircraft crew station to support training.

SUFFICIENT/
REC. CHANGE = The physical appearance of the ATD crew station is sufficiently similar to the aircraft crew station, but changes are recommended to support training.

NOT
SUFFICIENT = The physical appearance of the ATD crew station is not sufficiently similar to the aircraft crew station to support training; changes are required.

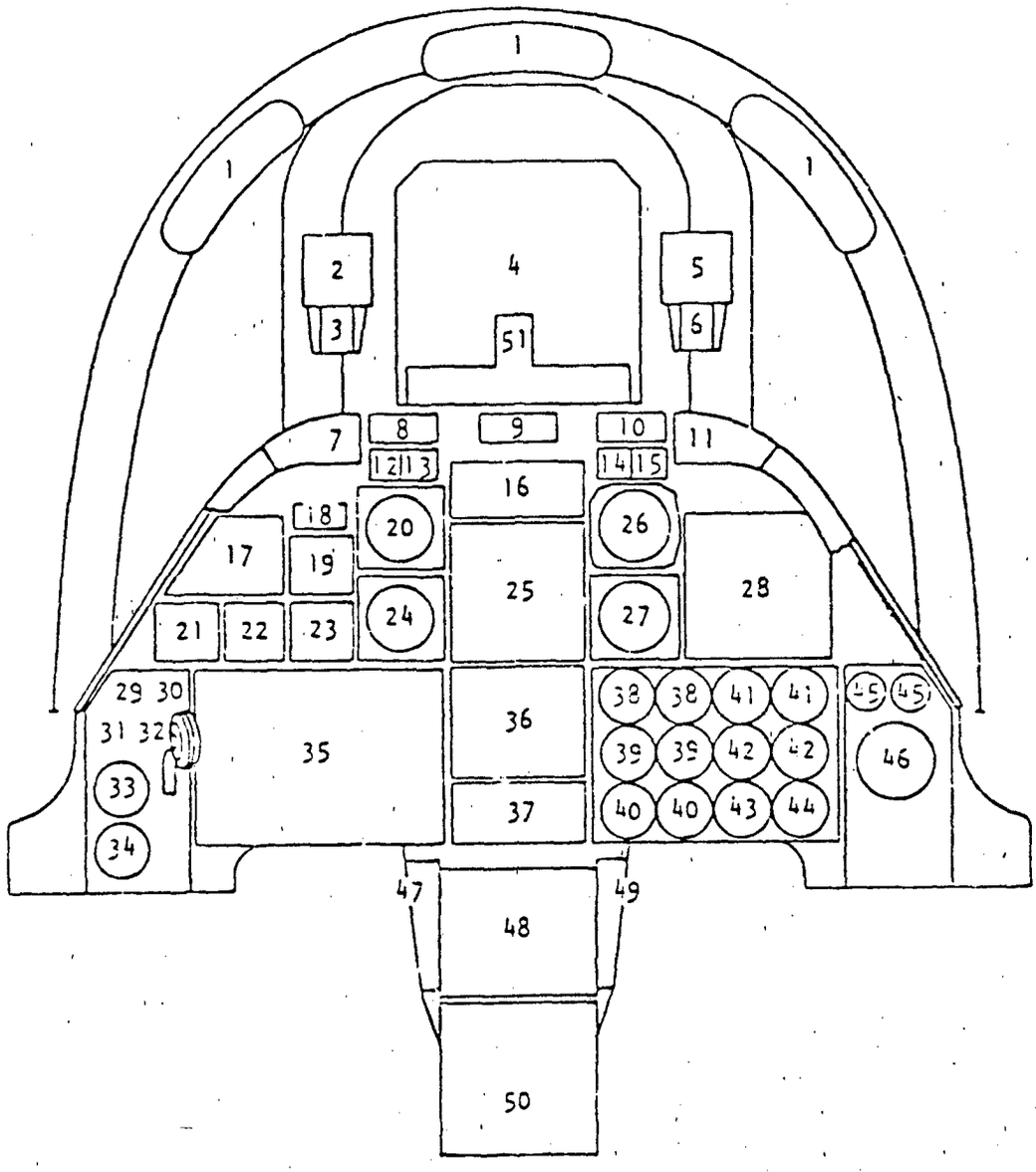


Figure 5-5. Sample line drawing of the forward cockpit of an A-10 ATD.

Figure 5-6 shows a sample data collection format for conducting the physical fidelity assessment of the crew station. Only two or three evaluators are required to conduct this part of the evaluation. Any recommended or required changes should be documented on the data collection form. Any items rated Not Sufficient must be reviewed to determine whether or not a service report is to be issued.

Pyschological Fidelity

After the evaluation of the crew station has been completed, the remainder of the fidelity evaluation should be conducted in the context of actual tasks/subtasks that are to be trained in the ATD and the adequacy of the device to provide the cues or information necessary to support effective training. As was stated above, high cue fidelity contributes heavily to the overall training capability of the ATD.

As an initial step to assessing psychological fidelity, list each task area the ATD is intended to train. Then list those tasks selected for evaluation under each task area. Divide each task into discrete segments or subtasks as appropriate.

For example, a takeoff task might be divided into the following segments:

T/O Roll

Rotation

Airborne

For each task segment, identify all of the cue groups (e.g., visual, motion, etc.) required to train that segment. For example,

Visual Cues (external to crew station)

Motion Cues

Instrument Cues

Aural Cues

Flight Controls

Other Aircraft Systems

Synchronization of Cues

STANDBY COMPASS

Present Absent

___ ___

Sufficient ☐

Sufficient/ Rec. Change ☐

Insufficient ☐

Recommended/Required Change: _____

_____

_____

GUN READY LIGHT

Present Absent

___ ___

Sufficient ☐

Sufficient/ Rec. Change ☐

Insufficient ☐

Recommended/Required Change: _____

_____

_____

AIRSPEED INDICATOR

Present Absent

___ ___

Sufficient ☐

Sufficient/ Rec. Change ☐

Insufficient ☐

Recommended/Required Change: _____

_____

_____

Figure 5-6. Specimen data collection form for crew station fidelity evaluation.

Next, for each cue group, identify the specific cues required to train the task or subtask. Using the Takeoff task as an example, this would provide the following:

TAKEOFF

Visual Cues

Runway centerline
Runway geometry
Rotation
End of runway
Horizon
Etc.

Motion Cues

Acceleration pressure
Runway bumps
Etc.

Instrument Cues

Engine instruments
Flight instruments
Etc.

Aural Cues

Engine noise
Slipstream noise
Etc.

Flight Controls

Stick response
Rudder control
Throttle
Etc.

Synchronization of Cues

Visual and flight controls synchronization
Motion and flight controls
Instrument and aural
Etc.

After cues have been identified for a given task segment, move to the next segment of the task until cues are identified for the entire task. The whole process is repeated for each task selected for evaluation.

In identifying the cues required to train a task, it is best to be as specific as possible. Increasing the specificity of the cue listing helps standardize the evaluation procedures. It also aids in isolating and correcting deficient cues.

Preparing a list of cues for each task/subtask to be evaluated will require considerable advance planning and time as well as the assistance of a group of subject matter experts. Instructional Systems Development (ISD) personnel and flight line instructors may be especially helpful in constructing the necessary cue listings. In the event that the required resources (e.g., time and personnel) are not available to allow listing of all of the specific cues for all of the tasks selected for evaluation, the level of cue specificity can be varied to accommodate available time and personnel. Obviously, the most desirable approach is to develop a highly specific list of task-relevant cues for each cue group. If necessary, only the cue groups themselves can be listed and rated for each task. Doing so, however, is a less desirable approach toward assessing the fidelity of an ATD because it may allow critically important, but deficient, cues to go undetected, and it does not allow as specific an identification of deficiencies and required corrective actions.

Cue fidelity rating scales. The primary purpose of conducting a fidelity evaluation is to identify for correction any deficiencies that are likely to impair/limit the training capability of the device. The cue is either adequate or needs correction. As a result, fairly simple rating scales can be used for these sorts of fidelity assessments. Some specific examples of such rating scales are shown below.

Option A

S = The cues provided by the ATD are Sufficiently similar to the cues required to allow training the task/subtask.

NS = The cues provided by the ATD are Not Sufficiently similar to the cues required to allow training the task/subtask.

Note that the distinction is between the sufficiency or insufficiency of the cues provided relative to those required for training to take place, not in the degree or extent of training that can be expected.

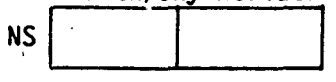
Example:

NORMAL TAKEOFF

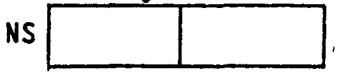
Visual Cues

Aircraft Nose

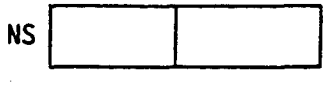
Earth/Sky Horizon



Runway Centerline



Ground Movement



Option B

This option expands the basic two-point scale of Option A to include a third category: cues that are sufficient to train a given task/subtask, but ones that, if improved, would increase significantly the ability to train that task/subtask.

S_1 = The cues provided by the ATD are Sufficiently similar to the aircraft cues required to train the task/subtask.

S_2 = The cues provided by the ATD are Sufficient to support training, but, if improved, would improve/enhance training.

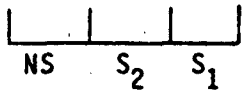
NS = The cues provided by the ATD are Not Sufficiently similar to the cues required to train the task/subtask.

Example:

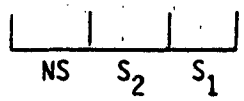
NORMAL TAKEOFF

Visual Cues

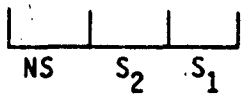
Aircraft Nose:



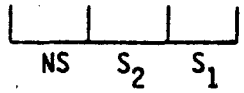
Earth/Sky Horizon:



Runway Centerline:



Ground Mov ent:



ATDs are often intended to provide training for groups of trainees with different skill levels. For example, Combat Crew Training (CCT) versus Continuation Training. Consequently, the quality or type of cues required to train different skill levels may vary. In such instances, it is necessary that the groups of raters used to evaluate the ATD be representative of or knowledgeable about the potential population of trainees. For example, if an ATD is intended for both CCT and continuation training, then one group of raters should be active CCTS instructors, and one group should be continuation training instructors. Although each group could use the same scale as shown above, the two groups might use different standards in judging precisely what constituted "Sufficient" cues. CCTS evaluators should evaluate the sufficiency of cues for CCTS proficiency level training, and the continuation training evaluators should evaluate the sufficiency of cues for continuation training.

Option C

When an ATD is to serve multiple training purposes for different skill level trainees, and if it is not possible to assemble different groups of evaluators representing each skill level, then a scale such as shown below can be used. It should be noted that this scale may be neither ordinal nor unidimensional. The cues represented by categories S_1 and S_2 may differ only in degree or amount, and thus give an ordinal and unidimensional scale, but they may differ in quality or type of cue, and thus give a non-ordinal and non-unidimensional scale.

S_1 = The cues provided by the ATD are Sufficiently similar to the cues required to provide (Continuation) training in the task/subtask.

S_2 = The cues provided by the ATD are Sufficiently similar to the cues provided by the aircraft to provide/allow (CCTS) training in the task/subtask.

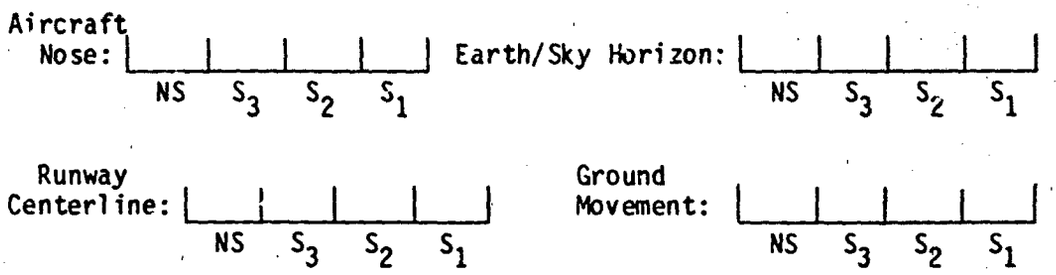
S_3 = The cues provided by the ATD are Sufficient overall to support training, but improvement of selected cues identified by the rating process would improve/enhance training.

NS = The cues provided by the ATD are Not Sufficiently similar to the cues required to train the task/subtask.

Example:

NORMAL TAKEOFF

Visual Cues



Option D

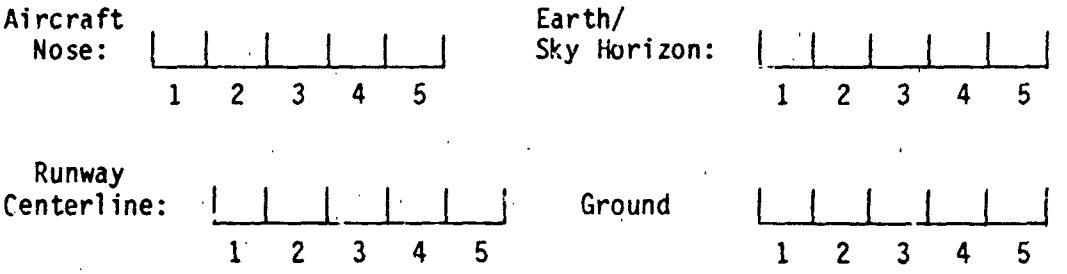
The last option to be described is a rating scale more typical of the type discussed in Section A of this chapter. This type of rating scale can be used for making fidelity assessments, but because evaluation decisions concerning fidelity are usually binary (i.e., a cue or cue group is sufficient to support training, or it is not), its use is usually an unnecessary rating sophistication. In the scale below, even though the degree of fidelity ranges from completely dissimilar to perfectly identical, the fidelity decision ultimately is binary.

Rating Number		Description
5	=	The cues provided by the ATD are identical to the cues provided by the aircraft.
4	=	The cues provided by the ATD are highly similar to the cues provided by the aircraft.
3	=	The cues provided by the ATD are moderately similar to the cues provided by the aircraft.
2	=	The cues provided by the ATD are slightly similar to the cues provided by the aircraft.
1	=	The cues provided by the ATD are not at all similar to the cues provided by the aircraft.
N/A	=	Cue not provided by the ATD.

Example:

NORMAL TAKEOFF

Visual Cues



Even though this scale has five scale points, ranging from no similarity to the aircraft to perfect similarity, a decision still must be made concerning what is sufficient and what is not sufficient. One possible way might be as follows:

	5 =	Identical
Sufficient	4 =	Highly Similar
	3 =	Moderately Similar
	2 =	Slightly Similar
Not Sufficient		
	1 =	No Similarity

Conduct of fidelity rating. The manner in which fidelity ratings are obtained is important. It is recommended that an evaluator rate only cues from one cue group at a time, except in cases where there are only a few cues to rate per cue group. On a task-by-task basis, the cues to be rated should be distributed equally among the raters such that any one rater evaluates all of the cues for a given cue group and that the cues for each group be evaluated by at least two different evaluators. It is best, however, to vary the cue groups to be rated among the evaluators. This reduces boredom and provides an opportunity for each evaluator to sample and assess the adequacy of all the cue groups provided by the ATD.

Rater procedures. Each rater should study the list of task cues he is going to rate before performing the task in the ATD. This will focus the attention of the rater on those cues specifically required to train the task. Each rater then performs the task in the ATD.

Immediately following task completion, he rates the sufficiency of each cue in his assigned cue group(s). The purpose of restricting a rater's attention to a selected set of cues is to reduce the bias of halo effects and to limit to a manageable number the cues that each evaluator has to observe and evaluate at one time.

Use of fidelity rating data. For each cue that is rated as Not Sufficient, regardless of which rating option is used, it must be determined whether or not the cue is critical to training. If it is judged as noncritical, it is simply listed as such, but no further action needs to be taken. However, if the cues are judged to be critical to training, then a service report is generated to initiate corrective action.

The fidelity rating data should be summarized in order to determine if the overall cue groups are sufficient or insufficient to support training for a specified task/subtask. Figure 5-7 shows a sample data summary sheet to be used for this purpose.

A second level of data summary involves estimating the percentage of each task/subtask segment for which ATD training can be provided. For tasks/subtasks where all of the cue groups are rated as Not Sufficient, the percentage of training is assumed to be 0; on the other hand, for tasks/subtasks where all of the cue groups are rated as Sufficient, it is assumed that training can be provided for 100% of that task/subtask. For all other cases, it will be necessary to estimate the relative importance of the deficient cues and to reduce the percentage of training accordingly.

This process obviously should be accomplished after individual ratings have been collected. It also should represent a consensus of the raters who participated in evaluating the task in question. Shown below is an example of what the output of this level of summary would be.

NORMAL TAKEOFF

Subtask/Segment	Estimated Percentage Trainable (EPT)
Segment 1	100%
Segment 2	20%
Segment 3	100%
Segment 4	70%

S = Sufficient NS = Not Sufficient

TASK	SUBTASKS/TASK SEGMENTS	CUE GROUPS	LIST OF ENHANCEMENTS/DEFICIENCIES
NORMAL TAKEOFF	Segment 1	Visual _____ Motion _____ Aural _____ Instrument _____ Others _____ _____	
	Segment 2	Visual _____ Motion _____ Aural _____ Instrument _____ Others _____ _____	
	Segment 3	Visual _____ Motion _____ Aural _____ Instrument _____ Others _____ _____	

Figure 5-7. Sample data summary sheet.

The fidelity data can be further summarized to show the percentage of each task for which ATD training can be provided. First, calculate the proportion of each task represented by each segment or subtask. For example:

NORMAL TAKEOFF

Subtask/Segment	Proportion of total task
Segment 1	.15
Segment 2	.25
Segment 3	.25
Segment 4	.35

Next, multiply the EPT for each task segment/subtask times the proportion of the total task represented by that segment/subtask. From the example above,

Segment EPT		Proportion of task		
100%	x	.15	=	15%
20%	x	.25	=	5%
100%	x	.25	=	25%
70%	x	.35	=	24.5%
		Task EPT	=	69.5%

TRAINING CAPABILITY EVALUATION

The training capability of an ATD refers to an estimate of the extent to which it can provide training for a specified set of tasks. Estimates of training capability are based on (1) an evaluation of the design characteristics (physical fidelity) of the ATD, and (2) the extent to which it provides the cues and response opportunities required to train a set of tasks (psychological fidelity). Such estimates are not usually made in the context of observed ongoing training or with actual trainees. In short, training capability is an estimate of expected ATD training utility, not a measure of demonstrated training ability (as is the case with a training effectiveness evaluation).

Training capability evaluations typically are conducted as part of in-plant I/QOT&E, although they can extend to on-site I/QOT&E efforts. The purpose of a training capability evaluation, therefore, is to provide a direct estimate of the extent or quality of training that can be provided by an ATD and to compare that estimate against its intended operational use. Data collected during a training capability evaluation also can be used to support and guide subsequent OT&E activities, and as an input for the development/refinement of an ATD training syllabus.

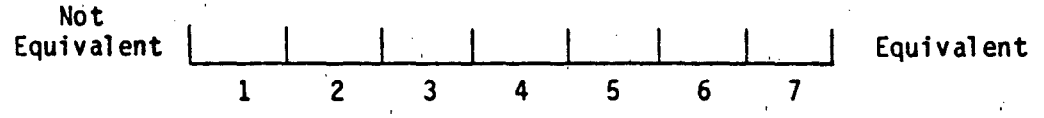
Training Capability Rating Scales

Any number of rating scales and rating scale formats may be designed to assess ATD training capability. Several candidate capability rating scales are provided below:

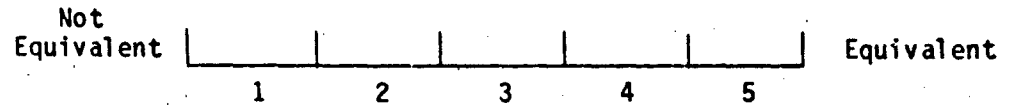
Option A. Option A is a standard seven-point scale of a form often used to rate training capability.

Rating Number		Description
7	=	Training provided by the ATD for this task is equivalent or superior to training provided in the aircraft.
1	=	Training provided by the ATD for this task is in no way similar to training provided in the aircraft; no positive training can be achieved for this task.

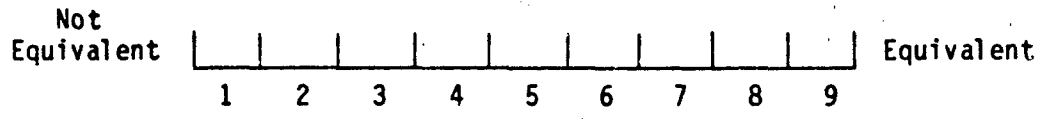
Raters should be instructed to regard the numbers between 1 and 7 as equal intervals or increments from no training similarity to equivalent training.



This basic scale can be expanded or contracted simply by changing the number of intermediate scale points. For example, the seven-point scale above becomes a five-point scale by deleting two intermediate points,



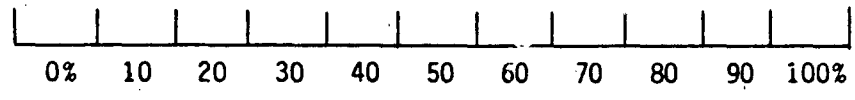
or, if the context warrants, it can be expanded to a nine-point scale by adding two intermediate points,



A difficulty encountered by using scales of this type is that of interpreting the data once it has been collected. For example, if 10 evaluators rated the ability of an ATD to train a takeoff task on a seven-point scale and the average rating turned out to be 5, it still remains to determine what an average rating of 5 means in terms of training c pability. This difficulty can partially be resolved by using a scale such as is shown in Option B below.

Option B. Option B asks the evaluator to estimate the percentage of the training requirement that can be satisfied by ATD training. Several different types of formats are possible. The rater can be asked to check (✓) his estimate with scales such as the following:

(a)



or, the rater can be provided discrete percentages and instructed to circle the appropriate percentage. For example,

(b) 0% 10 20 30 40 50 60 70 80 90 100%

The interval sizes on this type of scale can be increased or decreased as might fit the needs of the test, and as an alternative, the evaluator can be asked simply to estimate the percentage (write in the number) of the training requirement that can be satisfied by ATD training.

The advantage of Option B type scales is that they provide data which are easily analyzed and interpretable relative to the objectives of the evaluation. Percentages represent ratio level data; therefore any measure of central tendency or variability can be used to summarize and analyze data collected with these scales. In addition, because the data are presented as a percentage of task training requirements which can be satisfied in the ATD, they also indirectly indicate what percentage of the requirement must be satisfied in the aircraft.

Option C. Option C requires an a priori estimate of the number of trials or repetitions required to train each task in the aircraft without any prior ATD training. This methodology will take more time to construct, but has the advantage over options A and B in that it provides a more structured basis on which raters can provide training effectiveness estimates.

The basic procedure is described below, including the specific instructions to be given to the evaluators, several examples of the rating technique, and a sample data collection format.

INSTRUCTIONS

1. Below each maneuver/task to be rated is a string of consecutive numbers. The number in the box represents the number of aircraft trials or repetitions required by students, on the average, to achieve proficiency in performing the maneuver/task without any prior training in the ATD.

2. Assume that the same number of trials normally given in the aircraft (as indicated by the number in the box) was given first in the ATD, i.e., prior to training in the aircraft.

3. How many trials/repetitions in the aircraft, following ATD training, would then be required to achieve proficiency? Indicate your estimate of the number of remaining aircraft trials by circling the appropriate number.

The following two examples illustrate the basic procedure.

Example 1:

AILERON ROLL

1 2 (3) 4 5 6 7 8 9 10 11 12 13 [14]

Indicates aircraft trials required following 14 ATD trials

Average aircraft-only trials, and number of ATD trials

Example 1 indicates that student pilots typically require 14 trials in the aircraft, without prior training in the simulator, to

achieve proficiency in performing the Aileron Roll maneuver. However, if 14 trials are first given in the simulator, then only 3 trials are estimated to be required in the aircraft for the student to achieve the same level of proficiency. This indicates that the simulator is an effective training device for teaching the Aileron Roll maneuver.

Example 2:

CLOVERLEAF

1 2 3 4 5 6 7 8 (9) [10]

Indicates aircraft trials following 10 simulator trials.

Average aircraft-only trials, and the number of simulator trials

---

Example 2 indicates that student pilots typically require 10 trials in the aircraft, without prior training in the simulator, to achieve proficiency in performing a Cloverleaf. However, if 10 trials are first given in the ATD, then 9 trials are still required in the aircraft for the student to achieve the same level of proficiency. This indicates that the ATD is not a very effective training device for teaching a Cloverleaf maneuver.

In the examples above, the number of ATD trials is identical to the number of trials normally given in the aircraft. However, it is assumed that ATD training trials for a given maneuver would normally continue until the student is proficient in the ATD. In some cases you may feel that ATD proficiency could be attained with fewer trials, or would require more trials, than indicated. For these cases you can place your estimate of the required number of simulator trials in the box marked "Revised Number of Simulator Trials" that will accompany each maneuver to be rated. Example 3 shows this procedure.

Example 3 indicates, in addition to information already discussed, that the rater feels that proficiency in the simulator could actually be achieved by the typical student in 10 trials rather than 14.

Example 3:

DIVE RECOVERY

1 2 (3) 4 5 6 7 8 9 10 11 12 13 [14]

(3): Indicates aircraft trials required following 14 simulator trials

[14]: Average aircraft-only trials, and number of simulator trials

Revised number of simulator trials . . . [10]

Use this box to change the number of estimated simulator trials required for proficiency in the simulator.

Comparison of Options

Of the three options, A has the advantage of quick construction and relative flexibility across rating situations. It has a major disadvantage of providing data that are hard to interpret and difficult to relate to evaluation objectives. Option B type scales have the same advantage as Option A scales, but they produce data that are easier to interpret and easier to relate to evaluation objectives. Estimates of the percentage of training requirements that can be satisfied by ATD training provide meaningful measures of the ATD's training capability. Such measures can be used to guide and support subsequent OT&E activities; they also can be used as input for the design of ATD training programs.

Option C has the advantages of Option B, but it requires more time to construct such scales, because the number of trials required to achieve proficiency in the aircraft without prior ATD training must be obtained for each task to be rated. These data can either be collected directly prior to the evaluation, or they can be estimated by subject matter experts. (Note, the number of aircraft-only trials does not have to be extremely accurate for the rating procedure to work, because a ratio is calculated from that number. It only has to be a believable number so as not to bias the rater. Clearly, the more accurate the estimate, the better.)

Option C has additional advantages over standard rating scales, including the following:

- The approach used by the estimation technique is easily understood by operational flight instructors because the question is posed in a form they are familiar with, i.e., the number of trials or amount of time required to train a student to proficiency on a specified task.

- It provides a direct, quantitative estimate of ATD training capability that takes into account the efficiency of simulator training.

- It provides a structured format for evaluators to estimate the training capability of an ATD.

- If accurate estimates of the number of aircraft only training trials are used, then the data collected with this scale can be used directly to construct an initial ATD and aircraft training syllabus.

- The data produced by this approach are in a form that provides a meaningful measure of the capability of the device as compared against the aircraft.

Conduct of Training Capability Ratings

Several administrative decisions must be made concerning the actual collection of rating scale data. These include determining when evaluators are to assign ratings, the context in which ratings will be conducted, and the order in which individual tasks are to be rated. Normally, individual tasks to be rated are grouped together according to task area or phase of flight so that they constitute a mission profile or mission scenario. Figure 5-8 shows a sample ATD mission scenario. The basic data collection procedure involves completing each task or maneuver selected for evaluation in the ATD and then estimating the training capability of the ATD for that task. Each task should be rated immediately after it is completed. For the convenience of the evaluators, data collection forms can be constructed to fit a standard knee board, or voice recordings of the evaluator's estimates can be obtained.

Unlike the fidelity assessment, which requires that only two or three evaluators assess each task, all available evaluators should rate the training capability of all selected tasks/maneuvers for their crew station. If at all possible, the sequence of tasks for each mission profile should be varied for different evaluators. In addition, the order in which different mission profiles are experienced should be varied among evaluators. These variations help relieve boredom and control for the possible effects of bias that might result from the order in which evaluators experience the various tasks/maneuvers to be rated.

Initialize on ramp prior to taxiing
Pre-takeoff checks
Taxi
Takeoff (34,000 lbs 10 knots crosswind)
Reinitialize on runway
Takeoff (40,000 lbs 10 knots headwind)
Reinitialize on runway
Takeoff (30,000 lbs 20 knots crosswind)
SID departure to 5,000 feet AGL
FREEZE/ESTIMATE ATD TRAINING CAPABILITY
Reinitialize at 15,000 AFL
Stall Series
1. Power On
   FREEZE/ESTIMATE ATD TRAINING CAPABILITY
2. Power Off
   FREEZE/ESTIMATE ATD TRAINING CAPABILITY
3. Traffic Pattern Stalls
   a. Pitchout
   b. Nose high
   c. Nose low
   d. Flare
   FREEZE/ESTIMATE ATD TRAINING CAPABILITY
Barrel Roll
FREEZE/ESTIMATE ATD TRAINING CAPABILITY
Lazy 8
FREEZE/ESTIMATE ATD TRAINING CAPABILITY
Aileron Roll
FREEZE/ESTIMATE ATD TRAINING CAPABILITY
Loop
FREEZE/ESTIMATE ATD TRAINING CAPABILITY
Cuban 8
FREEZE/ESTIMATE ATD TRAINING CAPABILITY
Immelmann
FREEZE/ESTIMATE ATD TRAINING CAPABILITY
Split - S
FREEZE/ESTIMATE ATD TRAINING CAPABILITY
Reinitialize on 10 mile final
Pitchout
Touch and Go
FREEZE/ESTIMATE ATD TRAINING CAPABILITY
Reinitialize on initial at 1500 feet AGL
IFR Straight-in
FREEZE/ESTIMATE ATD TRAINING CAPABILITY

Figure 5-8. Sample mission profile of tasks to be rated.

Summarizing Training Capability Rating Data

After all the tasks/maneuvers have been rated by all of the evaluators, it will be necessary to summarize and format the collected data. The exact format and display of data normally depend upon the specific evaluation objectives addressed. However, guidelines for handling the data collected with the training capability scales for Options A, B, and C are provided below.

Option A: Standard rating scales. Data collected with these scales can be averaged among evaluators and displayed as average ratings for each task evaluated. They also can be averaged across tasks within a task group or area. However, averaging across tasks within a task area may obscure the meaning of data collected with this type of rating scale. Figure 5-9 shows a sample data summary form for Option A scales.

One problem with averaging ratings is that two extreme ratings, when averaged together, result in a moderate rating. For example, if two evaluators, using a seven-point scale, rate the effectiveness of an ATD for a particular task as 1 and 7, respectively, then the average rating is 4. This problem can be dealt with, in part, by providing measures of variability as well as the average ratings. (See Appendix A of this volume for instructions on how to calculate variability measures.)

Option B. Data collected with these scales can be averaged and displayed as percentages of training requirements which can be satisfied by ATD training. Figure 5-10 shows a sample data summary form for Option B scales. Average percentages and deviations can be shown for each task and across tasks within a task group or task area provided that the evaluated tasks are representative of the tasks contained in a task area. Care should be taken to identify highly variable ratings. Such ratings may indicate a problem with a particular rating procedure or with an individual rater.

Option C. Data collected with this scale can be used to calculate two measures of ATD training capability/effectiveness: Transfer-of-Training Ratios (TRs) and Transfer Effectiveness Ratios (see Chapter 6). A transfer ratio estimate can be calculated from the data collected with Option C as follows:

$$TR = \frac{AC_1 - AC_2}{AC_1}$$

TASK	EVALUATOR	RATINGS	AVERAGE RATING	AVERAGE DEVIATION	COMMENTS/DEVIATION FROM NORMAL PROCEDURES
POPUP ATTACK	A	4			
	B	3			
	C	4			
	D	5			
	E	4			
	F	2			
			3.67	.78	
LOFT ATTACK	A	5			
	B	7			
	C	7			
	D	6			
	E	4			
	F	3			
			5.33	1.33	
LEVEL/ LAYOUT ATTACK	A	2			
	B	1			
	C	1			
	D	3			
	E	2			
	F	2			
			1.83	.56	

Figure 5-9. Sample data reduction/summary sheet for Option A scales.

TASK	EVALUATOR	PERCENTAGE RATING SATISFIED BY ATD	AVERAGE PERCENTAGE	AVERAGE DEVIATION	COMMENTS/DEVIATIONS
POPUP ATTACK	A	60			
	B	40			
	C	60			
	D	70			
	E	40			
	F	40			
			51.67	11.67	
LOFT ATTACK	A	40			
	B	0			
	C	70			
	D	40			
	E	40			
	F	35			
			37.50	13.33	
LEVEL/ LAYOUT ATTACK	A	40			
	B	20			
	C	40			
	D	50			
	E	45			
	F	50			
			40.83	7.50	

Figure 5-10. Summary data reduction/summary sheet for Option B scales.

where:

AC_1 = Number of trials, or amount of time, to achieve proficiency in the aircraft without prior ATD training.

AC_2 = Number of trials, or amount of time, to achieve proficiency after ATD training.

For example, if it normally takes 14 trials in the aircraft without prior ATD training to achieve proficiency in performing an ILS approach, but only 3 trials in the aircraft after ATD training, then the estimated TR for that task would be,

$$TR = \frac{14 - 3}{14} = \frac{11}{14} = .79.$$

This measure indicates that .79, or 79%, of the trials previously assigned to aircraft training can be accomplished in the ATD; it also indicates that at least 21% of the previously assigned trials must still be trained in the aircraft.

The TR estimate is a useful indicator of ATD training capability, but it does not take into account the amount of ATD training required to achieve a given transfer effect. It does not measure the efficiency of ATD training. This shortcoming can be resolved by calculating an estimated Transfer Effectiveness Ratio (TER) from the data collected with Option C scale as follows:

$$TER = \frac{AC_1 - AC_2}{ATD}$$

where:

AC_1 = Number of trials, or amount of time, to achieve proficiency in the aircraft without prior ATD training.

AC_2 = Number of trials, or amount of time, to achieve proficiency in the aircraft after ATD training.

ATD = Number of trials, or amount of time, to achieve proficiency in the ATD.

For example, if it normally takes 14 trials in the aircraft, without prior ATD training, to achieve proficiency in performing an ILS approach, but only 3 trials after 10 trials in the ATD, then the estimated TER for that task would be,

$$TER = \frac{14 - 3}{10} = \frac{11}{10} = 1.10.$$

Estimated transfer and transfer effectiveness ratios can be calculated for each rater, averaged across raters, and displayed for each task. Figure 5-11 shows a sample data summary form for the Option C scale. Transfer ratios and transfer effectiveness ratios also can be averaged across task areas provided that the evaluated tasks are representative of the tasks to be trained in that task area. Figure 5-12 shows a sample format for summarizing the data from individual tasks that can be used either for data collected with the Option C scale or with Option A or B type scales.

TASK	EVALUATOR	TR	TER	AVERAGE TR	AVERAGE DEVIATION	AVERAGE TER	AVERAGE DEVIATION	COMMENTS
POPUP ATTACK	A	.43	.43					
	B	.43	.43					
	C	.69	1.20					
	D	.57	.80					
	E	.43	.60					
	F	.57	.57					
				.52	.09	.67	.22	
LOFT ATTACK	A	.33	.33					
	B	.40	.40					
	C	.40	.50					
	D	.67	.67					
	E	.00	.00					
	F	.40	.38					
				.37	.13	.38	.14	
LEVEL/ LAYOUT ATTACK	A	.48	.48					
	B	.44	.50					
	C	.48	.48					
	D	.40	.40					
	E	.20	.30					
	F	.40	.40					
				.40	.07	.43	.06	

Figure 5-11. Summary data reduction/summary sheet for Option C scales.

TASK	TASK GROUP	TASK AREA	AVERAGE TR	AVERAGE DEVIATION	AVERAGE TER	AVERAGE DEVIATION
Pop Up Attack			.52	.09	.67	.22
Loft/LADD Attack			.37	.13	.38	.14
Level/Laydown Attack			.40	.07	.43	.06
	Attack Maneuvers		.45	.10	.49	.14
Ordnance Visually						
CCIP Mode						
Rockets Using						
CCIP Mode						
Strafe Using						
CCIP Mode						
Ordnance Manually						
Rockets Manually						
	Ordnance Delivery					
Egress						
Controlled Range Departure						
Uncontrolled Range Departure						
	Range Procedures					
Air-to-Surface Tactical Formation						
Locate Known Target						
	Air-to-Surface Combat					
		AIR-TO-SURFACE				

Figure 5-12. Sample data summary sheet.

C. QUESTIONNAIRES

Questionnaires are widely used in the evaluation of aircrew training programs and/or aircrew training devices. They provide an attractive means of collecting information during an ATD OT&E because they are relatively inexpensive to develop and administer, yet allow large quantities of data to be collected quickly. Questionnaires are used for many purposes, varying from collecting background information about a group of prospective ATD evaluators to determining user opinions about a quality of a training program or the training capabilities of an ATD. In addition to the following material, the test director is referred to ARI P-77-1, entitled "Questionnaire Construction Manual," July 1976, for additional guidance in constructing questionnaires for use during ATD evaluation.

In general, two types of questionnaires are used as part of ATD OT&E. The first type is used to gather factual information; the second type is used to collect opinions or subjective estimates. Questionnaires used to collect factual data typically are those which ask the respondent to provide background information concerning his personal/professional experience, including personal demographic information. Questionnaires used to collect opinion data typically are those which ask the respondent to provide information concerning his opinion, or subjective estimate, concerning an event or observation which he has experienced.

The exact form of a questionnaire, its complexity, and its specificity depend upon its purpose. Consequently, some types of questionnaires are relatively simple to develop whereas others are much more complex. For example, the construction of a questionnaire for the purpose of gathering factual, background information is a fairly straightforward process. On the other hand, a questionnaire for the purpose of gathering opinion data is more difficult to construct properly because of the complex nature of opinions themselves.

Regardless of the complexity of a questionnaire, it is important that it be constructed properly. Poorly constructed questionnaires are likely to provide distorted data and may lead to the drawing of erroneous conclusions. It is important, therefore, that the ATD OT&E test director be provided specific guidance concerning the proper procedures to follow in questionnaire construction.

The discussion that follows provides specific guidance on the construction and formatting of questionnaires. Because the construction of background questionnaires is relatively simple technically, the primary emphasis in this chapter is on the construction of opinion questionnaires.

CONSTRUCTION OF QUESTIONNAIRE STATEMENTS

Questionnaire statements consist of questions or statements to which the respondent is asked either to provide an answer or to respond in a manner which is specified by the questionnaire. The way in which a questionnaire statement is posed often affects the kind of answer received. Therefore, it is extremely important when constructing questionnaire statements that they be stated in a way that does not bias or load the answers given. Questionnaire statements must be simple, clear, relevant, and unbiased if responses to them are to provide the test director with data useful to an OT&E effort. The following sections discuss various issues that relate to simplicity, clarify, relevance, and bias aspects of questionnaire construction.

Structured vs. Open-ended Questions

Questions may require answers that are structured, open, or a combination of both. A question that requires a structured answer is one that limits the potential answer to choices among several specified alternatives. A questionnaire statement that requires an open-ended response is one that leaves the content and structure of the answer entirely up to the respondent.

An example of a questionnaire statement with structured answers might be as follows:

Training in ATDs should be conducted by:

_____ Fellow Pilots

_____ Instructor Pilots

_____ NCOs/Airmen

_____ Civilians

The same statement in an open-ended answer format might read:

Who do you feel should conduct ATD training? Why?

Structured statements. Structured statements limit the number and breadth of possible responses, but they have the advantage of making the use of questionnaires less time consuming and less ambiguous than open-ended statements. Because the time to respond to structured statements typically is shorter than that required for open-ended statements, it is possible to obtain more responses in a given period of time than is usually possible with open-ended statements. Thus, a structured questionnaire usually provides more information.

Open-ended statements. Open-ended statements have the advantage of giving the respondent a chance to present views which might not have occurred to the test director, but such statements have a number of drawbacks: (1) The answers can be time consuming, both for the person answering the question and for the person who must interpret and categorize those answers; (2) the answers may be difficult to interpret or categorize; (3) the answers may not be relevant to the concerns of the questionnaire; (4) only a limited number of open-ended questions may be asked because of the time required to provide answers to open-ended questions; and (5) because of time constraints, motivation, or a simple inability to express oneself, open-ended statements may be left blank or may contain insufficient information to allow categorization of the answer.

Ambiguous Statements

An ambiguous statement is one that can be interpreted in more than one way. It is difficult to write short, clear statements that will elicit only the information of real interest. One problem is differences in interpretation among respondents, e.g., one person's "often" could be another person's "seldom." Another potential problem is that different words may imply different meanings depending on context, even though they are related. For example, the words "generous" and "extravagant" may both refer to the freedom with which one spends money. However, "generous" conveys a positive value, whereas "extravagance" conveys a negative value.

Double Negatives

Double negatives should be avoided in constructing questionnaire statements. For example, consider a questionnaire item such as:

Are you against not restricting the use of ATDs?

_____ 1. Yes

_____ 2. No

A "yes" or "no" response to such a statement could mean almost anything. People tend to respond "yes" to those things with which they agree and "no" to those things with which they disagree. Double negatives only serve to confuse the issue."

Double-barreled Statements

It is important to address only one issue in a single questionnaire statement. To do otherwise makes it difficult or impossible to determine to which issue the subject has responded.

Consider, for example, the following statement:

> Simulation can be used to enhance the readiness of the Air Force; therefore, we should decrease flying hour allocations.

If this statement evokes an "agree" response, it could be because the respondent believes that simulation enhances readiness, because he feels that we should decrease flying hour allocations, or perhaps both. How does a subject respond if he agrees with one part of the statement, but disagrees with the other part? For example, one might be favorably inclined toward simulation, but still feel that it would be unwise to decrease flying time.

The best way to handle double-barreled statements is to rewrite them into two separate statements. For example, the above statement can be rewritten as:

_____ 1. Simulation can be used to enhance the readiness of the Air Force.

_____ 2. Flying hour allocations should be decreased.

Social Desirability Effect

If a respondent feels that either positive or negative consequences may be associated with his expressing or holding a particular viewpoint, he will likely respond in accord with those consequences. Such consequences may be of several sorts. One is adverse peer pressure that can result from revealing personal, unpopular, or controversial viewpoints on various topics. Another involves adverse outcomes that can result from holding a viewpoint contrary to one's superiors. In addition, almost everyone wants to feel that his opinions are acceptable to others.

Consider the following statement.

> Please give a self-appraisal of your overall pilot skills before entry into Training Program X.

_____ Excellent

_____ Good

_____ Average

_____ Fair

_____ Poor

The wording of this statement may cause the respondent to mark one of the three middle choices instead of either end point, in an attempt to avoid making an extreme statement about himself. Further, it should be no surprise to find that most subjects responding to this statement would rate their skills as "good." It is likely that few pilots believe they have only average or below average skills, and even fewer would be willing to admit it publicly.

Here is another illustration of the social desirability problem in a questionnaire statement that might be given an instructor pilot:

> Do you feel you effectively communicate course objectives to your students?
>
> _____ Always
>
> _____ Often
>
> _____ Sometimes
>
> _____ Rarely
>
> _____ Never

"Always" sounds boastful, whereas the last three alternatives ask the instructor to admit that he is inadequate. Thus, the instructor's response might be heavily influenced by how he thinks his peers or superiors might perceive his response to this item.

"Loading" the Statement

Loaded or leading statements may convey something about the test director's attitudes or opinions (or those of superiors, the Air Force, etc.) which may bias the answer given by the respondent. Such items are sometimes labeled as "loaded." At best, they are leading and likely induce bias. In other words, they are not neutral, and they may suggest what the subject's response should be, or indicate the test director's (or others') viewpoint on the item content.

For example, consider the following two statements:

1. Most pilots agree that ATDs provide effective training; do you agree, or disagree?

2. What do you see as the primary benefits of Training Program X?

The first of these statements is clearly loaded because it conveys a conclusion (pilot agreement) that ATDs provide effective

training, instead of simply asking the respondent to indicate how he views the effectiveness of most ATDs currently in use. This initial positive statement about ATD effectiveness may, in turn, influence the manner in which the subject subsequently responds to the statement. He may indicate a positive attitude towards ATD effectiveness because he believes the test director feels that way or because he considers himself like "most pilots." The second statement is also loaded or biased because it implies that Training Program X is, indeed, beneficial, and it is simply the task of the respondent to provide a listing of those that are primary among the benefits. There is no way for the respondent to indicate drawbacks associated with Training Program X. However, pairing this item with another that allows time to list "primary drawbacks" would eliminate the loading problem.

Questionnaire statements also may be loaded through limiting the range of available answers. Consider the following example:

> Rate the effectiveness of Training Program X by circling your choice.
>
> ____ Excellent
>
> ____ Very Good
>
> ____ Good
>
> ____ Poor

Three of the four response alternatives represent favorable evaluations of the program. As a consequence, one might not be surprised if responses to this item indicate some type of positive reaction to Training Program X.

A final example illustrates how a leading question may be combined with limited response alternatives to load an item and potentially bias responses to it.

> Have the positive aspects of the training program outweighed the negative aspects?
>
> ____ Yes
>
> ____ No

Both "yes" and "no" responses might be interpreted as indicating that there were, in fact, positive aspects to the training program. The positive orientation of the question and the limitation of the two response alternatives increases the probability that subjects

will respond "yes," especially if they are not sure of their answer. Including an "Undecided" category would reduce this probability, but a preferred solution would be to reword the statement and have subjects rate the program. One possibility might be as follows:

> Indicate by checkmark your feeling as to the overall effect of the training program on pilot skills.

Positive Neutral Negative

It also should be pointed out that specific words or phrases within an overall questionnaire item may be considered loaded, leading, or biased. A loaded word or phrase is one which evokes strong, emotional feelings of a predetermined nature, or one which automatically elicits an automatic feeling of approval or disapproval.

FORMAT OF THE QUESTIONNAIRE

An often overlooked but important issue is that of the format and appearance of the questionnaire. Format and appearance can be important issues in that they potentially affect the mental set of the respondent and, therefore, can influence the reliability or integrity of the questionnaire itself.

General Format Suggestions

A number of suggestions designed to improve the appearance and acceptability of a questionnaire are suggested below. Although these suggestions must be interpreted in light of the particular purposes of the questionnaire and the responses to it, they provide worthwhile general guidance.

1. The introduction should be as brief as possible, but long enough to include all of the relevant information, including instructions. Such useful information should include:

   - The topic of the questionnaire
   - The source or origin of the study
   - The use that will be made of the results
   - The availability of results to the respondent

- Identification of the authorizing agency (if appropriate), a point of contact, and a phone number

2. Directions should be set in capital letters or in boxes, or both.

3. Care should be taken in the arrangement of items on a page. Proper arrangement can make the questionnaire more effective. For example:

   - Leave ample room; do not crowd the page. Too much typing on a page may make it difficult to read.

   - Group all the answers clearly. Use of a format such as this,

     1.___ 4.___

     2.___ 5.___

     3.___

     runs the risk of the fourth and fifth categories being overlooked.

   - Never let the list of alternatives for a single statement continue on to another page.

4. Each of the items on the questionnaire should be numbered; this will save time and increase the accuracy of data analysis.

5. The questionnaire should be proofread carefully for content, textual, and spelling errors (at least two persons should proof it).

6. The questionnaire should be neat. People are more likely to attend seriously to a "professional" appearing questionnaire than to one that is sloppy or one that looks ill-prepared.

Ordering of Questions

The general principle governing the order in which questions are asked in a questionnaire is that the respondent should be led through the area to be covered in as clear and logical a manner as possible. Questions should follow in a natural sequence, and transitions from one subject area to another should be made as easy for the respondent as possible. Questionnaire items should be sequenced so as to form

meaningful beginning, middle, and end segments. The early questions should serve to introduce the questionnaire to the respondent, to engage his interest, and to assure him that subsequent questions will not be too taxing or embarrassing. They should, therefore, be representative of the subsequent questions to be asked, and they should lead logically into the main body of the questionnaire. Finally, the last items included on the questionnaire should be sequenced so as to provide a logical ending point and closure for the topics covered.

The overall objective of correct ordering is to allow the respondent to be led through a subject area without breaks in his train of thought. This may also be accomplished by the insertion of appropriate explanatory bridging material. However, proper item ordering is preferable because excessive bridging material is likely to confuse the respondent and will add to the overall length of the questionnaire.

Pretesting the Questionnaire

The effectiveness of a questionnaire will be enhanced by pretesting. Ideally, the questionnaire should be pretested on a sample of people taken from the intended audience. However, if time or available resources preclude such pretesting, people as similar to the intended audience as possible should read over the entire questionnaire and comment on all parts of it (introduction, instructions, questions/statements, format, appearance). Emphasize that you want advice, not endorsement, and question your sample of respondents about each answer choice. Be prepared to take advice and constructive criticism. You may find that a respondent's understanding of a statement or an answer is quite different from what you intended.

It is important that the test director allow enough time to construct and pretest the questionnaire. Shortcuts and questionnaires that have not been reviewed often lead to useless results and considerable wasted effort. Worst of all, they may lead to erroneous or misleading information and could seriously compromise the integrity of an OT&E effort.

CHAPTER 6

TRANSFER OF TRAINING EVALUATION METHODS

INTRODUCTION

This chapter covers the design, conduct, and interpretation of transfer-of-training studies as a method for assessing the operational effectiveness of an ATD.

Definition of Transfer of Training

Transfer of training (TOT) refers to the effects of training in one task or situation upon performance in another task or situation. For example, the skills or information acquired in a classroom, or ATD, might make subsequent learning in the aircraft easier than if those skills or information had *not* been acquired. There are three ways in which we can characterize the manner in which learning in one situation can affect performance in a subsequent situation:

> Positive transfer occurs when training received on one task facilitates subsequent learning or performance of a second task. For example, if the amount of time, number of trials, or number of errors incurred during training in an aircraft are reduced by virtue of having had prior training in an ATD, then the training in the ATD can be said to have transferred positively to learning in the aircraft. Similarly, if learning to drive a car aids in learning to drive a truck, positive transfer occurs.
>
> Negative transfer occurs when training received on one task hinders learning or performance of a second task. For example, if the amount of time, number of trials, or number of errors incurred during training in an aircraft are increased by prior learning or experience in an ATD, then the training in the ATD can be said to have transferred negatively to learning in the aircraft.
>
> Zero transfer occurs when training on one task has no effect on the subsequent learning or performance of a second task. For example, if prior training in an ATD has no affect on subsequent learning in the aircraft, then training in the ATD has failed to transfer, either positively or negatively, to the aircraft.

Transfer-of-training (TOT) studies are used to measure the direction and amount of transfer between two tasks or situations. A TOT study, thus, can be used to evaluate the training effectiveness of an ATD. A TOT evaluation of an ATD, in its simplest form, involves comparing the performance of two groups of trainees: an "ATD Group" that receives some amount of ATD training prior to training in the aircraft; and a "Control Group" that does not receive ATD training prior to training in the aircraft. The effectiveness of an ATD can be assessed by comparing the number of trials or amount of time required by the ATD Group to learn a task in the aircraft against the number of trials or amount of time required by the Control Group to learn the same task in the aircraft. (The Transfer Ratio discussed in Chapter 6 is one means of making such comparisons.)

Differences in trials/time to attain criterion performance between the two groups can be attributed to the use of the ATD, provided that a number of conditions are met. These conditions involve assuring that factors that may affect performance, such as prior flying experience, quality of instruction, conditions under which performance is evaluated, etc., are kept as similar as possible (i.e., controlled) for both groups. Otherwise, it would be difficult to determine whether or not a difference in performance between the groups was due to ATD training, or to those other factors. Similarly, real differences in performance may be obscured if factors that affect performance are not kept constant for both groups.

Advantages of TOT Method

A TOT study, as a method for evaluating the training effectiveness of an ATD, has several distinct advantages over analytical evaluation methods. First, the design of a TOT evaluation is founded upon the basic concept underlying the use of ATDs, i.e., that training received in an ATD will transfer to the aircraft. A TOT study seeks to determine the nature (positive or negative) and extent of that transfer.

Second, a TOT study, unlike an analytical evaluation, allows a direct, quantitative determination of the effectiveness of ATD training compared against the effectiveness of training without the ATD. A properly conducted TOT study, therefore, can provide a direct measure of the contribution of an ATD within the context of a specific, ongoing training program.

Third, a TOT study can provide data for determining the effects of trainee characteristics (e.g., experience, prior performance, sex, etc.) upon subsequent performance. It can provide, therefore, the basis for developing different training strategies for groups of trainees that differ markedly in the extent to which they may benefit from ATD training.

Limitations of TOT Method

The TOT method of assessing the effectiveness of ATDs also has some limitations. First, it requires the ability to equate and control those factors that may affect performance between the ATD group(s) and the control group. This requires extensive support and cooperation from the using command as well as the individuals involved in the study (e.g., instructors, evaluators, schedulers, etc.). Second, a TOT study usually requires that "normal" training practices be disrupted, or at the very least, that two separate tracks of training be provided. Third, TOT studies typically require fairly extensive personnel and materiel resources, a good deal of time and effort to plan, and, frequently, a relatively long time to complete. Fourth, the actual conduct of a TOT must be monitored and supervised continuously to ensure that specified instructions, conditions, and procedures are followed. Finally, a TOT study cannot be used to assess the effectiveness of an ATD for training tasks that cannot be taught in the aircraft (e.g., some emergency procedures).

Purpose of this Chapter

The TOT method is a useful method of evaluating ATD training effectiveness, if its technical requirements can be accommodated sufficiently. This chapter describes in detail the TOT method of evaluating an ATD. It is subdivided into two sections, A and B. The first section offers a general discussion of the various activities that are involved in planning, conducting, and interpreting the data for a TOT evaluation of an ATD. The discussion is somewhat tutorial. It is intended to provide the ATD OT&E test director with an understanding of the method, of what is entailed in conducting a TOT study, and of which factors might compromise his ability to complete a successful TOT. In the event that he intends to conduct a TOT, the second section then describes four specific TOT study designs and provides guidelines for the proper application of each design.

A. GENERAL TOT EVALUATION PROCEDURES

TOT evaluation of an ATD involves three phases: a Planning Phase which involves designing the TOT evaluation and planning for its implementation; a Test Execution Phase which involves the collection of performance data during ATD and aircraft training; and a Post-test Phase which involves the reduction, analysis, interpretation, and reporting of the performance data collected during the evaluation. A detailed discussion of each of these phases and its component activities is provided below.

PLANNING PHASE

Probably the most important phase of any TOT study is the Planning Phase. In it, the design for the complete TOT study and detailed plans for its implementation must be carefully worked out and documented. Unless careful and complete planning is given to each of the activities involved, it is extremely unlikely that a meaningful TOT evaluation of ATD training effectiveness can be carried out.

Planning Phase activities include: estimation and coordination of all required resources; specification of OT&E test objectives; selection of tasks to be evaluated; development of performance measures, standards, and criteria for each selected task; development of a program of instruction for each group contained in the evaluation; determination of data collection procedures and format; selection of trainees and their assignment to groups; selection and assignment of instructors/evaluators; instructor/evaluator training; pretesting data collection forms and procedures; and development of a TOT Study Plan.

Resource Identification and Coordination

Identifying and coordinating the resources required to conduct a TOT evaluation is an activity that normally will continue throughout the Planning Phase. The initial detailed resource requirements list prepared by the test director likely will be incomplete; consequently, a basic strategy or framework for progressively updating resource requirements should be formulated by the test director early in the planning. The basic resource requirements list can be updated and refined after each new input is received and/or after each planning or design decision is made.

Some of the resources needed to conduct a TOT evaluation include:

- Technical support. The complexities of planning and conducting a TOT evaluation require the assistance of a qualified

behavioral scientist. This assistance is essential in selecting an appropriate TOT design, in determining adequate measures of performance, and in determining data reduction, appropriate analysis, and interpretation procedures.

- Trainee pool. The backgrounds of trainees needed to participate in the evaluation may differ. Trainees with backgrounds appropriate to the evaluation objectives of the ATD OT&E in question must be available to participate. The total number of trainees required for the evaluation as well as the number of trainees required for each different group that will participate must be specified.

- ATD availability. A TOT evaluation that may require the ATD be available for evaluation purposes for large segments of each day over a period of time ranging from several months to over a year.

- Aircraft availability. All trainee groups in a TOT study receive training in the operational aircraft. The test director must assure that adequate flying hours will be available to support the OT&E.

- Instructor/evaluator availability. Sufficient instructors and/or evaluators must be available to conduct the ATD training and gather performance data in both the ATD and in the aircraft. Backup and replacement personnel should be identified as early as possible.

- Data analysis support. Analysis of the data gathered during a TOT evaluation may require a sophisticated calculator or a computer. The need for such devices should be anticipated well in advance and their availability arranged.

- Key personnel. All key personnel should be identified and coordinated with throughout the evaluation. Key personnel include schedulers, personnel assigned to monitor/manage the test implementation phase, personnel to train instructors/ evaluators, etc.

- Contingency plans. Contingency plans should be developed to cover those chance events that might disrupt the planned evaluation. For example, absent or transferred instructors or evaluators, ATD equipment failure, changes in aircraft scheduling, flying hour availability, bad weather, etc.

Identifying required resources for a TOT evaluation is one thing; coordinating for their use in a TOT study is a different matter. The

controlling organization or agency responsible for each required resource should be identified in the test plan outline (TPO) as early as possible and updated as appropriate. The use of all resources must be coordinated with the controlling organization and agreements to use resources (both materiel and personnel) formalized in memoranda of agreement (MOA). Changes in resource requirements or schedule changes should be coordinated and reflected in updated MOAs. If the required resources cannot be secured, then an alternative evaluation approach (e.g., rating scale evaluation) should be considered, or the TOT evaluation should be postponed or aborted until the required resources can be secured. A TOT evaluation can be extremely time consuming and expensive; it should not, therefore, be attempted without sufficient resources to support its successful implementation.

Program schedule. One of the most overlooked elements during the planning phase is the need to schedule adequate time to complete the evaluation. Overly optimistic schedules usually do not consider the real world problems of testing, the late delivery of test equipment, mission failures, instrumentation problems, bad weather, availability of subjects, etc.

A program schedule, therefore, should be established as a first priority in planning the evaluation. Graphic schedule and milestone charts will help the test director see the critical portions of the planned test. The schedule should show when specific resources will be required. For example, the schedule should show the anticipated number of students available for the study and the required number of instructors and evaluators (both for the ATD and aircraft) for each week of the planned evaluation. Any special events that might affect the successful implementation of the TOT evaluation also should be included.

A carefully constructed schedule can help identify milestones for specific program events, unknown factors in the schedule, and possible schedule conflicts and critical path events. It also can aid in determining realistic amounts of time for accomplishing the evaluation and for determining the "lead time" required for some activities (e.g., instructor/evaluator training). The experience of the test director, test team members, and that of other knowledgeable personnel can be used to guide the development of a realistic schedule that will mesh with the overall program goals and evaluation constraints.

Specification of TOT Evaluation Objectives

The critical question addressed by a TOT evaluation is the extent to which ATD training transfers to subsequent performance in the aircraft. Although this question, as stated, is straightforward, it does not provide the test director with the specific guidance necessary to

plan and conduct a particular TOT evaluation. This guidance or lack thereof, will be reflected in both general and specific test objectives.

Test objectives should be as clear and specific as possible, and they should be stated in a manner that both suggests what should be measured and the format of the data to be collected. This will aid in assessing whether or not, and to what degree, OT&E test objectives are met.

General test objectives can be formulated from the questions which initially give rise to the evaluation. Some important sources include the Statement of Intended Operational Employment for the ATD, any identified Critical Questions, interactions with representatives of the Using Command, and results from previous OT&E activities. Specific test objectives then can be formulated from those general objectives. The procedure for deriving evaluation objectives is described in Chapter 5. Figure 6-1 provides additional illustration of the process of how general test objectives can be successively transformed into more specific objectives.

The general test objectives derived from the above sources should be reviewed to ensure that they meet the following criteria:

- The general test objective(s) should be stated so that their fulfillment will provide answers to the original questions.
- The general test objectives should be operationally meaningful in the sense that their fulfillment will provide useful information relative to the original questions asked.
- The general test objectives should be stated clearly.
- The general test objectives should be testable within the operational context in which the evaluation will be conducted.

These same criteria must be met when the specific test objectives are derived from the general test objectives. A general test objective usually is comprised of a set of smaller, functionally related parts which can be identified and separated into a group of specific test objectives. The actual steps involved in doing this are:

1. Analyze the purpose of the general test objective and identify the smaller, functionally related parts.
2. Subdivide the general objective into groups of smaller elements, e.g., tasks.

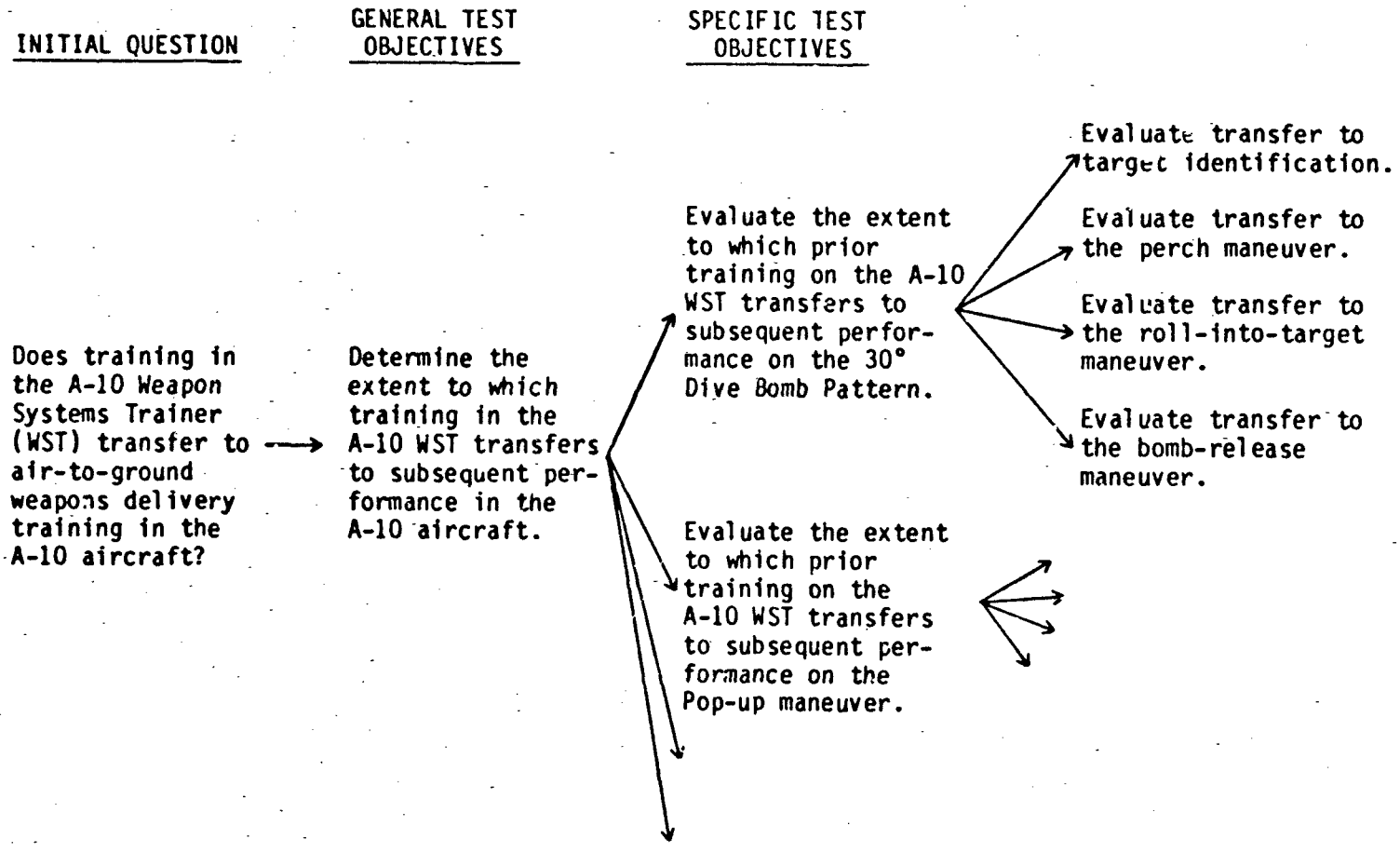


Figure 6-1. Development of test objectives from general to specific.

3. Analyze this new group of more specifically stated objectives and continue to subdivide them until objectives are obtained that can be addressed by a single performance.

The development of specific test objectives should be done in conjunction with selecting specific tasks to be evaluated. The topic of task selection is discussed below.

Task Selection

It usually will be impractical, during a TOT evaluation, to measure trainee performance on all the tasks normally trained by the ATD because of limitations on time and resources. Therefore, it becomes an important planning activity to select the tasks to be included in the evaluation. The first step in the process of task selection is to list all of the tasks potentially trainable in the device; individual tasks should be grouped according to task area (e.g., instruments, air-to-surface, etc.). Specific tasks then can be selected from this listing in accord with the procedure described in Chapter 5.

Selection/Development of Performance Measures

The purpose of a transfer of training study is to determine the extent to which ATD training affects subsequent performance in the aircraft. A TOT study typically involves comparing the performance in the aircraft of trainees who received prior training in the ATD against the aircraft performance of trainees who did not receive prior ATD training. The success of a TOT study, therefore, is determined, in large part, by the "quality" of the measures of trainee performance that are used.

For these reasons, the task in planning a TOT study of selecting from among existing measures of trainee performance those measures appropriate for TOT and/or to develop additional performance measures is an extremely critical task. In either event, the overall goal is to establish a set of performance measures that accurately and objectively reflect trainee performance and that are sensitive to differences in performance. Several criteria should be used to guide the selection/development of performance measures, i.e., measure validity, reliability, sensitivity, and acceptability.

Validity. The validity of a measure is the degree to which it measures what it is supposed to measure. Validity is the single most important aspect of a measure; if a measure does not have some degree of validity, its other characteristics are irrelevant. Although simple to state, the achievement, and satisfactory demonstration of performance measurement validity often is hard to accomplish. Refer to the statistical appendix of this volume for further discussion of validity.

One indication of performance measure validity is the ability of the measure to reflect changes in performance that result from repeated exposure to, or experience in, the training environment (e.g., as reflected by number of training trials or time in training). A measure which demonstrates behavior change over time as a function of training given is reflective of learning that has occurred; such a measure has a higher probability of validly measuring trainee performance than one that consistently fails to reflect such change.

Reliability. The reliability of a measure is the extent to which it remains constant, i.e., provides consistent information, over repeated applications. There are two aspects of measure reliability that are relevant to a TOT evaluation. The first aspect concerns inter-rater reliability, i.e., the extent to which a measure used by two or more raters, who observe the same performance, yields consistent results. The second aspect involves the extent to which a measure used to assess identical performances, occurring on different occasions, yields consistent results. Both aspects of measure reliability are important; both can be increased by providing objective measurement standards and criteria for each task to be evaluated, and by careful training of raters, observers, and others who are involved in data gathering. Refer to the statistical appendix of this volume for further discussion of reliability.

Sensitivity. The sensitivity of a measure is its ability to discriminate between different levels of trainee performance. An insensitive measure obscures real differences in performance. It is important, therefore, to develop/select measures that are sufficient to detect differences that may exist between the performance of ATD and Control group trainees.

Acceptability. A final consideration of measurement criteria, and a most important one, is its acceptance by the user. Even if a performance measure meets all the requirements of validity, reliability, and sensitivity, it will be of little use if instructors refuse to use it, or worse, refuse to use it appropriately. Assessing trainee performance for the purpose of supporting a TOT evaluation likely will increase IP/evaluator workload; a successful evaluation depends critically upon their support. This must be kept in mind clearly when developing performance measures. The developed measures must be credible to the IP/evaluator community, and they must not require a greater increase in workload than that community will reasonably tolerate. The development of performance measures, therefore, should be coordinated with the individuals who will be expected to use them.

Several other considerations should be kept in mind when selecting performance measures to be employed in a TOT evaluation. These include:

- Measures should be chosen that are relevant to both the objectives of the particular simulator training program and to the objectives of the ATD evaluation.

- Measures should be selected that will provide management with information that will allow them to make decisions associated with the TOT evaluation.

- Measures selected for use must be analyzable within the capabilities and resources available to the TOT evaluation team.

- Measures must exist in both the ATD and the aircraft.

- Measures must be feasible to implement within the context of the evaluation.

- Measures must be safe to implement and use during the evaluation.

- Measures must be cost effective to implement during the evaluation.

- Measures should be quantifiable to be useful for the purposes of most TOT evaluations. Thus, measures such as altitude, heading, climb, deviation scores, etc. are preferred over subjective ratings such as "Good vs. Bad" or "Pass vs. Fail."

Within the context of the criteria for selecting/developing performance measures discussed above, the test director must decide what aspects of trainee performance will be measured and evaluated, when performance will be measured, and how performance will be measured. Each of these aspects of performance measurement is discussed below.

What aspects of trainee performance should be measured. The current Air Force grading system typically evaluates performance on a four-point scale:

1 - Unsatisfactory or unable to accomplish

2 - Fair

3 - Good

4 - Excellent

The grades obtained from this type of scale during training may be adequate for student management purposes on a day-to-day basis, but they usually are not sufficiently discriminating to be used in a TOT

study. That is, the sensitivity of such scores is probably insufficient to detect real differences in the performance of students who received ATD training and those who did not. In addition, grades often are assigned according to a "normal" reference point, i.e., the performance expected from the average student, rather than according to an absolute standard. Consequently, a "good" grade may be assigned early in training, not because the performance itself met the criteria or standards desired, but instead because performance was good compared to the performance of the average student at that point in training.

One of the first decisions to be made relative to which aspects of trainee performance should be measured involves determining how specific the performance measures should be. The specificity of performance measures can range from a single performance score per maneuver to multiple, independent measures of all possible flight parameters.

One approach, which is recommended here, involves identifying the critical parameters that comprise a task or maneuver and determining acceptable standards or values for those parameters. Information then can be collected from scores which reflect student deviations from prescribed maneuver parameters and procedures.

Deviations from these established tolerances can be characterized as "errors," in the sense that acceptable performance is defined as being within specified tolerances, and unacceptable performance (i.e., error) as being outside the tolerance limits. These error scores have both content and face validity, insofar as they are based on those task elements identified as critical and defined as required for acceptable performance. For example, altitude at the end of final turn in a landing approach is critical. If "good" altitude control is established by consensus as ranging from 735 to 425 feet at rollout, then any deviation above or below becomes an error, an error of significance in terms of defined task performance. The sum of such errors for a maneuver thus becomes an index of the quality of overall maneuver performance.

When such a specification of required performance does not exist or cannot be obtained, error scores can be derived by establishing arbitrary optimal values and tolerances for meaningful aircraft control parameters such as heading, altitude, airspeed, and angle of attack. By defining smaller tolerance ranges on either side of these optimal values, estimates of error severity can be obtained by assigning 0 to the optimal, 1 to the next acceptable range, and 2 to the next, and so on for all ranges of interest. Figure 6-2 shows an example of how critical performance parameters might be scored for a takeoff maneuver; Figure 6-3 provides a similar example for a barrel roll maneuver.

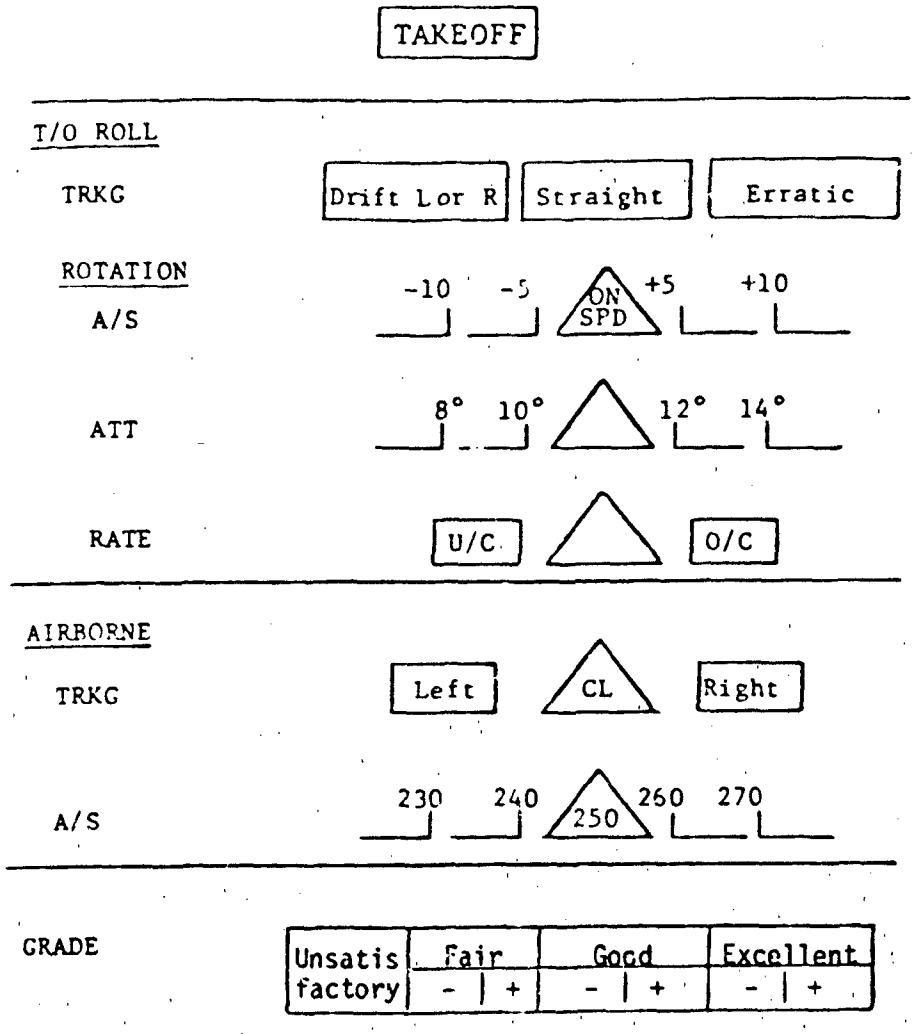


Figure 6-2. Sample performance measurement form for the takeoff maneuver.

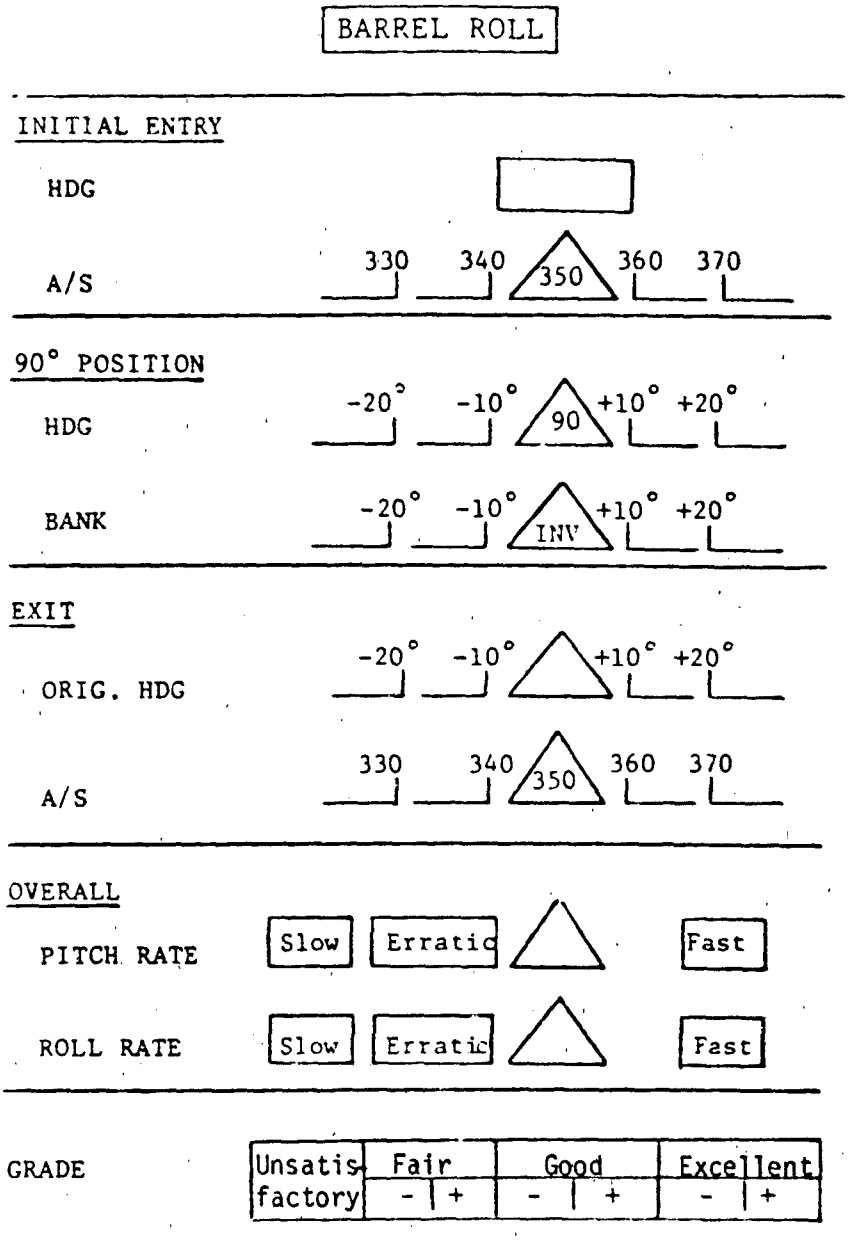


Figure 6-3. Sample performance measurement form for the barrel roll maneuver.

In addition to scoring critical parameters of performance, an overall maneuver grade/score can also be assigned. Although there are numerous ways in which overall grades could be assigned, one form is that shown in Figures 6-2 and 6-3. That is, the standard four-point grading scale used by the Air Force has been expanded to a seven-point scale by adding "pluses" and "minuses" to the grade categories of "Fair," "Good," and "Excellent." This potentially increases the sensitivity of the grading scale to distinguish smaller differences in trainee performance, while maintaining the basic integrity of the original grading scale. The revised seven-point scale helps accommodate the need for a relatively sensitive measure of performance required by a TOT study. It also is a scale with which Air Force instructors/evaluators are familiar and, hence, one they will likely accept.

This approach to measuring student performance has been used successfully by many previous investigators. It is based on the ability of experienced and trained instructor pilots to observe and record student pilot performance in terms of adherence to, or deviations from, critical aircraft performance parameters at specified points throughout a maneuver.

The performance measures are developed by analyzing each selected maneuver to identify critical performance parameter tolerances and the key maneuver points. Information concerning these matters can be developed through consulting sources such as IPs, standardization/evaluation personnel, and published standards. Then for each of the key points, parameter "scales" or boundaries are developed and arranged in an initial performance recording form.

These initial estimates/values for each maneuver to be measured then can be reproduced in several copies and flight tested by IPs to verify that the parameters are both observable and recordable in real-time flight. Any required changes then can be made and the parameters retested.

The development of performance parameters and standards, as described above, may not always be possible due to limited resources or operational constraints. It may, therefore, be necessary to rely solely upon overall maneuver grades. In such cases it is extremely important that overall maneuver grades be sensitive to differences in performance and that "proficiency" standards for maneuver grades be stated as clearly and objectively as possible.

When to measure performance. Trainee performance must be measured both for ATD and aircraft training. Ideally, performance should be measured on each training trial. Trial-by-trial measurement of performance is important for several reasons. First, comparing performance on the last trial of training against performance on the first aircraft trial indicates the amount of performance improvement,

or gain, that results from training. Second, showing the rate of performance improvement over successive training trials allows decisions to be made about when ATD training should terminate, i.e., it shows when the expected increment in performance does not justify the additional training time. Finally, differences in ATD and Control Group performance may be reflected in the rate at which learning takes place as opposed to the actual number of trials required to achieve a specified proficiency level.

Performance measures (as was the case with rating scale measures) also should be recorded as soon after they occur as possible to assure accurate measurement. In some cases (e.g., safety reasons), an IP may not be able to record trainee performance on a maneuver until after the maneuver is completed, or even until the training sortie is over.

Because performance data will be collected in a training environment, there may be times when performance measures should not be collected. For example, a trainee practicing a complex task in either the ATD or the aircraft often will receive instruction, coaching, or advice from his instructor. The trainee's performance for that trial should not be measured for evaluation purposes, because his performance is being directly affected by the input he is receiving from the instructor. A trial in which the trainee receives instruction is referred to as an instructional trial. Performance should not be measured during instructional trials.

Performance data should be collected on trials or at times when the instructor does not influence trainee performance. These trials are referred to as measurement trials. During measurement trials the trainee should be allowed to perform/complete the maneuver uninterrupted, unless safety factors make it impractical to do so. A training session will usually consist of some combination of instructional and measurement trials. Before data collection begins, the test director should determine with the IPs involved in the training program precisely when trainee performance will be measured. Performance measurement may occur on every trial, on every third trial, or whatever is best suited to the evaluation objectives.

How performance should be measured. There are two general issues to be discussed within the context of how trainee performance should be measured. The first concerns the "mechanics" of how performance should be measured; the second concerns issues that arise depending on whether or not training is continued until a specified criterion is reached.

Specific determinations about how to measure trainee performance are driven by considerations of what can be measured. In addition, the objectivity and reliability of performance measurement is a matter of the methods and procedures used in acquiring and recording the

performance data of interest. Good results usually are obtained when the measurement data are recorded automatically. However, in most cases, automatic performance recording equipment is not available, though some ATDs do have automatic performance measurement capability. If this ATD capability is used, care must be taken to ensure that performance criteria and measures used in the ATD are the same as those used subsequently in the aircraft. Good results also can be obtained when performance data are recorded manually, provided that data collection is accomplished with a standardized, structured data collection form that clearly specifies standards and criteria of performance, and that those gathering the data have been trained in the appropriate procedures. The performance measurement forms shown in Figures 6-2 and 6-3 are examples of structured data collection forms. More specific instructions concerning manual data collection procedures and forms are discussed in a later subsection.

A second consideration that affects how trainee performance will be measured is whether training for a given task continues until specified criteria are achieved, or for a fixed number of trials or for a fixed amount of time. If a "train to criteria" procedure is used, training on a given task is continued until standard criteria are achieved, and the number of trials or amount of time spent on that task is recorded. Trials or time to reach criterion are the primary measures of performance; they can be supplemented by a count of the number and magnitude of errors or deviations from desired performance parameters. On the other hand, if training for a given task is continued until a fixed number of trials has been given, or until a fixed amount of time has elapsed, then an absolute measure of terminal performance for each task must be recorded. This can be accomplished by assigning an overall grade to performance. A measure of terminal (end of training) performance for each task can be expressed as the average score or grade obtained by a trainee on the last two or three trials for that task. This overall measure can be supplemented by a count of the number and magnitude of errors or deviations from desired performance for those trials.

Program of Instruction for ATD and Aircraft Training

The transfer of training from ATD to aircraft is a function of how the ATD is utilized within a specific training context. Consequently, the sequencing and content of materials used in a training program will greatly affect the extent to which skills learned in the ATD transfer to subsequent performance in the aircraft. The sequence in which tasks are trained is important, because ATD skills learned early in a training program will have transfer relationships not only to training in the aircraft, but to later ATD training as well. The content of the training materials involved will determine the adequacy of the foundation upon which further skill acquisition occurs.

Training material content and the sequence and structure of training must be carefully considered when constructing a program of instruction (POI) to be employed in a TOT evaluation. To the extent that there are design differences between the ATD and aircraft, POIs should recognize such differences. While the initial inclination may be to make ATD and aircraft training as similar as possible, it should be kept in mind that adoption of aircraft training procedures as the model for ATD training is not appropriate. ATD training POIs should be designed so as to utilize the training capabilities of the ATD to greatest advantage. However, there should be commonality of content between ATD and aircraft POIs to allow evaluation of common performance parameters.

The first step in developing a POI involves deciding upon the specific tasks, maneuvers or skills to be trained, and upon those to be measured during the TOT evaluation (task selection was discussed above). The specific maneuvers or skills included in a POI and their sequence of training depend heavily on three factors:

- The stated objectives of the particular evaluation.
- The characteristics of the ATD to be evaluated.
- The experience and background of the ATD and Control Groups.

Once the specific content of the POI has been decided upon, the sequence in which this content will be taught for both the ATD and Control Groups must be determined. It is important to make sure that the training sequence used in the POI is constructed so that skills already learned will provide a foundation for the learning of subsequent material.

Sequence of ATD and aircraft training. There are basically two general ways to sequence ATD and aircraft training for those students who will receive part of their training in the ATD. In the first, all the ATD training to be administered is completed before training in the aircraft is conducted. In the second, a set of maneuvers, or a block/stage of training, is taught in the ATD and then in the aircraft, followed by the next set of maneuvers or block/stage of training in the ATD and then in the aircraft, etc. This results in a series of alternating ATD and aircraft training episodes.

The first type of sequencing will allow inexperienced trainees to be employed in the evaluation and thus allow the impact of ATD training on initial flight skill acquisition to be assessed. However, this design does not allow the evaluator to ascertain to what degree each separate ATD sortie, with its particular mission segments, has an impact on subsequent performance in the aircraft. This is because all

ATD training is received prior to the trainee's introduction to the aircraft, and the cumulative effect of the total ATD training program is all that can be evaluated. In addition, there is the risk that maneuvers learned early in ATD training, if not practiced subsequently, may deteriorate by the time aircraft training begins and, hence, result in lower estimates of transfer than are warranted.

The second sequencing, in which ATD and aircraft training are intermixed, allows the individual impact of smaller portions of the ATD training program to be assessed, because performance in the aircraft is measured after each training objective in the ATD is attained. If this method is used, the decision remains concerning how much ATD training will be given before aircraft training. Smaller numbers of maneuvers trained prior to aircraft training allow more specific assessment of ATD training effectiveness. This also minimizes the time interval separating performance acquisition and testing. However, there are several factors that constrain the "size" of ATD training blocks. First, there are aircraft scheduling factors that must be recognized, and the whole mission aspect of groupings of maneuvers and activities must be considered. Second, aircraft training sorties must be constructed that maximize the training value of flying time. Third, all ATD training for a given maneuver, or set of maneuvers, must be completed before aircraft training (and assessment) of those maneuvers begins. Finally, safety, cost, or other factors may limit, or even prevent, the alternated sequencing of ATD and aircraft training.

Structure of training. Another consideration that impacts the development of a POI is how ATD and aircraft training will be structured. This training can be structured in several ways: (1) trials to criterion, (2) fixed-trials procedure, or (3) fixed-time procedure.

Trials to criterion. With this procedure training is continued for as many trials as are necessary for standard criterion performance to be attained. This approach ensures that each trainee reaches a predetermined proficiency level on every task before progressing to a new set of material or before practicing those tasks in the aircraft. While the trials to criterion approach has many advantages, it has two limitations that should be considered. These are: (1) the approach may take a considerable amount of time because individual trainees learn at different rates; and (2) variable rates of learning may cause scheduling problems. This approach also requires that suitable and objective performance criteria be developed for each task to be taught and evaluated in the study.

Fixed-trials procedure. With this approach, each task is trained for a fixed number of trials only. The number of trials should be based on extensive prior experience in training the task, and it should reflect the optimum number of trials needed for the average student to reach task proficiency. An advantage of this approach is that scheduling and instruction can proceed without variability in the rate at which individual trainees reach proficiency. One disadvantage is that some trainees will not reach proficiency on each task before moving on to the next task, whereas other trainees will continue to train on tasks after they already have reached proficiency. If optimal numbers of trials required to train a task to proficiency cannot be determined, two or more values can be used for each task. One value should be less than the estimated optimum number and one value should be greater than the estimated optimum number. This allows the effect of additional trials beyond some minimum to be determined empirically by the results of the study. Note, however, that this approach requires two ATD groups: one group which receives the lower fixed number of ATD training trials and one which receives the higher number of trials. This approach requires a total of three groups: two separate ATD training groups and one Control Group that receives no ATD training.

Fixed-time procedure. This procedure is identical to the "fixed trials" approach, except that each task is trained for a prespecified amount of time rather than a specific number of trials. The fixed time approach is generally the least preferred approach because it provides the least standardized ATD training, i.e., the number of trials or repetitions given per unit of instruction may vary from student to student.

In some cases, the selection of a training format may not be up to the test director because he may not have sufficient control over syllabus development. Nonetheless, he should attempt to construct/select a format that best meets the evaluation objectives.

Development of training sorties. After the content, sequencing, and structure of training have been determined for ATD and Control Groups, the training material must be grouped into training sorties. That is, the amount of material to be learned during each training session for each group must be decided upon. The amount of material to be included in each training sortie will be determined largely by the amount of time devoted to each. For example, training material will be grouped one way if it is to be presented in six 90-minute

training sessions, and in a very different way if it is to be presented in nine 60-minute sessions. The length of a training sortie will, in turn, be determined by such factors as the total number of ATD or aircraft hours available, the number of trainees in the ATD and Control Groups, and the complexity of the material to be learned relative to the background of the trainees involved.

Once the above items have been completed, a detailed plan for each training sortie can be developed. A detailed description of each task to be completed in the training sortie and their sequence must be written down in a format which can be utilized by the trainee and instructor. When this step has been completed, a written POI consisting of a particular sequencing of training sorties or sessions, which in turn consist of a sequence of individual tasks or maneuvers, for both the ATD and Control Groups will have been developed.

Data Collection/Analysis Format and Methodology

After a POI has been developed, and performance parameters and measures have been decided upon, data collection format and procedures must be developed. This involves constructing data collection forms, developing data collection and handling procedures, and developing data reduction and analysis strategies.

Data collection forms. A data collection format should be constructed which will allow the instructor or evaluator to record trainee performance data while performance is being observed. When performance data are to be collected manually (which will almost always be the case), the data collection format should be structured so that performance data can be quickly and easily recorded. The data sheets should fit easily on a clipboard or knee board. One approach is to use a series of formatted data sheets which contain a space for training each task to be performed and evaluated. The sequencing of tasks on the sheets is determined by the order in which trainees perform them during the maneuver, how the maneuvers are sequenced in the sortie, and how the sorties are sequenced during the training program itself.

Data handling and management procedures. Data handling and management procedures must be developed before data collection begins. These procedures include identifying who will be responsible for handling data collection forms and how the data will be handled. Some recommendations concerning who should be responsible for data handling are given below.

- Data collection forms may be held at training facilities and trainees may be made responsible for obtaining and bringing data collection forms to training.

- Instructors/evaluators should be responsible for turning in the completed data forms to the person managing data collection activities. This should be the test director or his appointed assistant(s).

- There should be personnel dedicated to managing data collection activities, including responsibility for receiving and recording of completed data forms, and for preparing/issuing data collection forms to trainees.

In addition to identifying personnel responsibilities, specific procedures must be developed for handling the collected data. These procedures involve where data collection forms will be issued, where they will be collected, how data contained on the data forms will be recorded and stored, and how "missing" data will be accommodated. (The actual management/monitoring of data collection activities is discussed in the section covering test execution.)

Data analysis. The specific tests and analyses to be performed on the collected data also should be specified before the TOT evaluation begins. However, the specific analyses to be used will vary depending on a host of factors, including how the TOT study is designed, (e.g., the number of evaluation groups), the type and nature of performance measures, the specific evaluation objectives, etc. It is recommended that data analysis procedures be developed with the assistance of a qualified behavioral scientist who has experience with statistical procedures and transfer-of-training designs.

Assignment of Trainees to Groups

A critically important part of conducting a TOT evaluation is to make sure that the ATD and Control Groups employed are as similar to one another as possible with respect to factors that are likely to affect performance. By assembling groups that are as similar as possible initially, and then training them in different ways (ATD training versus no ATD training), differences in performance between the groups can be attributed to differences in the training they received.

An initial similarity among the trainees in the ATD and Control Groups is required for an accurate evaluation of an ATD's effectiveness within the context of an ongoing training program. It also affects the number of trainees required for the evaluation; as the degree of initial similarity between trainees in ATD and Control Groups increase, the required number of trainees required for each group decreases. This is because trainee differences, other than those which result from the training received, will contribute less to subsequent performance differences between the groups if the groups

were similar to begin with. Having groups comprised of trainees of similar characteristics reduces the likelihood that the contribution of ATD training will be obscured by the impact of other factors; therefore, a smaller number of subjects can be used to evaluate its effectiveness.

Three techniques can be used to assemble equivalent evaluation groups through trainee assignment. The first technique involves the random assignment of individual trainees to the various evaluation groups. The second technique involves the random assignment of intact groups of trainees passing through a training program to the various evaluation groups. The third technique involves assembling evaluation groups comprised of trainees who are matched on one or more individual factors which are believed likely to affect performance. The procedural details, advantages, and disadvantages of each technique are described below.

Random individual assignment. One technique for assigning trainees to evaluation groups is to assign individuals randomly to the ATD and Control Groups from the overall pool of available trainees. Random assignment means that each trainee has an equally likely chance of being assigned to each group. Random assignment can be accomplished in a number of ways. If the TOT evaluation requires only two groups of trainees, each trainee may be assigned to either the ATD or Control Group by the flip of a coin. If more than two evaluation groups are to be employed, the same random assignment process can be accomplished with the use of dice. For example, if a TOT evaluation is to employ four groups of trainees, a die can be rolled for each trainee until a 1, 2, 3, or 4 comes up. On the first roll in which one of these numbers comes up, the first trainee is assigned accordingly to the first, second, third, or fourth evaluation group. The process is repeated until all trainees are assigned to a group. One restriction normally imposed on the process of randomly assigning trainees to groups is that each group contains an equal, or nearly equal, number of trainees.

The process of random assignment is intended to ensure that all groups will be essentially equivalent in terms of individual characteristics. However, this assumption is safe only when group sizes are relatively large. Problems arise when group sizes are small, because small randomly chosen evaluation groups are not likely to be equivalent on every trainee characteristic likely to affect performance. Therefore, it is recommended that random assignment be used only when it is possible to have a minimum of 20 to 25 trainees in each evaluation group.

Random intact group assignment. A second technique for assigning trainees to evaluation groups is to use intact groups in which groups

of trainees passing through a training program are assigned to either an ATD or Control Group. Intact groups may be assigned to ATD or Control Groups in the same manner described above for assigning individuals to groups.

This procedure is a convenient way to assemble evaluation groups; however, one problem is that any disruptive or unscheduled event which occurs to the group before or during the evaluation may differentially affect performance of the entire evaluation group. Another problem is that the members of an intact group of trainees may have characteristics in common with each other that they do not have in common with trainees in the other evaluation groups; this may bias the outcome of the evaluation. One way to minimize this possibility is to assign each group member to an evaluation group randomly as they pass through the training program. However, it may be logistically difficult or impossible to stagger the construction of evaluation groups in this manner.

As with the first procedure, this random assignment of groups of trainees to evaluation groups should only be done when a substantial number of trainees are to be used in the evaluation. It is recommended that intact groups of trainees be randomly assigned to evaluation groups only when each evaluation group can be comprised of a minimum of 30 trainees. Overall, the intact group is the least preferred technique for assigning trainees to groups.

Matched individual assignment. A third technique for assembling similar groups of trainees is to match or equate individual trainee characteristics among groups. That is, for each trainee with given characteristics in one group, there will be a trainee of similar (identical if possible) characteristics (e.g., amount of flight experience) in each of the other groups. This matching procedure is especially effective when few trainees are available for an ATD evaluation. It involves identifying trainees that are similar to one another on one or more characteristics likely to affect performance, and then assigning one of these trainees to each evaluation group. Doing this for all available trainees will create evaluation groups that are completely matched on the individual characteristics involved. Some individual characteristics which are likely to affect trainee performance are: (1) total trainee flying time; (2) the type of previous flying experience; (3) recency of previous flying experience; and (4) grades or performance scores associated with previous flying experience (e.g., UPT scores).

An example illustrating the matched individual procedure for assigning trainees to evaluation groups is described as follows: A test director is assigned the responsibility for evaluating the effectiveness of an ATD used in training air-to-air combat skills. The

evaluation objectives require that two groups of trainees be used, a group that receives its training in both the aircraft and the ATD, and a Control Group that receives all its training in the aircraft. A total of 20 trainees are available to be used in the evaluation. Some of these trainees are transition pilots with varying amounts of flying experience and others have just graduated from UPT; they also have varying amounts of flying experience.

There are too few trainees available to use a procedure for randomly assigning trainees to evaluation groups; therefore, groups are constructed by matching them on one or more factors which may affect performance. Because the type and amount of previous flying experience is likely to affect performance during training, it is decided to construct two equivalent evaluation groups by matching them on these two factors.

The trainees available for assignment are listed below.

Trainee	Type of experience	Previous flying hours
1	UPT	320
2	Transition	1200
3	Transition	1180
4	UPT	350
5	Transition	1050
6	Transition	1063
7	UPT	290
8	UPT	306
9	Transition	835
10	UPT	415
11	Transition	1185
12	UPT	315
13	Transition	860
14	Transition	910
15	Transition	915
16	UPT	351
17	UPT	305
18	UPT	420
19	Transition	1210
20	UPT	287

The first step in constructing two matched groups with this list of trainees is to divide the above list into separate lists of Transition and UPT trainees. Doing this, we get:

UPT		TRANSITION	
Trainee	Flying hours	Trainee	Flying hours
1	320	2	1200
4	350	3	1180
7	290	5	1050
8	306	6	1063
10	415	9	835
12	315	11	1185
16	351	13	860
17	305	14	910
18	420	15	915
20	287	19	1210

The second step in constructing two matched groups is to rank order each of the above groups from least to most previous experience. Doing this, we get:

UPT		TRANSITION	
Trainee	Flying hours	Trainee	Flying hours
20	287	9	835
7	290	13	860
17	305	14	910
8	306	15	915
12	315	5	1050
1	320	6	1063
4	350	3	1180
16	351	11	1185
10	415	2	1200
18	420	19	1210

The third step in constructing matched groups is to draw a line under every other trainee in each group which will create five pairs of subjects in each group. If three evaluation groups were to be employed, the line would be drawn under every third trainee instead of every second. Doing this, we get:

UPT		TRANSITION	
Trainee	Flying hours	Trainee	Flying hours
20	287	9	835
7	290	13	860
17	305	14	910
8	306	15	915
12	315	5	1050
1	320	6	1063
4	350	3	1180
16	351	11	1185
10	415	2	1200
18	420	19	1210

The final step in this process is to take each pair of trainees that has not been created and randomly assign one member of the pair to the ATD evaluation group and the other member of the pair to the control evaluation group. This can be accomplished by flipping a coin. If heads comes up, the first member of the pair goes to the ATD Group and the second goes to the Control Group. If tails comes up, the first member of the pair goes to the Control Group and the second member goes to the ATD Group. Doing this for the example above, the following ATD and Control Groups were obtained:

ATD EVALUATION GROUP			CONTROL EVALUATION GROUP		
Trainee	Type	Flying hours	Trainee	Type	Flying hours
20	UPT	287	7	UPT	290
8	UPT	306	17	UPT	305
1	UPT	320	12	UPT	315
4	UPT	350	16	UPT	351
10	UPT	415	18	UPT	420
9	TRANSITION	835	13	TRANSITION	860
15	TRANSITION	915	14	TRANSITION	910
6	TRANSITION	1063	5	TRANSITION	1050
11	TRANSITION	1185	3	TRANSITION	1180
2	TRANSITION	1200	19	TRANSITION	1210

This process allowed two evaluation groups to be created which are nearly identical on the type and amount of previous flying experience. These two equivalent evaluation groups can now be employed in the TOT evaluation and any differences in performance obtained for them can be confidently attributed to the different types of training received. Note that a random group construction process could have allowed the ATD Group to be comprised largely of UPT pilots and the Control Group to be comprised largely of Transition pilots. If this had happened, the difference in previous flying experience for the two groups may have biased the evaluation results and perhaps caused the test director to draw inaccurate conclusions from the training data obtained.

Selection and Assignment of Instructors/Evaluators

In most cases, the test director will not have the option of selecting the instructors/evaluators who will participate in the TOT study. However, he can specify some minimum criteria for their selection. Such criteria should include minimum experience level as an instructor and some minimum qualification as an instructor/evaluator for that stage of training or type of performance he is to instruct or evaluate. There also should be some provision for releasing those instructors who clearly show strong negative feelings toward being a participant in the study.

Assignment of instructors/evaluators. An important aspect of conducting a TOT study is to keep factors that may affect trainee performance as similar as possible for ATD and Control Groups. Otherwise it will be difficult to determine whether or not differences in performance between the groups are due to ATD training, or to those other factors. Trainee characteristics represent one factor that significantly affects performance; procedures for maintaining similar trainee characteristics between evaluation groups were discussed in the previous section. Another factor that significantly affects measured performance is the quality of instruction and the manner in which performance is evaluated. It is important, therefore, to match or equate instructor/evaluator characteristics across evaluation groups so that individual differences of instructors/evaluators won't bias the outcome of the study.

There are two basic methods that can be used to control for the effects of individual differences of instructors/evaluators. First, a separate group of instructors/evaluators can be used for each ATD and Control Group. Individual instructors/evaluators then could be assigned to the various groups randomly, or they could be matched according to those characteristics/factors that are likely to affect the quality of their instruction or the way in which they evaluate performance. Some of these factors may include:

- Flying experience
- Experience with the ATD being evaluated
- Instructional experience
- IP attitude toward ATDs

A second method for controlling the potential effects of individual differences among instructors/evaluators is to have each instructor/evaluator work with an equal number of trainees from each evaluation group. For example, if the TOT study included two ATD Groups and one Control Group, then each instructor would instruct an equal number of trainees from each group, and each evaluator would evaluate the performance of an equal number of trainees from each group. This has the effect of evenly distributing the effect or influence of a given instructor/evaluator across the ATD and Control Groups.

Instructor/Evaluator Training

All participating instructors and evaluators must be adequately trained in the procedures to be used during the TOT study, and they must faithfully execute those procedures if the study results are to have rigor and integrity. Instructor/evaluator training also should include a brief course covering the operation of the ATD, the use of the instructor's console, and the use of any special instructional features that are to be used during the study.

An instructor/evaluator course should be constructed and given to all instructors/evaluators who are scheduled to participate in the study. Provisions also should be made for administering the course to those instructors/evaluators who become participants after the study begins, i.e., replacements. The exact content of an instructor training course will vary from study to study; however, a brief outline of the issues to be covered by such a course is given below:

Overview. All of the instructors and evaluators should be instructed concerning the overall perspective of the TOT study to be conducted. This should state the purpose of the study, how long it will last, how the results will be used, and the role of instructors and evaluators. The agency or organization with responsibility for conducting the study also should be identified as well as key personnel and points of contact.

Training procedures. An overview of the procedures to be used for conducting ATD and aircraft training should be given that emphasizes the importance of maintaining the structure and standards as specified in the Program of Instruction. Specific issues include how ATD and aircraft training will be terminated, use of special ATD instructional features, the use of "instructional" trials, and so on.

Performance measurement. The manner in which trainee performance is to be measured should be specified clearly. Specific issues include when performance will be measured, i.e., which trials will be included as measurement trials and when during those trials performance will be measured. How performance is to be measured also must be specified, as well as those aspects of performance which are to be measured. In cases where specific maneuver standards/criteria have been developed, they should be reviewed and discussed with instructors/evaluators.

Data collection. All aspects of collecting performance measurement data should be discussed, including (1) how to fill out the data collection form; (2) who is responsible for bringing the data collection form to training; (3) who receives the completed data collection forms; and (4) who is responsible for overall data collection activities.

Deviations. Instructor/evaluator training also should include a section on how to handle deviations from specified procedures/schedule. Although deviations should be held to

a minimum, undoubtedly there will be occasions when specified procedures cannot be followed. Guidelines for handling these contingencies should be developed and briefed to the instructors. It also is recommended that a master log book be maintained in which all deviations are dated and recorded.

Testing. Finally, instructor/evaluator training should terminate with some type of "test" designed to assess the extent to which instructors/evaluators understand the procedures to be used during the TOT study. Ideally, this will consist of performance of all instructor duties and tasks under test director supervision (i.e., a kind of "practice teaching"), including actual practice with all data collection forms and procedures. Use of real students like those to be subjects in the OT&E is preferred in this instructor "testing."

Pretest Data Collection Forms and Procedures

One of the most frequently committed errors in conducting field research is the failure to pretest data collection instruments and procedures; it also can be one of the most costly errors. If possible the data collection instruments and procedures should be pretested by the actual personnel selected to use them and, to the extent possible, in an environment similar to the one in which the instruments are to be used. This introduces the data collector to some of the potential problems of using the instrument that may not have occurred to its designer. Frequently, data collectors are requested to record more data than they can accommodate in real time, or to monitor more activities than is reasonably possible. Pretesting data collection instruments typically reveals such areas of difficulty that, for whatever reasons, may have been overlooked or not dealt with sufficiently during the development of the measures. Pretesting frequently will identify unanticipated contingencies which could unfavorably impact data collection.

Pretesting data collection forms and procedures is especially important for the measurement of airborne performance. Pretesting also can be used to validate and refine any standards or criteria of performance that have been developed for the study.

The procedures for handling, reducing, and analyzing the collected data also should be pretested. This involves either collecting some sample data or generating "mock" data. Pretesting data collection, handling, reduction, and analysis procedures allows a realistic assessment of the efficiency and practicality of specified procedures and a good estimate of the manpower and time required.

Preparation of a TOT Study Plan

The activities of the Planning Phase culminate in the preparation of a TOT Study Plan. This plan documents the outcome of all the issues discussed in this section. A summary outline of the contents of a study plan is shown below. The outline can be used as a guide in writing a TOT study plan; it also can be used as a planning checklist.

TEST EXECUTION PHASE

The Test Execution Phase refers to the actual execution or conduct of the transfer of training study. It involves principally the collection of trainee performance data during ATD and aircraft training using the structure and procedures developed during the Planning Phase.

The purpose of this section, therefore, is to acquaint the test director with some of the contingencies that are likely to arise during a TOT study and which may affect its successful completion. General guidelines for anticipating, identifying, and handling contingencies when they arise are outlined for the two major components of test execution: data collection and test management.

Data Collection

The entire product of a TOT study consists of the trainee performance data collected during the course of the study. The interpretations, conclusions, and recommendations that come from a TOT study are based on these data. Consequently, every possible precaution must be taken to ensure the integrity, objectivity, and reliability of the data collected.

A critically important activity for the test director during the Execution Phase is to monitor, on a daily basis, the actual collection of the TOT evaluation data. One aspect of this monitoring process is to see that all data are collected and handled according to the procedures specified in the TOT Study Plan. Another aspect is to ensure that any deviations from specified procedures that may become necessary do not adversely affect or bias the outcome of the study.

Daily collection and inspection of data forms will help identify problems that may arise during data collection activities. Some of these problems include:

TOT STUDY PLAN OUTLINE

ESTIMATION AND COORDINATION OF REQUIRED RESOURCES

Schedule

- Personnel Requirements as a function of time
- Materiel Requirements as a function of time

Materiel

- Aircraft (number, type, availability)
- ATD (number, type, availability)
- Other

Personnel

- Trainees
- Instructors
- Evaluators
- Test Manager(s)
- Support

Data Collection, Reduction and Analysis Support

Contingency Plans

STATEMENT OF EVALUATION OBJECTIVES

General Objectives

Specific Objectives

LIST OF TASKS/MANEUVERS TO BE EVALUATED

PERFORMANCE MEASURES

What performance should be measured

Measure Parameters

Standards

Criteria

How performance will be measured

- Aircraft
- ATD

When performance will be measured

- Aircraft
- ATD

Data Collection

- Form
- Procedures
  - Data handling
  - Data reduction
  - Data analysis

Management of Data Collection Activities

PROGRAM OF INSTRUCTION

Content - training session descriptions

Structure of Training

- How training will terminate (trials to criterion, fixed trials or fixed time)
- Use of instructional trials
- Use of special ATD instructional features

Sequencing of Aircraft and ATD training

Training Sorties

ASSIGNMENT OF TRAINEES TO GROUPS

Type Method Used (random, intact group, or matched assignment)

Selection of Instructors/Evaluators

- Criteria for selection

SELECTION/ASSIGNMENT OF INSTRUCTOR/EVALUATORS

Type method used

INSTRUCTOR/EVALUATOR COURSE

Course Outline

PRETESTING PROCEDURES

- Data sheets not being filled out in the format or manner that was originally planned.
- Data sheets not being completely filled out (missing data).
- Wrong information being collected because POI is not being adhered to (i.e., wrong maneuvers being flown).
- Data sheets not being turned in at the right place or to the right person.
- Idiosyncratic differences in the ways that IPs record data on the data sheets (e.g., IP markings cover two item alternatives).
- IPs not filling out rating sheets accurately or conscientiously throughout the study.
- IPs not rating performance in a way that allows for differences in trainee performance to be distinguished.
- IPs developing different opinions on what constitutes correct performance or instruction on a maneuver.

Strict monitoring of data collection activities at the beginning of the TOT Study is especially important. First, it will take a while for instructors/evaluators to become accustomed to the new data collection format and procedures. Second, if the data collection forms and procedures were not pretested, modifications may be required for their successful implementation in the operational environment. Such changes should be identified and made as early as possible.

Test Management

Aside from carefully monitoring the data collection process, the test director has overall management responsibility for the TOT evaluation. This responsibility includes maintaining direct control over the TOT evaluation and ensuring test continuity.

Maintenance of direct control. The test director must maintain direct control of the TOT evaluation for its entire duration. Liberal delegation of authority or responsibility for the conduct of all or part of the evaluation may result in critical tasks not being completed satisfactorily. Test management is not a clerical function to be delegated, but an important part of the test director's job duties.

A major responsibility of the test director is to ensure that all procedures for conducting the TOT, as specified in the TOT Study Plan,

are followed. However, unanticipated contingencies may arise which prevent the exact implementation of specified procedures. For these instances the test director must evaluate the potential impact of any proposed deviation on the successful outcome of the study, i.e., the proposed deviation must be evaluated according to whether or not, and to what degree, it will bias the results of the TOT study. Seemingly minor changes in procedure potentially can invalidate an entire TOT study. Because of the importance of this type of decision, it is desirable that the test director have as part of his team, or readily available to him, a qualified behavioral scientist familiar with TOT evaluation issues.

A second aspect of maintaining control over the TOT evaluation involves monitoring and coordinating all the personnel and materiel resources required to complete the study. Especially critical personnel resources include trainees, instructors/evaluators, and any data handling/recording personnel. The test director must coordinate instructor/evaluator replacement and training activities. He also must monitor and coordinate the use of all materiel resources including aircraft, ATD, and support equipment (e.g., automated data recording/storage equipment).

Test continuity. The test director also is responsible for the smooth and efficient completion of the TOT evaluation. Any number of unplanned events can influence test continuity. Some of these can be planned for, whereas others cannot. However, some likely events are listed below:

- Weather conditions which result in the grounding of aircraft may cause unplanned delays and/or sequencing of tasks in training the ATD group(s).
- Equipment failure (both ATD and aircraft) also can disrupt the continuity of training.
- Trainer unavailability.
- Instructor/evaluator unavailability or transfer.
- Untrained instructors/evaluators.
- Unavailability of support personnel.

To the extent possible, precautions should be taken to minimize the affects of unplanned disruptions to training. Special care should be taken to minimize the time separating the completion of ATD training and the start of aircraft training. Long delays can reduce significantly the transfer of ATD training to aircraft performance. In addition, if two or more ATD groups are used in a TOT study, delays

between ATD and aircraft training should be similar for both groups. Otherwise, the obtained results may be biased.

POST-TEST PHASE

Activities conducted during the Post-Test Phase of a TOT study include the reduction, or summarization, of individual and group performance data, statistical analyses of the collected data, and an interpretation of the results.

Data summary. One of the first steps in summarizing the collected data is to organize the performance data by trainee for each task or maneuver employed in the study. A separate data sheet for each task or maneuver should be constructed for each task that was included in the TOT study. Both aircraft and ATD performance measures should be displayed for those trainees who were ATD Group members. Figure 6-4 shows a sample data sheet for summarizing task or maneuver performance data.

The second step is to summarize the data for each evaluation group. This involves averaging the performance measures of all the trainees in each group. In addition to providing average data on individual performance, some measure also should be provided of the extent to which individual performance varied. (The statistical appendix at the end of this volume provides instruction on how to "average" data and how to calculate measures of variability.) Figure 6-5 shows a sample data sheet for organizing the average data for each evaluation group. Figure 6-6 shows one way of graphically displaying average performance data that allows an easy task-by-task comparison of ATD and Control Group performance.

Data analysis. The procedures discussed above are useful for summarizing, or describing, the performance data collected during a TOT study. If there are any apparent differences in performance measures (e.g., trials to criterion), the test director must determine whether or not the obtained differences in performance are statistically reliable (i.e., due to factors other than chance). Statistical reliability refers to the likelihood that the same results would be obtained again if the study were repeated using the same subjects and procedures as before. Sometimes, obtained differences may be due to normal variations in performance that would not necessarily occur again. If statistical reliability of performance differences is not assessed, such normal variations in performance could be mistaken as "real" differences.

SUMMARY DATA SHEET FOR TASK/MANEUVER _____

	ATD PERFORMANCE		AIRCRAFT PERFORMANCE	
TRAINEE	TRIALS TO CRITERION	NO. OF ERRORS	TRIALS TO CRITERION	NO. OF ERRORS
1				
2				
3				
.				
.				
.				
.				
.				
.				
.				
.				
.				
.				
.				
n				
$\bar{X}$				
SD				

Figure 6-4. Sample data sheet for summarizing trainee maneuver performance.

EVALUATION GROUP (ATD OR CONTROL)

TASK/ MANEUVER	ATD PERFORMANCE				AIRCRAFT PERFORMANCE			
	MEAN TRIALS* TO CRITERION	STANDARD DEVIATION	MEAN NO. OF ERRORS	STANDARD DEVIATION	MEAN TRIALS* TO CRITERION	STANDARD DEVIATION	MEAN NO. OF ERRORS	STANDARD DEVIATION
A								
B								
C								
D								
E								
.								
.								
.								
.								
.								
Z								

*Performance data also can be expressed as the amount of training time required to reach criterion.

Figure 6-5. Sample data sheet for summarizing performance for an entire evaluation group.

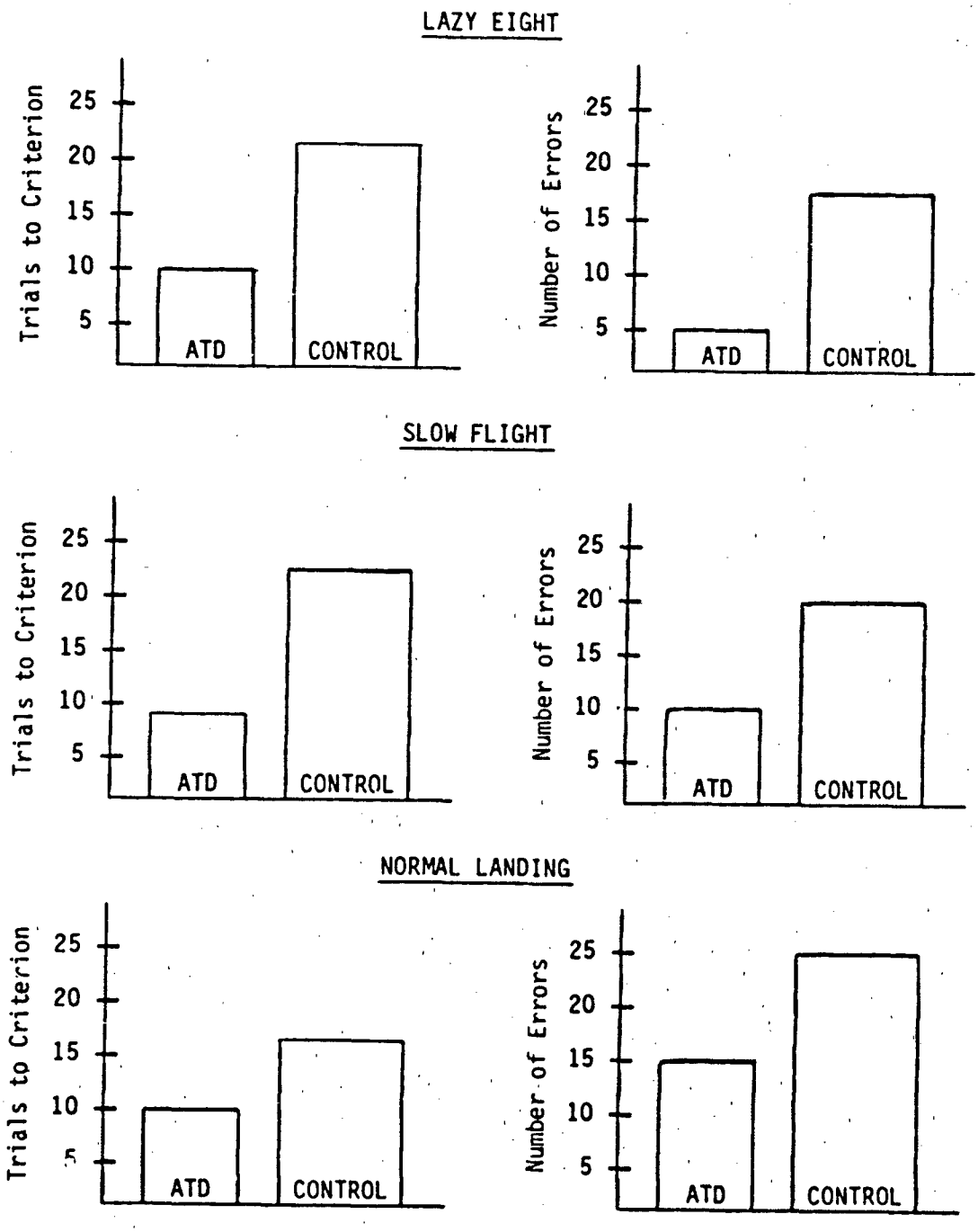


Figure 6-6. Specimen display of ATD and Control Group data.

A statistic frequently used to assess the reliability of obtained differences in performance is referred to as analysis of variance (ANOVA). Instructions on how to calculate a simple ANOVA are contained in the Inferential Statistics section of the statistics appendix to this volume. However, ANOVA procedures can involve very complex analyses, especially if: (1) multiple evaluation groups are used in the TOT study; (2) if multiple measures of performance are collected; or (3) evaluation groups are comprised of different numbers of trainees. It is recommended, therefore, that assistance be sought from a qualified behavioral scientist or statistician with appropriate expertise in statistical analysis.

Transfer of Training Measures

The ultimate purpose of a transfer study is to determine the extent to which training received in an ATD affects subsequent performance in the aircraft. This normally is accomplished by comparing aircraft performance of trainees who received some ATD training against the aircraft performance of trainees who did not receive ATD training. Performance, in this case, refers to the amount (trials or time) of aircraft training required to achieve a specified performance level or criterion. Differences in the amount of training required to attain criterion performance between the two groups can be attributed to ATD training, provided that sufficient control precautions have been taken to prevent bias.

The Transfer Ratio (TR)

A Transfer Ratio (TR) is a measure of the proportion of a training requirement that can be satisfied by ATD training. A TR is calculated as follows:

$$TR = \frac{CON - ATD}{CON}$$

where:

CON = Trials, time, or errors required by a Control Group to reach criterion performance in the aircraft without prior ATD training.

ATD = Trials, time, or errors required by an ATD Group to reach criterion performance in the aircraft after receiving some ATD training.

To illustrate, if it took 10 trials for Control Group trainees to reach criterion performance on a takeoff task without prior ATD training, but required only 4 trials for trainees in the ATD Group to

reach criterion performance on the same task, then the TR for that task would be 0.60. That is,

$$TR = \frac{10 - 4}{10} = 0.60$$

This transfer ratio is an index of the savings in aircraft training time that can be achieved through ATD training. Whereas it normally requires 10 trials of aircraft only training to attain criterion performance, equivalent performance can be attained with ATD training plus only 4 trials of aircraft training--a reduction in aircraft trials of 60%.

The Transfer Effectiveness Ratio (TER)

One limitation of the Transfer Ratio is that it does not take into account the efficiency of ATD training. To offset this limitation another measure of transfer, Transfer Effectiveness Ratio (TER), can be calculated as follows:

$$TER = \frac{CON - ATD_1}{ATD_2}$$

where:

CON = Trials, time, or errors required by a Control Group to reach criterion performance in the aircraft without prior ATD training.

ATD_1 = Trials, time, or errors required to reach criterion performance in the aircraft after receiving ATD training.

ATD_2 = Trials, time, or errors given in the ATD prior to training in the aircraft.

The TER takes into account both the effectiveness and efficiency of ATD training as measured against the effectiveness and efficiency of aircraft training.

Using the example above, if 7 ATD training trials first were given to the ATD Group before aircraft training, then the TER would be:

$$TER = \frac{10 - 4}{7} = 0.86$$

If, however, the ATD Group received 10 trials in the trainer and still required 4 aircraft trials to reach criterion, then the TER would be:

$$TER = \frac{10 - 4}{10} = .60$$

Transfer ratios and transfer effectiveness ratios can be calculated for each task evaluated as part of the TOT study. These ratios then can be displayed in tabular form, i.e., a simple listing of TRs and TERs for each task, or they can be displayed graphically.

The Incremental Transfer Effectiveness Ratio (ITER)

Another measure of transfer of training is the incremental transfer effectiveness ratio (ITER). It is used to show the change, or increment, in transfer that results from successive units (trials or time) of ATD training. The ITER is an especially useful measure in determining the most cost-effective allocation of ATD training time. Since each successive unit (trial or time) of ATD training given for a specific task typically results in a progressively smaller increment in the amount of training that is transferred to the aircraft, there comes a point when further ATD training for a given task is no longer efficient or cost-effective, i.e., a point of diminishing returns is reached.

The following example illustrates how ITERs are calculated. Assume a TOT study was conducted using four evaluation groups: one control group that did not receive any ATD training prior to aircraft training, and three ATD groups that received 10, 20, and 30 trials, respectively, of ATD training prior to aircraft training. ATD and aircraft training trials are shown below:

	Trials given in the ATD			
	0	10	20	30
Trials to criterion in the aircraft	15	12	10	9
Aircraft trials saved by ATD training	0	3	5	6

The ITER is calculated as follows:

$$ITER = \frac{\text{Incremental savings in trials or time to achieve criterion performance in the aircraft attained as a result of increased ATD training}}{\text{Increment in the amount (trials or time) of ATD training given}}$$

The increment in training effectiveness attained from increasing ATD from 0 (Control Group) to 10 trials thus is:

$$\text{ITER} = \frac{15 - 12}{10} = \frac{3}{10} = 0.30;$$

the further increment in training effectiveness attained from increasing ATD training from 10 to 20 trials is:

$$\text{ITER} = \frac{12 - 10}{10} = \frac{2}{10} = 0.20; \text{ and}$$

the final increment in training effectiveness attained from increasing ATD training from 20 to 30 trials is:

$$\text{ITER} = \frac{10 - 9}{10} = \frac{1}{10} = 0.10.$$

Table 6-1 shows TRs, TERs, and ITERs calculated for the above example to illustrate the relationship among these three measures of transfer. As can be seen, the obtained transfer ratio increased from 0.20 to 0.40 as the number of ATD training trials increased from 10 to 30, as reflected in a reduced number of aircraft training trials subsequently required to achieve criterion performance. On the other hand, the increment in transfer decreased from 0.30 to 0.10 as the number of ATD training trials increased from 10 to 30.

TABLE 6-1. EXAMPLE ILLUSTRATING THE RELATIONSHIP BETWEEN DIFFERENT MEASURES OF TRANSFER

Transfer measure	Trials given in ATD			
	0	10	20	30
Trials to criterion in the aircraft	15	12	10	9
Aircraft trials saved by ATD training	--	3	5	6
Transfer Ratio (TR)	--	.20	.33	.40
Transfer Effectiveness Ratio (TER)	--	.30	.25	.20
Incremental Transfer Effectiveness Ratio (ITER)	--	.30	.20	.10

B. SPECIFIC TOT STUDY DESIGNS

The preceding material has addressed a number of general topics and issues that relate to the planning, execution, and reporting of any transfer-of-training study. The following discussion treats four specific TOT designs that relate to the types of situations most likely to be encountered in ATD OT&E where a TOT evaluation may be applicable.

TOT DESIGNS

The four TOT designs to be discussed, shown in Figure 6-7, are all fairly similar. Each involves comparing the performance of two or more groups of aircrewmen who have had training programs that differ in the amount, or type, of ATD-based training. Selection of the TOT design appropriate to be employed in a particular ATD OT&E will depend largely on two factors--the evaluation objectives which need to be met, and the resources that are available to the test director to carry out the TOT study. The discussion that follows contains information about the specific application of each of the four TOT designs. It is intended to aid the test planner in selecting the design most appropriate to his particular test objectives and resource availability constraints. If none of the four TOT designs addressed here corresponds adequately to a particular ATD OT&E situation, assistance should be obtained (e.g., from AFHRL) to help in the modification of one of the designs, or to develop an adequate design.

For each of these four TOT designs, the discussion below contains a brief summary of the design in question, identifies its application, and presents an hypothetical example showing how resultant data might be summarized and interpreted.

Basic TOT Design

The Basic (and simplest) TOT design requires two groups of subjects. One group is an "ATD Group" that receives ATD training prior to training in the aircraft. The other is a "Control Group" that receives all of its training in the aircraft.

Applications. The basic TOT design may be applied in several situations. A common example for ATD OT&E would be when an ATD is added to an aircraft-only training program. Another example might be when a visual system is added to a nonvisual ATD. If the OT&E objective is to evaluate only the added value of the visual system, the appropriate control group would be a group of students who receive ATD training without the visual system. The ATD group would receive a

	DESIGN	GROUP	ATD	AIR-CRAFT
1	BASIC TOT DESIGN	C	NONE	Y_c
		A	X	Y_a
2	THREE GROUP TOT DESIGN	A_1	X_1	Y_1
		A_2	X_2	Y_2
		A_3	X_3	Y_3
3	DOUBLE TOT DESIGN	C_1	NONE	Y_{C1}
		A_1	NONE	Y_1
		C_2	X_1	Y_{C2}
		A_2	X_2	Y_2
4	ATD--COMPARISON TOT DESIGN	C	NONE	Y_c
		A_a	X_a	Y_a
		A_b	X_b	Y_b

C = Control Group. Receives aircraft-only training.

A = ATD Group. Receives some level of ATD training plus aircraft training.

X = Number of trials, errors, or amount of time spent in the ATD.

Y = Number of trials, errors, or amount of time spent in the aircraft.

Figure 6-7. Four specific TOT designs.

certain amount of training in the simulator with the visual system added. Should the OT&E objective, however, be to evaluate the total contribution of the new ATD to training, the most appropriate control group would be students who receive all their training in the aircraft.

Hypothetical example. The new AX-OFT is a full mission aircrew training device that is to be added to the current TAC AX aircraft-only training program. It is intended to be used for transition training in the AX weapon system (an attack system) program. For its transition training mission, it is intended that the AX-OFT be used for training basic airwork and aerobatics, landings, emergency procedures, low-level navigation, and air-to-surface weapons delivery. The AX-OFT is fitted with a forward looking 90° field of view, color visual system of high resolution and scene detail.

In an earlier conducted QOT&E, a rating scale procedure was employed that suggested the device should have high training capability for aerobatic maneuvers, emergency procedures, straight-in landings, and air-to-surface weapons switchology (procedures); moderate training capability for air-to-surface weapons delivery; but low training capability for low-level navigation and normal traffic pattern landings. The lower ratings on those maneuvers were attributed to the somewhat constrained field of view of the visual system.

During FOT&E, a basic transfer of training study was conducted to provide a data base for integrating the OFT into the ACX transition training syllabus. Fifteen tasks were selected for the TOT evaluation. The tasks were selected to be representative of the range of tasks intended to be trained in the device. Emergency procedures were excluded because they were not currently trained inflight.

A total of 28 trainees participated in the evaluation (Control Group N = 15; ATD Group N = 13). The Control Group trainees went through the current AX aircraft-only syllabus, during which time a specially constructed 5-point criterion-referenced grading procedure was employed for those tasks/maneuvers selected for study. Both number of trials and time spent in attainment of criterion performance were compiled for each selected task. The ATD Group trainees went through a specially adapted program of instruction (POI) wherein the selected tasks/maneuvers were introduced and trained to criterion level in the OFT prior to training in the aircraft. That POI followed the basic AX syllabus, but allowed for interspersion of OFT training and for any resultant "efficiencies" in aircraft training. The same criterion-referenced grading procedure was employed with the ATD Group for both the ATD and subsequent aircraft training.

Hypothetical results of the TOT are shown in Table 6-2. Transfer effectiveness ratios were calculated for both trials and time. Only

the Time and TER were calculated for low-level navigation, because "trials" per se did not apply to that task. The Trials TERs were calculated from the mean number of trials spent training in the simulator and the mean number of trials to criterion level spent training in the aircraft for each task/maneuver. The Time TERs were calculated using the mean number of hours (to nearest tenth) required for training the tasks/maneuvers to criterion level in both simulator and aircraft.

TABLE 6-2. TRANSFER EFFECTIVENESS RATIOS (TERs) BY TASK/MANEUVER FROM THE AX-OFT TO THE AX AIRCRAFT (RANK ORDERED BY TER-TRIALS) (Hypothetical Data)

Task/Maneuver	TER-Trials	TER-Time
Lazy Eight	.85	.86
Slow Flight	.82	.80
Barrel Roll	.79	.86
Straight-in Landing (D)	.73	.83
Takeoff (light)	.70	.90
Cuban Eight	.65	.70
30° High-angle Strafe	.65	.69
Takeoff (Heavy)	.63	.75
30° Dive Bomb	.60	.72
Power-on Stall	.59	.60
Level Bomb Pass	.50	.71
Steep Turn	.38	.40
Pop-up Attack	.25	.30
Straight-in Landing (N)	.22	.42
Low-level Navigation	NA	.21

In addition to the TERs, learning curves were plotted for both groups of trainees for each task/maneuver (Figure 6-8, example shown for Lazy Eight maneuver). The learning curves were constructed to show the 16th, 50th, and 84th percentiles of the groups.

These TOT data are largely consistent with the hypothetical QOT&E rating scale data. It would appear that airwork and aerobatics are among the OFT's better training capabilities. Contact maneuvers, including day landings and air-to-surface weapons delivery tasks, were shown in the middle range of transfer. The low TER achieved for night straight-in landing might be attributable to the fact that only a limited number of aircraft trials were given, regardless of proficiency level. The other lower TER tasks of Steep Turn, Pop-Up Attack, and Low-level Navigation may be attributed to the somewhat constrained 90° field of view.

From the total training program standpoint, these data indicate that the AX-OFT should fulfill its intended transition training mission well. It appears that the device could greatly facilitate attainment of basic airwork skills and familiarity with the AX aircraft handling characteristics. Also, the device appears to be effective for training combat skills including weapons switchology and delivery. The forward-looking visual system limits the effectiveness of the device for training low-level navigation because terrain objects to the sides cannot be utilized for determining position. Addition of side windows to the visual system would be required to enable effective training for that task area.

Three Group TOT Design

This TOT design incorporates three groups of subjects, each of which receives a different amount of ATD training prior to aircraft training. The three group design enables ATD training effectiveness to be determined in a way that allows transfer effectiveness functions to be plotted. These TER functions may be plotted with the transfer effectiveness ratio on the vertical axis and the amount of ATD training on the horizontal axis. Alternatively, aircraft hours following ATD training may be plotted on the vertical axis and ATD training on the horizontal axis. The advantage of this method of presentation is that it shows the decreasing effectiveness of ATD training over successive training trials.

Applications. The Three Group TOT design may be applied in a number of situations. For example, it can be used when some amount of ATD training is required of all trainees. Such might be the case when a new training device is replacing an old one and a zero-ATD control group is not feasible. In this case, the lowest level of ATD training might be kept as close to zero as practical. The middle and upper levels of ATD training should then be spaced at logical intervals

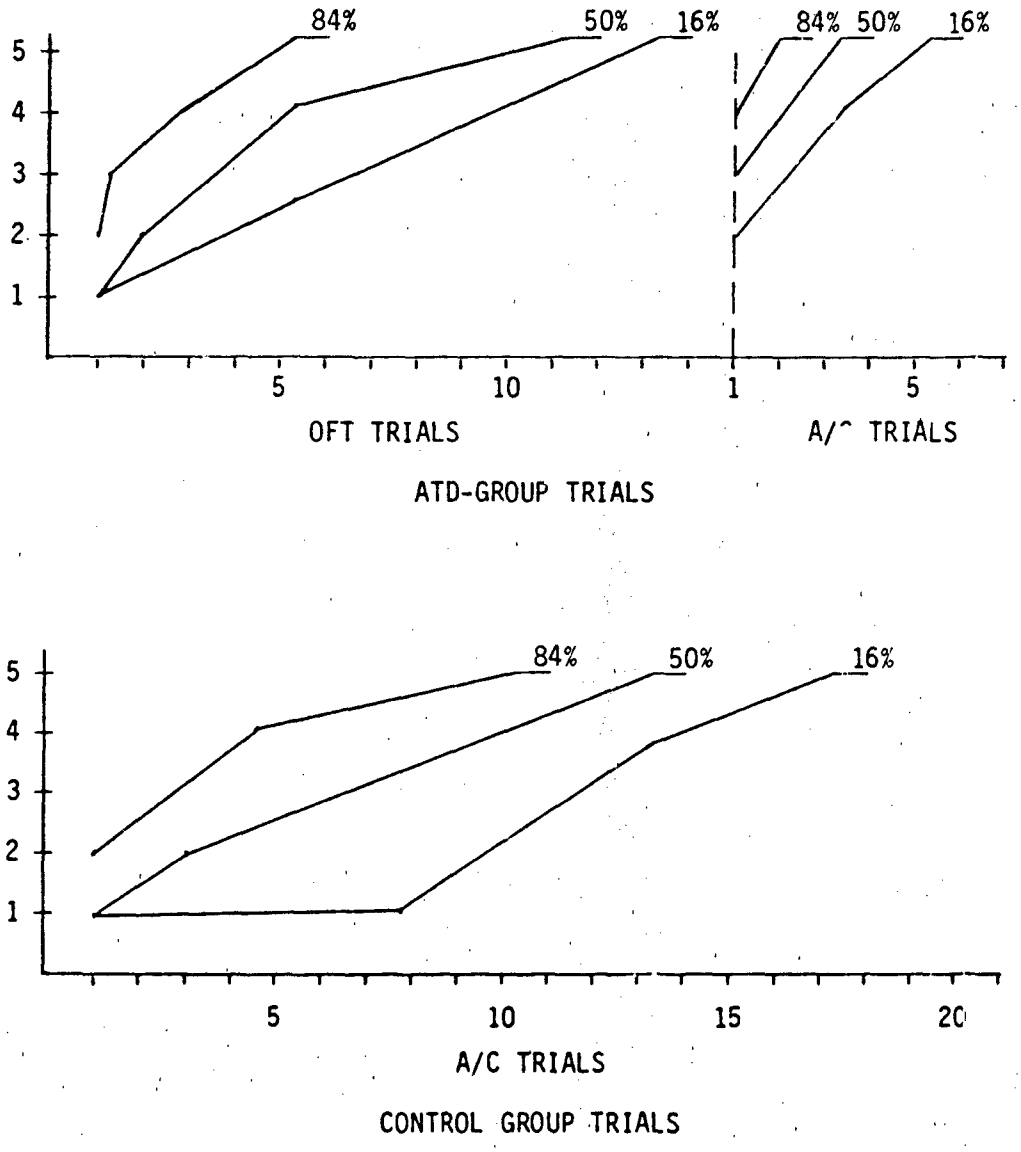


Figure 6-8. Hypothetical learning curves for lazy eight maneuver.

depending upon the training tasks in question and Subject Matter Expert (SME) judgment of anticipated training times/trials for those tasks. Another application of the three group TOT design might be to have a zero-ATD training control group, as in the Basic TOT design, and two levels of training for the ATD groups.

Hypothetical example. A new ATDX special task trainer has been implemented into an operational squadron to replace its old air refueling trainer. Because the current unit training syllabus for air refueling training calls for practice in the ATD prior to inflight training on that task, no zero ATDX group can be possible, i.e., some minimal level of ATDX training will be necessary for all trainees. For the TOT study, three levels of ATD training were selected as shown in Table 6-3.

TABLE 6-3. ATD AND AIRCRAFT TRAINING LEVELS FOR THREE GROUPS OF TRAINEES (Hypothetical Data)

Groups	ATD trials	A/C trials to criterion
ATD 1	5	10
ATD 2	15	7
ATD 3	25	5

The mean number of trials to reach criterion for each group also is shown in Table 6-3. These results can be shown graphically as depicted in Figure 6-9. This way of plotting the data shows the decreasing effectiveness of ATD training as time/trials in the device are increased. Another way of plotting these data is shown in Figure 6-10. Here, transfer effectiveness ratios are plotted on the vertical axis with ATD training on the horizontal. In order to obtain these TER values, a zero-point for aircraft training had to be interpolated. A conservative value for that zero-point aircraft training level was calculated based upon the slope of the function between 5 and 15 ATD trials (value = 11.5).

These data suggest a fairly consistent TER for the ATDX in training air refueling, although the expected decrement in effectiveness did occur to a small degree. The most significant result to note is the decrease in required aircraft training that resulted from additional ATD training trials.

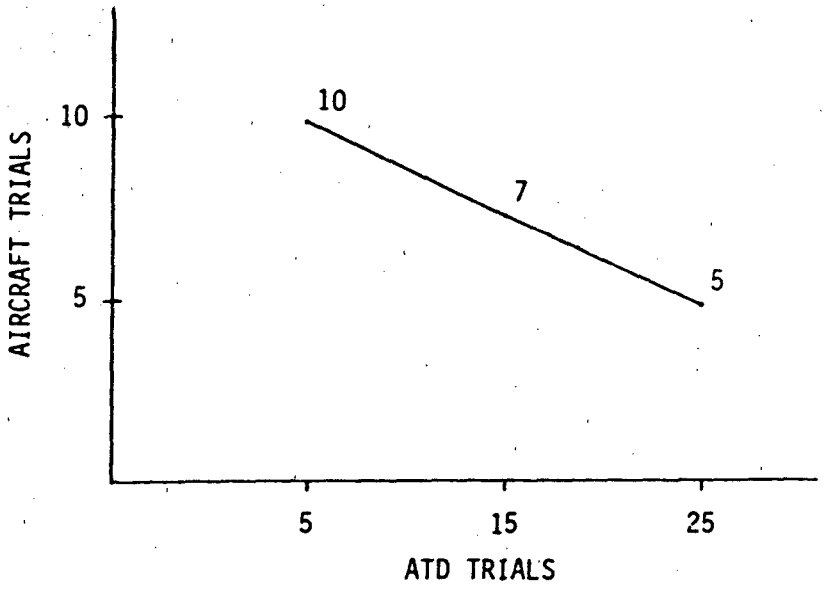


Figure 6-9. Transfer function showing relationship of ATD training to aircraft training.

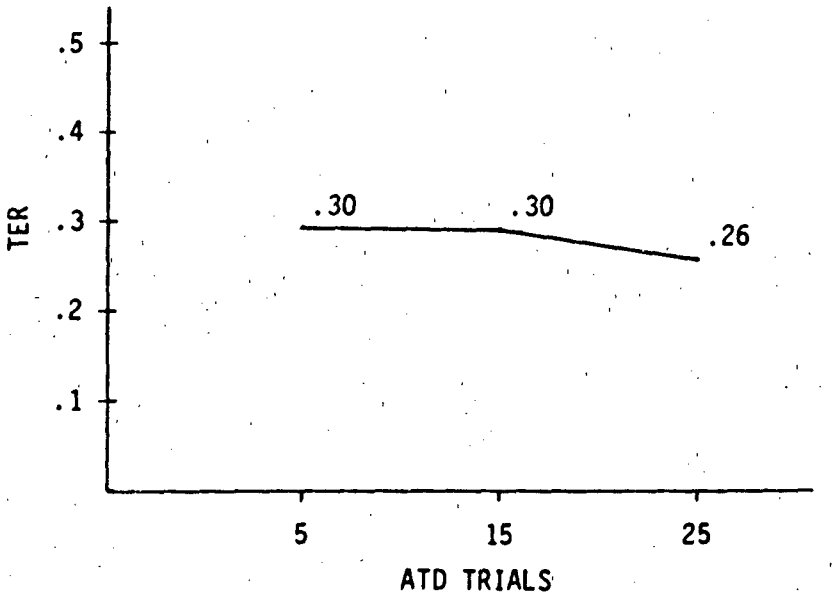


Figure 6-10. Transfer function showing relationship of ATD training to TERs.

Double TOT Design

The Double TOT design is actually nothing more than a combination of two simultaneous Basic TOT (two-group) studies. This design incorporates an "ATD Group" and a "Control Group" for each of two different populations of trainees. Accordingly, the Double TOT design requires essentially twice the planning and resource coordination effort that is necessary for the Basic TOT design. For example, two different sets of POIs will need to be developed; one appropriate for each trainee group. In addition, the execution phase activities of scheduling and coordinating data collection for two different populations of trainees can be much more complicated. Finally, data analyses and interpretations of results must be accomplished separately for each group, because the effectiveness of an ATD may be quite different for the two trainee populations.

Applications. The Double TOT design may be applied in situations where two different populations of trainees plan to incorporate a single ATD into their respective training syllabi, such as, for example, initial transition trainees and requalification trainees. This may often be the case in operational training squadrons. Drawbacks to use of the Double TOT design have to do with the added complexity for planning, executing, and reporting the study, as noted above.

Hypothetical example. The ATDX-WST is an air-to-air and formation trainer to be employed in training both initial transition trainees (out of UPT and fighter lead-in) and requalification transition trainees that have past operational experience in other similar weapon systems. In this dual role, the ATDX-WST could provide a convenient and safe mechanism for developing necessary competencies before training in the aircraft.

A Double TOT study was devised to evaluate the training effectiveness of the WST for both populations of trainees. Four groups were needed, i.e., an ATD and a Control Group for each trainee population. Subjects in associated ATD and Control Groups were matched on the basis of flight experience (hours) in the past six months. The Control Groups received no training in the WST prior to aircraft training. The ATD Groups were trained to criterion performance in the WST before transferring to the aircraft.

Table 6-4 shows the results of the TOT including mean training times to criterion for the formation task and mean numbers of engagements for air-to-air intercepts (the study was limited to addressing fingertip formation).

TABLE 6-4. RESULTS OF THE ATDX-WST TRANSFER-OF-TRAINING EVALUATION FOR INITIAL TRANSITION AND REQUALIFICATION TRAINEES (Hypothetical Data)

Groups	ATD training (mean values)	A/C training (mean values)	TERs
INITIAL TRANSITION:			
Control Group			
• Fingertip Formation	0	3 hours	-
• A/A Intercept	0	15 trials	-
ATD Group			
• Fingertip Formation	4 hours	1.5 hours	.38
• A/A Intercept	10 trials	9 trials	.60
REQUALIFICATION:			
Control Group			
• Fingertip Formation	0	1.5 hours	-
• A/A Intercept	0	5 trials	-
ATD Group			
• Fingertip Formation	2 hours	1.5 hours	0
• A/A Intercept	6 trials	4 trials	.16

From these results it would appear that the ATDX-WST is an effective trainer for Initial Transition trainees, but an ineffective trainer for Requalification trainees. The latter finding, however, does not necessarily reflect a deficiency in device design, but rather the fact that the Requalification trainees did not have as far to progress to reach criterion performance relative to the Initial Transition trainees.

ATD-Comparison TOT Design

The ATD-Comparison design incorporates three groups--a "Control Group" that receives no ATD-based training prior to aircraft training, and two "ATD Groups" that receive training in two different training devices prior to training in the aircraft. In this way, the relative training effectiveness achieved with either device can be used for making judgments and decisions about device configuration options, syllabus refinement, and the value or "worth" of some new technology from a training and cost standpoint.

Applications. As its name implies, the ATD Comparison TOT design applies to those instances where a comparison between two ATDs is desired. One example might be a comparison between two devices that employ different types of visual systems, e.g., CGI vs. camera/modelboard. Another application might be to compare some new training technology like CAI to a "conventional" cockpit-type ATD.

Hypothetical example. Two ATDs with different visual systems were procured to support squadron level training on the ACX transport aircraft. The visual systems in question varied in field of view (FOV) and scene detail. ATD-1 had a wide FOV and high density visual system, whereas ATD-2 had a comparatively narrow FOV and low density visual system. The devices were identical in all other respects. The critical question asked in the TOT study was, "what decrement in training effectiveness could be expected with the less expensive narrow FOV, low density visual system?" This was the critical question because the results of the TOT were to be used to aid in making a decision whether to procure the wide FOV or narrow FOV configuration for future Air Force-wide deployment.

The TOT study focused upon those tasks that required outside visual cues and orientation. Selected for study were takeoffs, landings, and low-level flight. Selected instrument tasks were also included to assure that the two ATDs were equivalent in this area. Control Group trainees went through the ACX aircraft-only transition training program. Both ATD Groups (ATD-1 and ATD-2) received identical training POIs that followed the basic ACX syllabus, but that allowed for interspersion of ATD training in the respective devices.

Results of the TOT evaluation are shown in Table 6-5. TERs were calculated for both trials and time. Only Time TERs were calculated for low-level flight, because "trials" per se did not apply to that task. These data indicate that the training provided by both devices for takeoffs and landings was roughly equivalent. For the low-level flight task, however, the wider field of view trainer (ATD-1) was clearly superior in training effectiveness. Data on both devices for instrument training were collected to assure equivalency in that area; no significant differences were found.

TABLE 6-5. RESULTS OF THE ATD COMPARISON TOT EVALUATION
(Hypothetical Data)

Groups (N)	ATD training trials/time (mean values)	A/C training trials/time (mean values)	TERs trials/time
Control Group (12)			
• Takeoff	0/0	14/1.5	-
• Landing	0/0	20/2.5	-
• Low-Level	0/0	NA/4.0	-
• ILS Approach	0/0	7/3.4	-
ATD-1 Group (10)			
• Takeoff	15/1.5	7/1.0	.47/.33
• Landing	15/1.5	12/1.5	.53/.67
• Low-Level	NA/3.0	NA/2.5	NA/.50*
• ILS Approach	5/2.0	3/1.5	.80/.95
ATD-2 Group (11)			
• Takeoff	16/1.	8/1.1	.38/.27
• Landing	20/1.7	13/1.5	.35/.59
• Low-Level	NA/3.0	NA/3.9	NA/.03*
• ILS Approach	5/2.0	4/1.6	.60/.90

*Significant difference at $p < .001$ between these two groups.

For a discussion of statistical significance, please refer to pages 241 and 242 of this volume.

CHAPTER 7

INSTRUCTOR/OPERATOR STATION EVALUATION

INTRODUCTION

The preceding chapters have treated evaluation approaches to ATD training effectiveness with primary emphasis upon the training effectiveness of the device trainee station, i.e., its capability to simulate the cues and responses necessary for training aircrew tasks. Ultimate ATD training effectiveness is, however, not only a function of the device's capability to simulate training tasks accurately, but also of its ability to operate as an effective instructional tool. Its effectiveness as an instructional tool depends on instructor/operator station (IOS) factors such as instructor/operator workload, performance monitoring and evaluation capabilities, and ease of training task/mission set-up. These and many other IOS capability factors must be taken into account during the operational testing and evaluation of an ATD.

Effective design and use of the IOS can greatly impact the potential and achieved training effectiveness of the device as a whole. The purpose of this chapter, therefore, is to provide guidance or evaluation of the ATD instructor/operator station.

This chapter contains four subsections. The first subsection provides a discussion on task requirements of the ATD instructor/operator for use in developing a task list for evaluation purposes. The second subsection addresses a number of common ATD instructional features with emphasis on their purposes and instructional use. The third subsection outlines the major evaluation concerns relative to the IOS during ATD OT&E. The fourth subsection provides an approach to IOS evaluation which employs a training scenario approach to enable assessments of its features and characteristics in a dynamic training framework.

INSTRUCTOR/OPERATOR TASKS

The evaluation of ATD training capabilities as presented in earlier chapters involved specifying the aircrew tasks to be trained in the device, and directing the evaluation to those specific tasks. Analytic procedures were defined for selecting among aircrew training tasks those for emphasis during the OT&E. The need for analysis to identify aircrew training tasks was based on the premise that, in an ATD OT&E, the emphasis should be upon evaluating the device in an "operational" training context. Likewise, in order to evaluate the

when a training mission involves repeated discrete trials of a given flight task or several discrete training segments on different flight tasks.

(3) Training Scenario/Problem Support. Training problem support tasks, unlike those in two previous I/O task categories, can be and are typically performed simultaneously, and are repeated as necessary to achieve desired instructional purposes.

Examples:

- Communications (e.g., role play ground/approach/tower, radio traffic features)
- Friendly aircraft control
- Threat aircraft control
- Environment modification

Note: Tasks are carried out in this category to support and implement the planned training scenario/problem and are modified as necessary in accordance with trainee performance and progress.

(4) Instruct. The tasks included in the instruct category are those that involve "active" instruction, i.e., the basic use of the ATD. It is through these tasks that the I/O functions as an active instructional process manager. Here, emphasis is on the instructor's role in controlling practice, providing guidance and feedback, "pacing" the training, and similar functions that involve contingent real-time manipulations of the instructional process. It is with reference to the instruct task category that IOS instructional features (e.g., freeze, replay, auto-demo, etc.) are addressed.[1]

Examples:

- Operate Record/Playback
- Store/Reset Current Conditions
- Demonstrate Task X

---

[1]Device instructional features are discussed more fully in the next subsection in this chapter.

- Manual/Auto Freeze
- Hardcopy

Note: Obviously, performance monitoring/evaluation tasks also relate closely to the instruct tasks, but they may be considered as functionally distinct and are, therefore, identified separately below.

(5) Performance Monitoring/Evaluation. I/O tasks in this category have to do with active monitoring and evaluation of trainee performance. Monitoring tasks may be differentiated from evaluation tasks in that the latter involve a decision making process. Performance monitoring provides much of the relevant information to be used in making those evaluative decisions.

Examples:

- Time/Position Monitoring
- Procedures Monitoring/Evaluation
- Cockpit Instruments Monitoring
- Visual System Monitoring
- Maneuver Scoring
- Weapons Delivery Scoring

Note: Evaluation tasks may involve use of both manual (instructor-monitored) and automated (computer-monitored/processed) performance data in maneuver scoring, weapons scoring, navigation scoring, etc.

(6) Trainee Debriefing. Trainee debriefing occurs typically at the end of the ATD training session, but may occur at other major transitions within the training session. During debriefing, the I/O reviews the training session with the trainee, points out significant performance discrepancies, notes progress relative to learning objectives, and makes recommendations to help the trainee.

Examples:

- Review Training Session
- Replay Significant Events
- Obtain Hardcopy

Note: In the process of debriefing, the I/O may wish to utilize various instructional features (e.g., hardcopy, replay) to facilitate communication with the trainee and otherwise enhance the debriefing. Logically, trainee briefing, which is an important instructor task, might have been included in this listing also. However, since the usual pre-instructional briefing seldom involves the IOS in a major active way, the briefing task is not considered basic to IOS evaluation in an ATD OT&E. In contrast, the IOS is much more likely to be used in an active way during debriefing, and, hence, the extent to which it is or is not facilitative of the debriefing task is of concern in ATD OT&E. Should the device and its IOS be used more actively in briefing, then briefing would properly be added to the present I/O task listing.

(7) Safety Monitoring. Safety monitoring has to do with both I/O and trainee safety in use of the ATD. As such, it is not an "actively" performed instructional task but, rather, consists of routine safety checks and monitoring of possible hazardous conditions.

Examples of Monitored Systems:

- Halon System
- Hydraulic/Pneumatic System
- Emergency Egress

(8) Program/Syllabus Support. These I/O tasks are performed "off-line," i.e., not during actual training. Included are those functions which involve preparation of the "IOS data base" that the I/O will use during periods dedicated to training. All of these functions involve semi-permanent storage and thus should not be enabled during period dedicated to training.

Examples:

- Demonstration Preparation
- Target Set Preparation
- Display Page Generation/Modification

Note: While portions of all of these tasks will be performed by I/Os, other portions of these functions must be performed by persons with sufficient computer programming and operations expertise.

ATD INSTRUCTIONAL FEATURES

ATD instructional features are those special capabilities of a device intended to facilitate and enhance its instructional utility. Properly employed, these features can significantly increase the effectiveness and efficiency of training. Instructional features can take many forms, ranging from the simple freeze control to more complex custom-tailored capabilities for demonstration of specific aircrew tasks (e.g., specific maneuver and weapons delivery demonstrations). Instructor/operator use of device instructional features is of particular interest during OT&E IOS evaluation. To evaluate such features, however, requires that the operation and intended use of those features be understood fully so that the content and structure of the evaluation may be properly directed. Also, instructional features are usually accessible to a limited degree from the ATD trainee station and, accordingly, are of interest in trainee station evaluation.

A representative listing of such ATD instructional features is shown below. This listing is based upon a recent effort that was directed toward developing a means for communicating the training use of instructional features to the ATD design community.[1] An understanding of the intended instructional function of each of these features would be required of OT&E personnel responsible for their evaluation. The point of concern for ATD OT&E is that the evaluation of such features must be cast in terms of their intended instructional use in their real-world setting. For example, record/playback may be of high technical quality, but if the I/O must wait for an extended time in order to reach (access) a desired playback segment, that feature, as implemented, might be judged instructionally unsatisfactory.

(1) Record/Playback

(2) Store/Reset Current Conditions

(3) Remote Display

(4) Hardcopy

(5) Manual Freeze

(6) Automatic Freeze

(7) Parameter Freeze

(8) Demonstration

(9) Malfunction Simulation

(10) Automatic Malfunction Insertion

The following paragraphs describe the general instructional function and use of the above features.

(1) Record/Playback. Record/Playback (R/P) is an ATD instructional feature that permits the I/O to replay a recent or immediately preceding segment of simulated flight. During a playback, all events which occurred as a consequence of trainee input to the ATD's controls will be reproduced.

The purpose of the R/P feature is to enable the trainee to examine his own performance and to aid the I/O instructor in critiquing trainee performance. R/P provides a faithful reproduction of performance that can be examined in detail at a pace determined by the instructor, repeatedly if necessary, while that performance is simultaneously being reviewed by the trainee. Its use will permit relationships between control inputs and system responses to be examined, and thus it can be employed with trainees having particular difficulty mastering a specific task. The most frequent use of the R/P feature will follow an error or a less than satisfactory performance by the trainee. Rather than waiting until a post-training period debriefing to critique that performance, the I/O will interrupt the simulated flight to replay the performance in question.

(2) Store/Reset Current Conditions. Store/Reset Current Conditions (S/R) is an ATD instructional feature that permits the simulation to be returned or reset to a set of conditions that existed at an earlier point in time. The primary purpose of the S/R feature is to permit a trainee to be returned to a previously encountered set of simulated conditions in order that he may repeat a maneuver or flight segment attempted earlier. The S/R feature provides a means of increasing the efficiency of the ATD instructional process by enabling the rapid and easy return to the exact conditions needed for a particular instructional event.

(3) Remote Display. The Remote Display (RD) feature permits alphanumeric and graphic data on an IOS display to be displayed simultaneously at the trainee station. The purpose of the RD feature is to enable the instructor at the IOS and trainee at the trainee station to view displayed information simultaneously. The feature will be employed to facilitate communication between the instructor and the trainee, particularly when the communication involves reference to graphic or symbolic information.

(4) Hardcopy. Hardcopy is an ATD instructional feature that enables the I/O to reproduce on paper data displayed at the IOS. The feature provides a copy of those data as they existed at the time the Hardcopy was initiated by the I/O. The copied display may be used by the instructor to compare the performance of a trainee at two points in time during a single instructional period or over several such periods, to compare the performance of several trainees on similar flight tasks, to aid the instructor in subsequent review of a trainee's performance, and/or to provide objective information for permanent record purposes.

(5) Manual Freeze. Manual Freeze (MF) enables the I/O or the trainee to freeze or suspend ongoing simulated activity resulting from input to the aircraft's controls (at the trainee station and at the IOS). During the period of Freeze, the simulated conditions existent at the onset of MF will be preserved, and the suspended activity may be resumed at the option of the I/O or the trainee. The primary purpose of the MF feature is to permit the interruption of the simulation so that other instructional or supporting activities may take place or to provide a break in the instruction. The secondary purpose of this feature is to provide a stable condition while the simulator is "on" that will allow necessary setup or simulation modification functions to be performed and cockpit ingress/egress.

(6) Automatic Freeze. The Automatic Freeze (AF) feature automatically freezes or suspends ongoing simulated activity when predetermined conditions are met. The purpose of the AF feature is to place the simulator in freeze status immediately upon the occurrence of specified events, and to do so without intervention by personnel at the IOS or trainee station.

(7) Parameter Freeze. Parameter Freeze (PF) enables the I/O to freeze one or more of the simulator flight parameters to its current value. When a parameter is in freeze status, all other parameters will be unaffected. The primary purpose of the PF feature is to enable the I/O to reduce the difficulty to the trainee of the task being performed. Such an approach might be employed to simplify aircraft control when a pilot is experiencing difficulty developing the skills required to fly the simulated aircraft, or while the pilot acquires skills at associated tasks such as tracking a missile on a target or learning to operate on-board avionics and associated displays.

(8) Demonstration. Demonstration (Demo) is an ATD instructional feature that consists of a prerecorded aircraft maneuver that provides

a model of the desired performance of the maneuver. The Demo reproduces all simulated flight conditions and aircraft performance that occurred when the maneuver was originally recorded. A Demo usually includes a synchronized audio briefing, explanation, and instructional commentary designed to facilitate the trainee's subsequent attempt to perform the maneuver. The purpose of the Demo feature is to provide standardized instruction in the performance of difficult and/or complex aircraft maneuvers.

(9) Malfunction Simulation. Malfunction Simulation (MS) enables the I/O to insert a failure, partially or totally, to a simulated aircraft component or to introduce an abnormal aircraft condition. When such a failure is inserted into the simulation, the consequences will duplicate the consequences of a corresponding failure in the aircraft and elicit trainee responses appropriate thereto. The purpose of the MS feature is to enable the I/O to simulate the occurrence of component malfunctions and failures so that the trainee may learn to determine that an abnormal condition has occurred, identify the condition, and take the prescribed corrective or compensating action.

(10) Automatic Malfunction Insertion. Automatic Malfunction Insertion (AMI) is an ATD instructional feature that automatically inserts malfunctions or failures of simulated aircraft components in response to previously selected conditions expected to occur during an instructional activity. These contingent conditions include events such as reaching a specified altitude or airspeed, passing a geographic position, releasing a weapon, exceeding time limit, or any combination of such events. The purpose of the AMI feature, in contrast to the nonautomatic Malfunction Simulation feature in which malfunctions are inserted manually by the instructor, is to cause selected malfunctions to be inserted automatically upon the first occurrence during a simulated flight of previously specified events.

IOS EVALUATION CONCERNS

As in all other aspects of ATD OT&E, the identification of operational deficiencies and determination of training capabilities are the chief evaluation concerns with regard to the IOS. In addressing those issues, however, it will be of use to identify two general classes of IOS evaluation concerns. The first has to do with the functional characteristics of the IOS relative to its use as an effective instructional tool. That is, does the IOS facilitate the job of the I/O, or does its design introduce unacceptable disruptions to the instructional process? The second evaluation concern is with the more traditional human factors engineering considerations as represented by standard reference documents such as MIL-STD-1472 [7] and the "Human Engineering Guide to Equipment Design" [8]. For both of these areas, the intent during an OT&E is to conduct the evaluation relative to I/O tasks as they are performed operationally. The dynamics of the I/O's tasks in using the IOS must be taken into account for a meaningful evaluation to be accomplished during OT&E. The following subsections address briefly the general nature of the defined areas of IOS evaluation concern.

Functional IOS Concerns

Assessments of IOS characteristics from the standpoint of instructional effectiveness and utility are of principal concern during ATD OT&E. Termed "functional" IOS considerations,[1] these assessments may deal with factors such as display informational content, I/O workload, or system inherent time delays that may be associated with I/O task performance. For example, the possibility that time to access a specific recorded segment utilizing the RECORD/PLAYBACK features might be excessive, as previously discussed, illustrates this area of assessment concern. Such a characteristic might be undesirable and negatively affect the instructing function.

The subjective and variable nature of these types of considerations necessitates that they be evaluated in the context of trainee task/flight task performance. The rationale for doing so is straightforward, and best communicated by example. Suppose that an evaluation of I/O workload is desired relative to a "training scenario/problem support" task. If the training problem to be supported is relatively uncomplicated and involves few variables, the I/O task may be likewise simplified and thereby introduce a very low level of workload.

---

[1] This terminology is not intended to suggest that "traditional" human factors considerations are nonfunctional, but rather to emphasize the need for ATD OT&E to give appropriate attention to instructional task functions.

However, if the training task involves numerous variables and is of longer duration and greater complexity, the I/O problem support task might then be more complex and demanding. Thus, the measure of I/O workload would vary as a function of the support requirements imposed by the specific trainee task in question. This example shows a fairly clear interrelationship and interdependency of trainee task and I/O task. In many instances, however, the connection may not be so obvious. For example, a complex trainee task might be supported with an automated IOS feature, thereby requiring very little I/O activity and effort to operate. Table 7-1 lists functional IOS concerns of interest in ATD OT&E.

TABLE 7-1. FUNCTIONAL IOS CONCERNS IN ATD OT&E

Operator Workload

- Activity Level
- Mental Effort Level
- Stress Level

System Inherent Time Delays

Control Task Actions

- Sequence of Action
- Error Frequency
- Ease of Error Detection/Correction
- Intra-Task Feedback

Displayed Information Sufficiency for:

- Trainee Monitoring/Evaluation
- Trainee Guidance (talk through)
- Identification of Errors/Developing Errors

Traditional Human Factors Considerations

Evaluation during an OT&E of the IOS with respect to "traditional" human factors considerations emphasizes those aspects of the IOS equipment and environment which impact its basic operation; e.g., control and display characteristics, anthropometry, standard console design, ingress/egress, environmental factors, and the like. As noted earlier, these traditional considerations are addressed in MIL-STD-1472, and as such are imposed as requirements to all contractors supplying equipment to the military.

During ATD Development Test and Evaluation (DT&E), tests may be conducted to determine adherence to those requirements; such tests typically are effected independently of the actual operational use (I/O tasks) of the device. For example, during DT&E, an evaluator would assess IOS writing surfaces with respect to the specific dimensional criteria as contained in MIL-STD-1472:

> Writing Surfaces - Where a writing surface is required on equipment consoles, it shall be at least 16 inches (400 mm) deep and should be 24 inches (610 mm) wide, when consistent with operator reach requirements.

In OT&E, in contrast, while the evaluator would likewise be concerned with writing surfaces, he would now be concerned principally in the context of their relevance to particular I/O task performance requirements. The point to note is that it would be possible to have IOS writing surfaces that comply with the Standard, but still do not fulfill specific I/O task requirements (e.g., access to maps, charts, grading forms and syllabi) as they relate to the instructing (as opposed to the writing) task functions.

Often, ATD OT&E is conducted in combination with DT&E to effect desired savings in test resources (time, cost, personnel, test equipment, etc.) [9]. In doing so, however, confusion can arise relative to the manner in which IOS human factors considerations are to be treated in concurrent and subsequent OT&Es. It is important to ensure that OT&E objectives do not become lost in the process. The measurement of IOS characteristics for strict compliance to quantitative criteria in a Standard has little direct relevance to the basic intent and objectives of operational test and evaluation, especially FOT&E. Using such an approach may give evaluators a feeling of test accuracy and comprehensiveness, but it may well not provide the kind of information desired from an OT&E, i.e., information that is explicitly I/O task performance related. A good example of correct emphasis in this regard is illustrated by the following excerpt from the SAAC FOT&E Final Report [10].

> The freeze feature was only used to terminate the flight scenario. The feature was not used by the IP for instructional purposes since it was controlled [only] at the console and thus involved a third party in the training equation. Comments

made by the IPs and the ACs in both groups indicated that the freeze capability would provide a valuable tool if its control button was available in the cockpit. Such an arrangement would allow the IP or student to freeze the simulator without delay when desired.

As this example illustrates, strict compliance with the Standard was not at issue. What was of concern for OT&E was whether or not the FREEZE control button was satisfactory relative to the specific I/O tasks it supports in the context of the operational training use of the ATD. Thus, what is of concern in OT&E with regard to IOS controls has to do with their accessibility to the instructor, their grouping with reference to instructional tasks, their size adequacy to carry necessary labels relating to their instructional function (e.g., CRASH/KILL OVRD or STORE CURRENT COND), and similar instructor task-related functions. Table 7-2 lists traditional human factors concerns for IOS evaluation during ATD OT&E.

Note: With reference to the listings of functional and traditional human factors IOS evaluation considerations shown in Tables 7-1 and 7-2, the relative length of the two listings should not be taken as indicative of the relative time and effort that will likely be involved. On the contrary, the execution of the functional IOS concerns evaluation typically poses much more of a challenge to the ATD OT&E team than does the traditional.

IOS EVALUATION METHODS

The preceding subsections have introduced a task model for the ATD instructor/operator, defined the general instructional function and use of device instructional features, and discussed two general classes of IOS evaluation concerns that should be addressed during an ATD OT&E. Given that basic framework and orientation, the present section treats specific approaches to IOS evaluation including definition of test objectives, data collection procedures and formats, and analysis and interpretation of results. The nature of IOS evaluation during OT&E may be generally described as involving collection of subjective rating data from relatively small groups of "evaluators." Those "evaluators" will generally consist of aircrews and instructor personnel highly experienced in the operation of the type of aircraft/weapon system that the subject ATD is designed to support. They are not necessarily trained or experienced in ATD evaluation. Therefore, the evaluation must be carefully planned and supervised so that desired objectives are met in an effective and efficient manner. Otherwise, the resulting evaluation may produce incomplete, unnecessary, or misleading results.

TABLE 7-2. TRADITIONAL HUMAN FACTORS CONSIDERATIONS IN ATD OT&E

CONTROLS

- Direction of movement
- Grouping/location
- Size
- Shape
- Action resistance
- Rapid operation
- Travel (displacement)
- Separation
- Positive indication
- Keyboard: layout, slope, height, relation to displays
- Accidental activation safeguards
- Joystick dimension, resistance, location
- Control labeling

VISUAL DISPLAYS

- Information: simplicity, format
- Location/arrangement: orientation, access, reflection, (glare), grouping, frequency of use, importance
- Viewing distance
- Legibility: character size, contrast, spacing

AUDIO DISPLAYS

- Unambiguous function
- Frequency
- Intensity: (too loud, not loud enough) volume control
- Signal/noise ratio
- Headset comfort
- Automatic/manual shut-off
- Voice communication system: speech intelligibility, noise, volume control

CONTROL/DISPLAY INTEGRATION

- Unambiguous relationship
- Functional group arrangement
- Access to more frequently used control(s)/display(s)
- Movement relationships: ratio, direction

ANTHROPOMETRY

- Control reach
- Display viewing distance
- Seat: vertical adjustment, backrest, cushioning, armrests, knee room, special positions
- Ingress/egress

WORKSPACE

- Kickspace
- Workspace: depth, width, height
- Storage space
- Panel slope
- Display placement

ENVIRONMENT

- Temperature comfort
- Ventilation comfort
- Humidity comfort
- Illuminance
- Ambient noise
- Vibration

The approach discussed here will enable efficient evaluation of IOS characteristics of greatest relevance to training. This approach advocates use of training scenarios wherein specific test activities are addressed. This type of approach allows concurrent evaluation of IOS functional and traditional human factors considerations in the context of specific training tasks.

Training Scenario Approach

This method is intended to provide the most realistic test of the IOS short of use in actual training. This method employs instructors and experienced aircrews who, in the interest of efficiency, role play as trainees (the experienced aircrews may also be instructors). These instructors and aircrews conduct a series of structured training scenarios. A skilled observer monitors the conduct of each scenario and collects data regarding the effectiveness and efficiency with which the system permits instructional functions to be performed. This method lends itself to use in selection of tasks for which the ATD is to be used for actual training. Note that a decision not to attempt to train a particular task or group of tasks in the ATD may be based not only on indications that the trainee station does not provide adequate cues and responses for the intended training tasks, but also on the fact that the system may lack sufficient instructional capability for the intended tasks and trainees.

A general flow of activity for the training scenario approach to IOS evaluation is presented in Figure 7-1. Each activity is treated further in associated paragraphs below.

Determination of objectives. The starting point for all evaluations should be a determination of objective(s), i.e., the goal(s) of the evaluation. Based upon the earlier discussion of IOS evaluation, two general objectives may be defined:

- To evaluate the IOS from the functional instructional standpoint.
- To evaluate the IOS from the traditional human factors standpoint.

The above general objectives do not, however, provide concrete bases for conducting the IOS evaluation. They are not specific enough: What specific I/O tasks are of interest, and in the context of which specific aircrew training tasks? Which functional and traditional human factors considerations are of interest? And so on. Under each general objective, then, a number of specific subobjectives must be explicitly defined. The format of those specific subobjectives should be similar to the following:

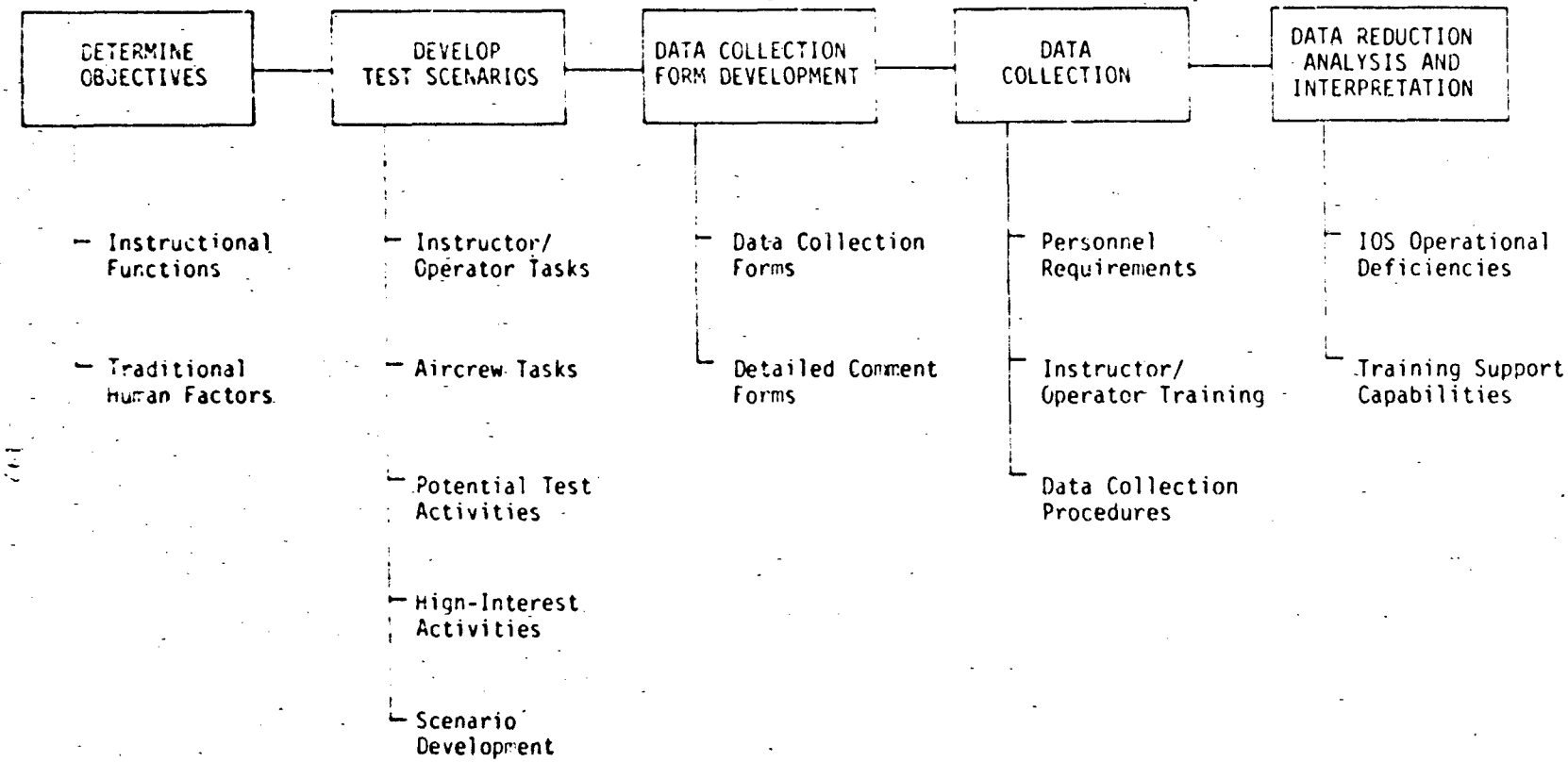


Figure 7-1. Training Scenario Approach to IOS evaluation during ATD OT&E.

Subobjective # : To evaluate IOS (evaluation concern) connected with (I/O task) in support of (aircrew training task).

Example 1:

Subobjective #1: To evaluate IOS displays/controls connected with Performance Monitoring/Evaluation in support of Formation-Training.

Example 2:

Subobjective #2: To evaluate I/O workload connected with Training Scenario/Problem Support for Low-Level Navigation Training.

Subobjectives such as those above can enable evaluators to focus much more precisely upon the specific elements of interest. A number of factors should be considered when developing evaluation objectives and subobjectives. These include the type of training to be conducted with the device, e.g., initial, transition, and/or continuation; the experience levels of the prospective device users (trainees), e.g., novice, recent UPT graduate, or highly experienced aircrews; the locus of instructional control, e.g., under control of an instructor at the IOS, or self-instruction in cockpit; and the amount and type of training to be given the I/Os prior to using the device, e.g., minimal training or a structured program of instruction.

It is unlikely, however, that time constraints for IOS evaluation during OT&E will allow every possible combination of I/O and trainee task to be evaluated. Therefore, it is necessary to identify those combinations of highest interest, and to design the evaluation training scenarios such that those high-interest activities are addressed in an efficient manner. Development of training scenarios for OT&E application involves five basic activities, and is described below.

Scenario development. Five activities are necessary in development of training scenarios:

(1) Identification of aircrew training task requirements

(2) Identification of I/O task requirements

(3) Determination of potential test activities

(4) Determination of high-interest test activities

(5) Development of test scenarios

The subsections which follow provide a detailed exposition of the content of each activity. Included are illustrative examples of the various steps involved.

(1) IDENTIFICATION OF AIRCREW TRAINING TASK REQUIREMENTS: The first step involves compiling a list of the aircrew tasks for which the ATD in question is to be utilized. This task list should be readily available, and be consistent with that used in evaluating the device training capabilities as discussed in preceding chapters. Figure 7-2 shows an example of this type of task list (the tasks listed in the example were taken from the F-5E IFS Test Plan [11]).

(2) IDENTIFICATION OF I/O TASK REQUIREMENTS: The second step in scenario development is to identify required Instructor/Operator tasks and/or subtasks. The I/O task categories described earlier should be used to organize I/O tasks into an appropriate task list. The result of this step would be a list of device-specific I/O tasks as well as general I/O task categories. The format of that list could follow the example in Figure 7-3.

(3) DETERMINATION OF POTENTIAL TEST ACTIVITIES: Not all I/O tasks will be applicable to all aircrew training tasks. For example, the I/O typically does not monitor weapons delivery when the trainee is performing an instrument task. To accomplish this step, it is useful to construct a matrix of potential test activities in a format such as shown in Figure 7-4. Once the matrix is constructed, each cell of the matrix should be examined logically and marked to indicate a potentially relevant combination of I/O and training task, as shown.

(4) DETERMINATION OF HIGH-INTEREST TEST ACTIVITIES: While all relevant IOS operations should undergo evaluation to the maximum extent possible, time and resource limitations that typically constrain an OT&E will make a sampling approach necessary. Thus, test planners will have to select those tasks which should receive principal emphasis during actual testing. This is not to suggest that any IOS operations should be totally eliminated--such would make the evaluation incomplete--but, rather, that the relative emphasis among possible activities must be predetermined so that those of highest interest are included in the resulting mock training scenarios.

AIRCREW TASK LIST

(Example)

INSTRUMENT PROCEDURES

Preflight/ground operations
Instrument ground checks
INS alignment procedures
Instrument takeoff (ITO)
Departure
Course interception
Performance monitoring of climb
Enroute TACAN navigation
TACAN holding
TACAN penetration
TACAN approach
Missed approach
ILS approach
Dragchute landing
Postflight/ground operations

TRANSITION TASKS

Preflight/ground operations
Departure
Arcing
Course interception
IFF procedures
Vertical "S"
Lazy eight
Steep turns
Unusual attitudes
Aileron rolls
TACAN point-to-point navigation
Maximum range descent
Straight-in approach
Missed approach

Figure 7-2. Example of aircrew training task list (partial).

INSTRUCTOR/OPERATOR TASK LIST

(Example)

- ATD set up
  - Power-on
  - Data clear
  - Mode select
    - Plot mode
    - Status mode
    - Program mode
  - Set initial conditions
- Training scenario/problem set up
  - Mission data set insertion
  - Malfunction select
    - Manual
    - Automatic
  - Set up gaming area
    - Radio nav stations
    - Threat (ground pos.)
- Training problem support
  - Communications
  - Air-to-air target control
  - Malfunction insertion
  - Environment modification
  - Weapons stores reload
- Instructional features
  - Store/reset
  - Record/replay
  - Hard copy
  - Freeze
  - Freeze override
  - Remote display
- Performance monitoring/evaluation
  - Maneuver scoring
  - Cockpit monitoring
  - Time/position monitoring
  - Procedures monitoring
- Training debriefing
- Safety monitoring
  - Emergency egress
  - Halon system
- Program/syllabus support
  - Auto-demonstration preparation
  - Target set preparation
  - Preprogrammed exercise preparation

Figure 7-3. Example of I/O task list format (partial).

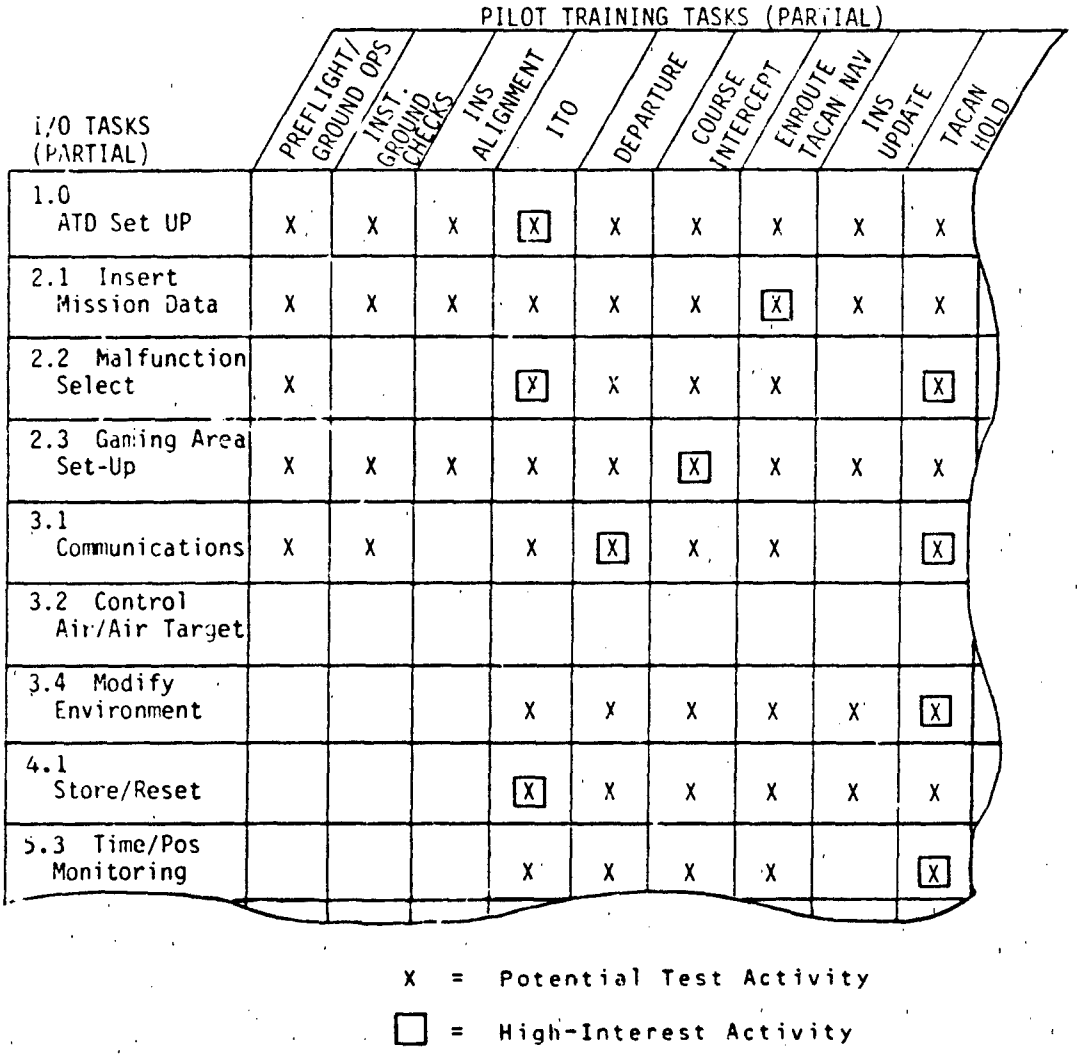


Figure 7-4. Illustration of selection of high-interest test activities.

To determine which I/O and training task combinations should be selected from those identified in the matrix, a rationale for selecting those tasks in terms of their criticality to ultimate ATD utilization must be established. The following is a list of types of criteria that may be employed for this purpose:

- Frequent Tasks: Tasks that will be performed frequently in the course of training are candidates for emphasis. An infrequently performed task may, however, be crucial to accomplishment of a particular training mission; if so, it may be selected as a "special purpose" task.

- Routine Tasks: Routine tasks, by their nature, will often be selected for inclusion in the scenarios.

- Difficult Tasks: Difficult I/O tasks should be represented. They would include those which require complex action sequences, multiple display integration, fine control manipulation, concurrent operations, etc.

- Effect of Task Error or Failure: Those tasks which, if performed incorrectly, lead to significant training disruptions, or are difficult or time consuming to recover from (excessive downtime) are candidates for emphasis.

In addition to test activities that are selected by these criteria, it would be desirable that representative tasks from all I/O task categories be included at least once in the interest of test comprehensiveness. Figure 7-4 illustrates selection of high-interest activities.

(5) DEVELOPMENT OF TEST SCENARIOS: Once the high-interest activities have been selected, it is necessary to combine those activities into a limited number of training scenarios. This development will require assistance from a subject matter expert in the aircraft/weapon system of interest (e.g., personnel from the appropriate MAJCOM ISD). Training scenarios should be constructed such that each corresponds to a distinct area or grouping of aircrew tasks. Task areas such as instruments, air-to-air weapons delivery, low-level navigation, basic aerobatics, etc., should be kept relatively separate from one another when constructing the training scenarios. In this way, the relative capability of the TOS to support training in these varied task groupings can be assessed. Examples of training scenarios are provided below:

INSTRUMENTS

SCENARIO #1
(Sample)

TRAINEE TASKS	I/O TASKS
	Perform ATD set up
Execute TACAN HOLD	Initialize at 10,000 in TACAN HOLD
	Store conditions
	Insert malfunction to ECS system
Maintain holding pattern and execute emergency procedure for ECS	
	Monitor procedures
	Monitor time/position
	Evaluate holding pattern and emergency recovery
	Remove malfunction
Execute TACAN approach	Role play approach control
	Modify environment (Lower ceiling to ground level)
	Monitor procedures
Execute missed approach	Monitor situation when trainee reaches decision height
	Reset to holding pattern
	Repeat as necessary
	End scenario

AIR-TO-AIR

SCENARIO #2
(Sample)

TRAINEE TASKS	I/O TASKS
	Perform simulator set up
	Initialize simulator 3000 ft. in trail on prerecorded AC (air-to-air armament)
	Select crash/kill override on
Pursue and close for a gun attack on a prerecorded aircraft performing a constant 3G turn	
	Monitor position and closure
	Monitor tracking stability
	Monitor weapons delivery accuracy
	Freeze
	Initialize simulator 6000 ft. in trail on prerecorded AC
	Select crash/kill override off
Pursue and close for an Aim-9 attack on a prerecorded aircraft performing a 3G turn	
	Monitor maneuvering and A/A missile results
Close for high angle gun attack	Monitor maneuvering and high angle gun results
	Freeze
	Initialize simulator 12,000 ft. in trail on prerecorded AC

AIR-TO-AIR

SCENARIO #2
(Sample)

TRAINEE TASKS	I/O TASKS
Maneuver for front quarter missile attack on a prerecorded aircraft performing a constant 4G turn	
	Monitor maneuvering and missile results
	Freeze
	Initialize simulator in perch position on joystick controlled AC
Attack a reactive target	Fly joystick AC defensively, attempting to overshoot attacker and reverse, or attempting to separate
React to countermeasures	
	Freeze
	Reset initial condition and perform second trial
	Critique results
	End scenario

Data collection form development. The need for some type of structure in data collection for this type of approach cannot be overemphasized. The fact that different scenarios will encompass different sets of I/O tasks necessitates that the evaluators have a means to identify the areas of test emphasis within each. Further, when it becomes necessary to prepare the OT&E report, having test data documented in a standard format will make those data easier to compile, analyze, and interpret than if one is forced to depend upon memory alone.

There are two data collection modes required in carrying out the training scenario approach to IOS evaluation. The first mode occurs during the active real-time conduct of training scenarios; the second mode occurs, logically, during those periods between and following active use of the IOS. The sheer number of IOS evaluation concerns listed earlier in Tables 7-1 and 7-2 makes their use practically infeasible, however, during the real-time data collection mode. Accordingly, data collected during the real-time mode must deal with those aspects of I/O task performance that can be observed and recorded quickly and, at the same time, are meaningful within the context of the total IOS evaluation. That is, what data are collected in real-time must be traceable to one or more specific areas of the functional or traditional human factors concerns. In addition to the fact that the number of real-time evaluation areas must be limited, the type rating required for each must be very simple and straightforward. Four elements to be addressed during the real-time mode are defined below:

OPERATOR ERROR: Any mistake in control input that impacts adversely the logical and smooth flow of desired instructional events. (For example, input of incorrect training/problem set up parameters; incorrect sequence of IOS control inputs; and incorrect IOS control procedures resulting in disruptions to the training process.)

INPUT TIME: Time required to set up a desired instructional event following correct or incorrect input procedures.

SYSTEM TIME DELAY: Time required for the ATD system to respond to desired IOS control inputs.

TRAINEE PERFORMANCE INFORMATION ADEQUACY: Trainee performance information (what the trainee is doing as he does it) should be supplied at the IOS for a number of I/O tasks. The adequacy of that information to support required trainee guidance/monitoring/evaluation is a key factor in the usability of the total system. Also, this area can pose significant levels of I/O workload when performance information must be integrated from a number of sources.

The rating scale suggested for use during the real-time data collection mode in the above four areas is a two category (acceptable/unacceptable) rating as defined below.

Rating

A	=	Acceptable	= No operator errors; input time reasonable; system time delays reasonable; trainee performance information adequate.
U	=	Unacceptable	= One or more operator errors; input time excessive; system time delays excessive; trainee performance information inadequate.

Figure 7-5 illustrates the type form appropriate for the real-time data collection. This data collection instrument has three basic regions as described below.

(1) HEADER: Header information includes basic identifying information including Date, Device, Scenario, Trial, Aircrew, Instructor/Operator, and Evaluator.

(2) TRAINING SCENARIO TASK SEQUENCE: The left half of the form contains columns to indicate the aircrew tasks, I/O tasks, and the relative sequence of instructional events associated with the desired training scenario. Also, an "X" column is used to denote the specific I/O tasks to be rated in the evaluation.

(3) RATING COLUMNS: The right half of the form contains columns corresponding to the real-time evaluation areas identified above.

In addition to the four areas of concern discussed for real-time rating, a column for recording workload ratings is included. While it is not a strict requirement, it would be suggested that this rating (discussed below) be collected as near real-time as possible within the constraints of the training scenario. In some cases, it will be possible to obtain this rating concurrently, while at other times it will be necessary to wait until the termination of the training scenario to obtain the required estimates of I/O workload.

WORKLOAD RATING: A rating of I/O workload can provide a meaningful index relative to the effectiveness and efficiency of IOS design to support I/O tasks required in conducting ATD training. From the standpoint that the ATD I/O is an "instructional process manager," it

IOS EVALUATION: DATA COLLECTION FORM							Page _ of _
DATE	DEVICE		SCENARIO			TRIAL	
AIRCREW	I/O		EVALUATOR				
TRAINEE TASK	I/O TASK	X	OPER. ERROR	INPUT TIME	SYSTEM DELAY	PERF. INFO.	WORK-LOAD

Figure 7-5. IOS evaluation data collection form.

is understandable that the design of his interface with the instructional system, namely the IOS, can significantly impact his effectiveness in that role. Certainly, an IOS that imposes high levels of confusion and strain upon the I/O in order to effect desired instructional events is less desirable than one which does not do so. Conversely, a design that automates too much of the instructor's task such that he is effectively taken out of the instructional loop is also undesirable. ATD IOS design must be such that the I/O remains an active in-the-loop instructional process controller and decision-maker, and, at the same time, be one that alleviates those task requirements which impose unwanted and unneeded levels of operator workload. In most instances, of course, low to moderate levels of workload are desired. The rare case where "too much" automation has been designed-in must be considered independently of the workload rating itself.

Before describing the suggested I/O workload rating scale, it will be useful to consider what factors can contribute to high levels of I/O workload. The term "workload" can be defined in many ways. For example, "difficulty," "stress and strain," "activity level," "fatigue," "perceptual load," "information load," and other terms connote different aspects of the operator's workload involved in performing a task. In recent work,[1] three attributes of operator workload have been suggested, each of which is believed to contribute independently to the operator's task. These are:

- FRACTION OF TIME BUSY: Portion of time during task performance actively doing some thinking, in a functional sense, on the task; ranges from doing essentially nothing to fully occupied.

- INTENSITY OF THINKING/INFORMATION-PROCESSING: The mental effort involved in performing the task; ranges from completely automatic to extreme effort and concentration.

- INTENSITY OF FEELING: The amount of anxiety or stress associated with task performance; ranges from relaxation to severe frustration, confusion, and stress.

---

[1] This work was conducted by Sheridan and Simpson at the MIT Flight Transportation Laboratory, and was directed to investigation of workload factors involved in IFR piloting tasks. The basic approach suggested by that work to conceptualizing and rating operator workload for IFR piloting has been adapted here as appropriate for ATD instructor/operator station evaluation purposes.

Any one of the above factors can by itself, or in combination with others, contribute to an unacceptable or impossible level of perceived operator workload in performing various I/O tasks. Design of the IOS, its physical and functional characteristics and features, should be such as to minimize adverse effects in these areas. In addition to IOS design factors, the training given to the I/Os participating in the evaluation can influence the workload rating. For example, an improperly trained I/O evaluator may not have learned correct IOS operation and control procedures, and may, in turn, experience confusion (and high workload) in attempting to perform certain I/O tasks. In this instance, what may at first appear to have been an IOS design problem may, in reality, be related to an instructor/operator training deficiency. This matter of attributing high workload ratings to either an IOS equipment design deficiency or an I/O training deficiency is discussed more fully in the later section dealing with interpretation of test results.

A suggested workload rating scale is shown in Figure 7-6. This scale is an adaptation of a Cooper-Harper type scale and considers I/O workload along a single dimension. Ten levels of I/O workload are defined, with a "10" indicating an "impossible" level of workload (meaning a level of operator workload so high as to make impossible performance of the required task). Scale points "1" through "9" indicate successively greater levels of I/O workload going from low (1) to high (9).

DETAILED COMMENT FORM: In addition to the real-time data collection form, detailed comment forms are needed for a comprehensive evaluation (Figure 7-7). These forms are used to record the more specific details as to the nature of any operational problems or deficiencies encountered. The information to be recorded on the detailed comment form is as follows:

- I/O task
- Specific IOS component in question (hardware component, display page, etc.)
- Description of the problem (Tables 7-1 and 7-2)
- Estimate of the nature of the problem; i.e., equipment problem or training problem, and suggested solution
- Judged priority of the problem (e.g., High, Medium, Low) relative to the overall training utility of the device, and the need for some type of modification or fix.

Data collection. The process of data collection must be such that the desired information may be acquired efficiently. Development

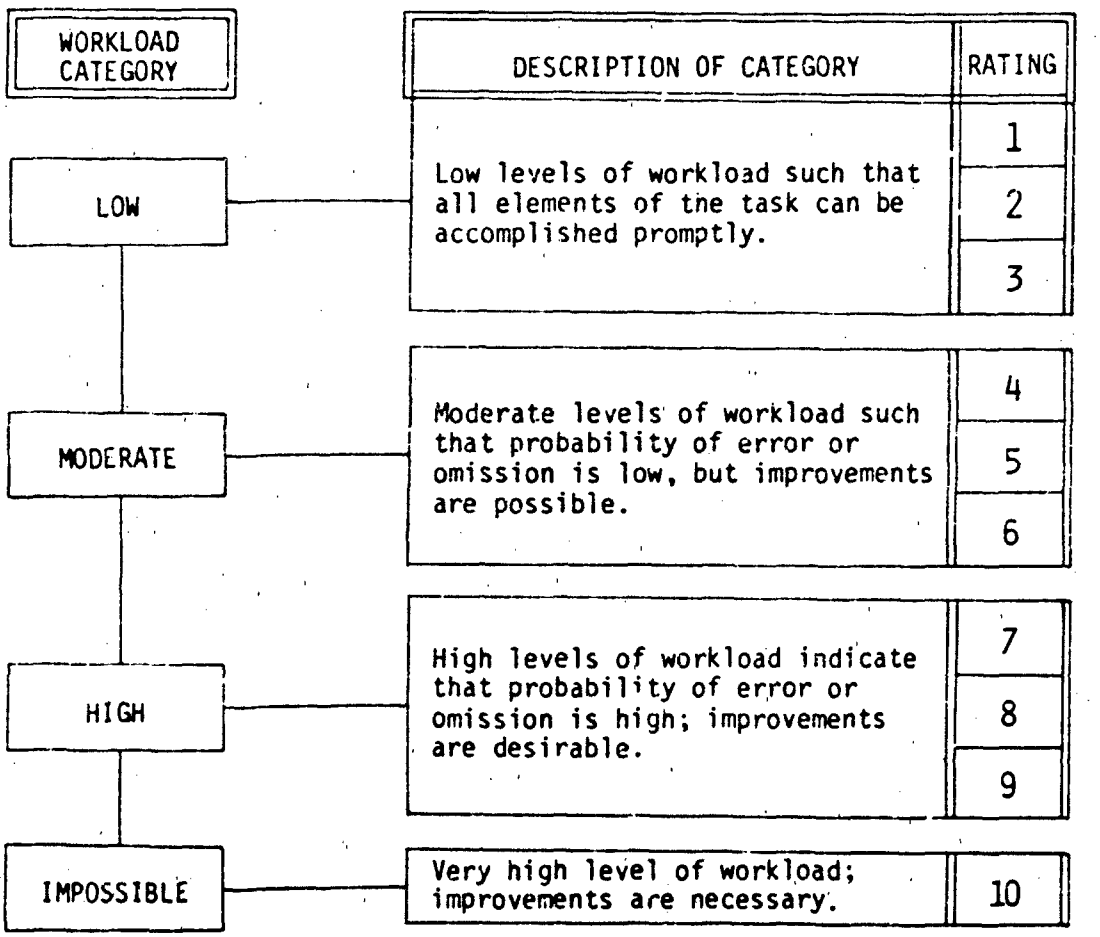


Figure 7-6. I/O workload rating scale for IOS evaluation during ATD OT&E.

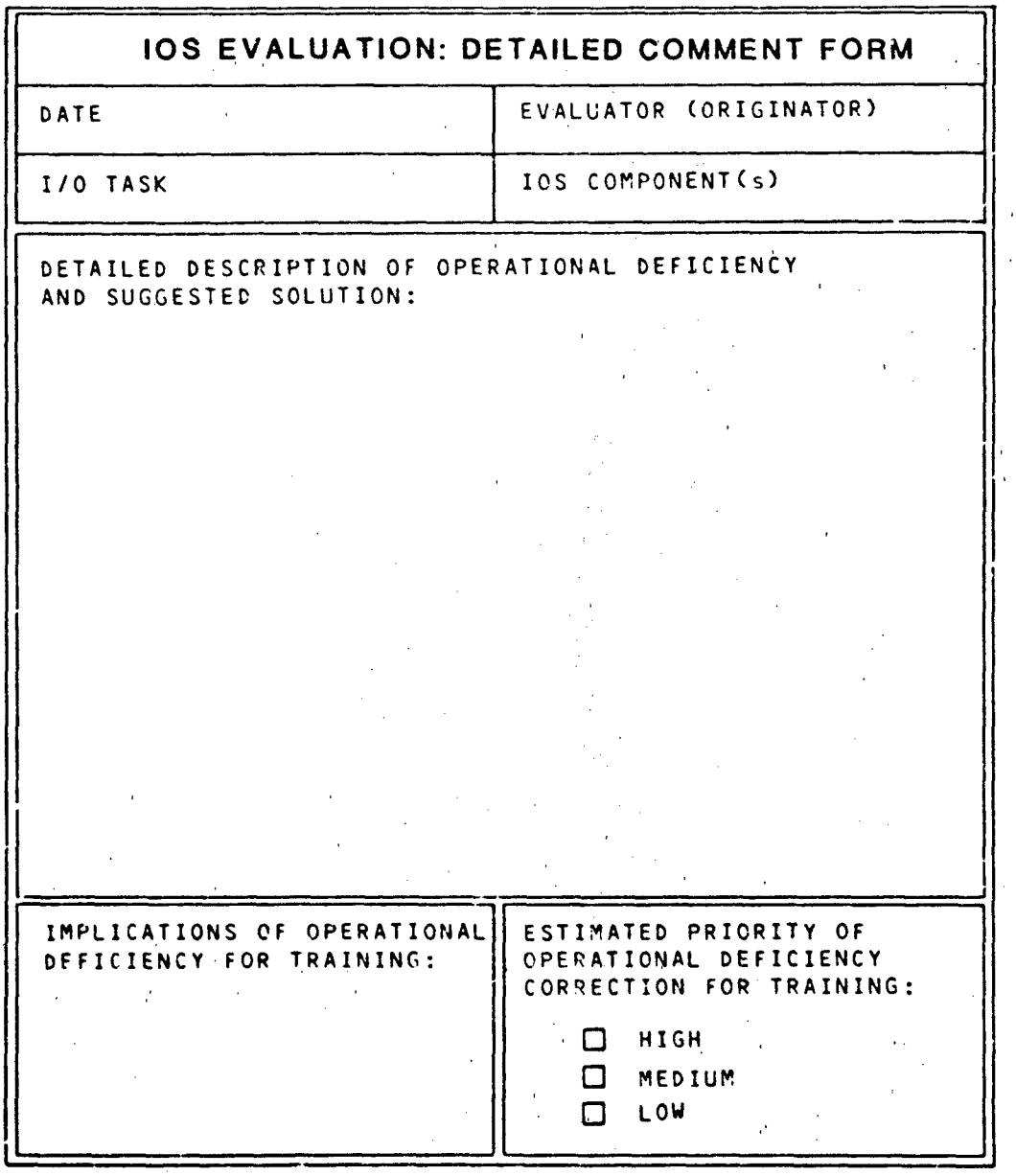


Figure 7-7. IOS evaluation detailed comment form.

of structured data collection instruments such as those described is a good starting point in this regard. The manner in which those instruments are utilized, however, can affect greatly the quality of the resultant data. As should be clear at this point, the training scenario approach to IOS evaluation stresses the operational training use of the total ATD "system"--trainee station, IOS, aircrew, and Instructor/Operator--rather than measuring adherence to predefined standards and human factors criteria intended for application outside the context of actual I/O task performance. As such, personnel requirements, I/O training, and data collection procedures are more extensive than would be the case with the more surface-level type evaluation. Each of these areas is discussed below.

PERSONNEL REQUIREMENTS: Personnel requirements will vary with device complexity and configuration; however, a minimum of three individuals are required to effect the basic data collection procedure:

- Two representative users (pilots, IPs, etc.)--one acting as the ATD "I/O"; the other acting as the "Trainee."
- One human factors evaluator--to manage the test, observe I/O activity, and make all entries to data collection forms. This individual should be knowledgeable about the purposes of the evaluation, all aspects of test scenario development, and data collection procedures.

As noted, the total personnel requirement will be a function of the complexity, configuration, and training uses of the device in question; for example, a full mission ATD that may be used to train a wide range of aircrew tasks across a number of task categories (e.g., instruments, air-to-air weapons, air-to-ground weapons, low-level navigation) will require a corresponding wide range of training scenarios to evaluate fully. On the other hand, a special task trainer to be used to train a single aircrew task or a small number of related tasks (e.g., air refueling trainer) would require a relatively limited number of training scenarios in order to exercise the full capability and range of IOS features. In addition to the basic number of training scenarios to be run in the evaluation, the total clocktime required to perform those scenarios must be taken into account relative to the number of I/Os and trainees to be needed. It is recommended that, at a minimum, each training scenario be performed by three (3) sets of I/Os and trainees. Of course, the same three sets of subjects might be able to perform all training scenarios, depending upon the total length (clocktime) of the various scenarios. Otherwise, additional sets would be required.

INSTRUCTOR/OPERATOR TRAINING: Pre-test I/O training may range from a "checkout" of the IOS consisting of basic familiarization with controls and displays, to a comprehensive program encompassing

training in instructional technology, and how and when to employ device instructional features. In any case, it is desirable in the present context to ensure that the prospective I/Os are trained such that the IOS evaluation results are not overly confounded by inadequate I/O training.

DATA COLLECTION PROCEDURES: Three sequential activities are required for each training scenario trial. The first activity during data collection involves a briefing (review and discussion) of the specific scenario to be conducted. During this preliminary step the Evaluator and I/O review together the specific elements (I/O tasks) to be evaluated during performance of the training scenario.

The second data collection activity occurs during conduct of the actual training scenario. During this period, the Evaluator observes I/O behavior and makes note of apparent errors and operational difficulties on the data collection form. The I/O carries out the planned scenario without attempting to record concurrently any evaluative data which might, in itself, introduce operational difficulties and disruptions to the training scenario. The I/O is instructed to note verbally to the Evaluator any difficulty that he encounters during the scenario. Also, if possible within the context of the training scenario, the Evaluator may request that the I/O provide workload ratings for the high-interest I/O tasks as they are performed. If this is not possible (i.e., it would disrupt the continuity of the scenario), this rating should be obtained immediately following the scenario.

The third activity occurs following completion of the scenario, during which time the data collection form is verified and completed. Ratings on the form will have been made to indicate any operational difficulties encountered and I/O workload ratings. Based upon the ratings obtained during this trial of the training scenario, detailed comment forms should be prepared. A detailed comment form should be generated for any "unacceptable" rating indicated on the data collection form. ("High" and "Impossible" I/O workload rating categories, Ratings 7-10, should be accompanied by a completed comment form which describes the specifics of the problem encountered.) Of course, detailed comment forms describing any IOS operational deficiency may also be prepared at this time, even if it pertains to an I/O task that was not specifically to be rated (an "X" item) during the scenario.

The above three data collection activities are repeated as necessary until all test scenarios have been performed and evaluated. Several factors should be taken into account when scheduling personnel and test scenarios to maintain quality data.

- Each grouping of test scenarios should encompass no more than 90 minutes, which includes the setup, trial, and post-trial data collection periods for each training scenario (see Figure 7-8).

- If possible, different I/O and trainee subject groups should be employed for each successive 90 minute period of testing to minimize fatigue and boredom effects.

- An hypothetical arrangement of twelve training scenarios and three subject groups is shown in Figure 7-9. This arrangement achieves the desired three trials for each scenario to minimize rater fatigue. Overall, three days of testing would be required.

Data reduction, analysis and interpretation. Once all training scenarios have been run and associated IOS data collection and detailed comment forms have been completed, what does all of this mean? How does one now proceed to derive some meaningful conclusions about the ATD IOS in question, and its capability to support the anticipated training use of the device? In what areas does the IOS provide a highly facilitative training tool? And, in what areas does the IOS not facilitate the job of the ATD Instructor/Operator? The extent to which these questions can be answered satisfactorily will, in effect, be a direct result of the quality of data collected. Poorly collected data, incomplete data, and otherwise marginal data cannot be transformed easily into meaningful results and conclusions. On the other hand, data which have been collected following the procedures described herein should be relatively straightforward to reduce and interpret in a meaningful way.

Basically, there are two areas for which data from the training scenario IOS evaluation are to be used. The first concerns any specific IOS operational deficiencies which are identified in the process. These deficiencies would relate to one or more of the IOS evaluation concerns identified earlier in Tables 7-1 and 7-2 (for example, an ambiguous relationship between an IOS control and its intended function). The second area of interest has to do with the relative capabilities of the IOS for supporting effective training across the range of aircrew tasks to be trained in the device. Reduction, analysis, and interpretation of data to serve these two areas are discussed below.

IOS OPERATIONAL DEFICIENCIES: An IOS operational deficiency can be anything related to its features and design characteristics which may impair or otherwise limit its effective utilization in an active training (i.e., operational) mode. The basic structure of the data collection process in the training scenario approach to IOS evaluation is primarily directed to identifying operational deficiencies. For

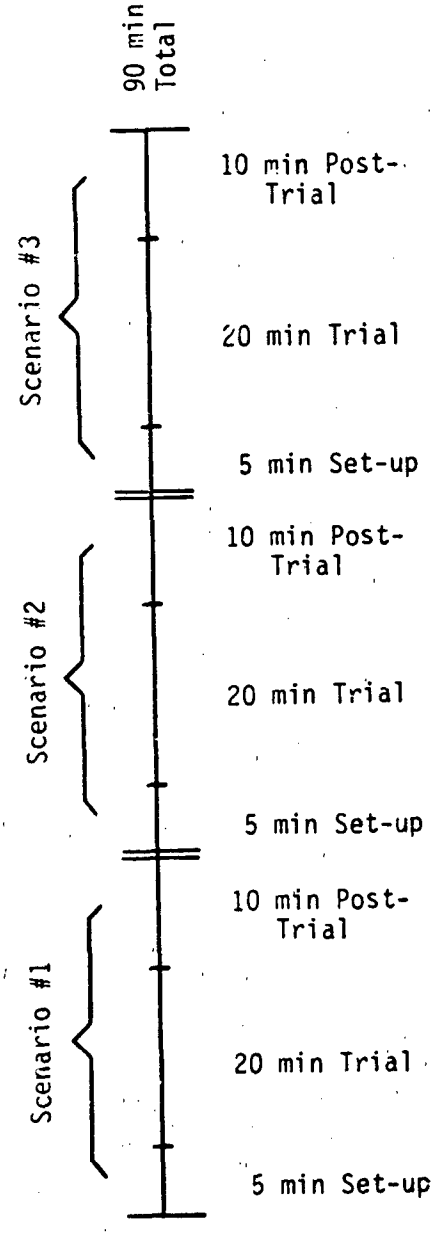


Figure 7-8. Scheduling of test training scenarios for IOS evaluation.

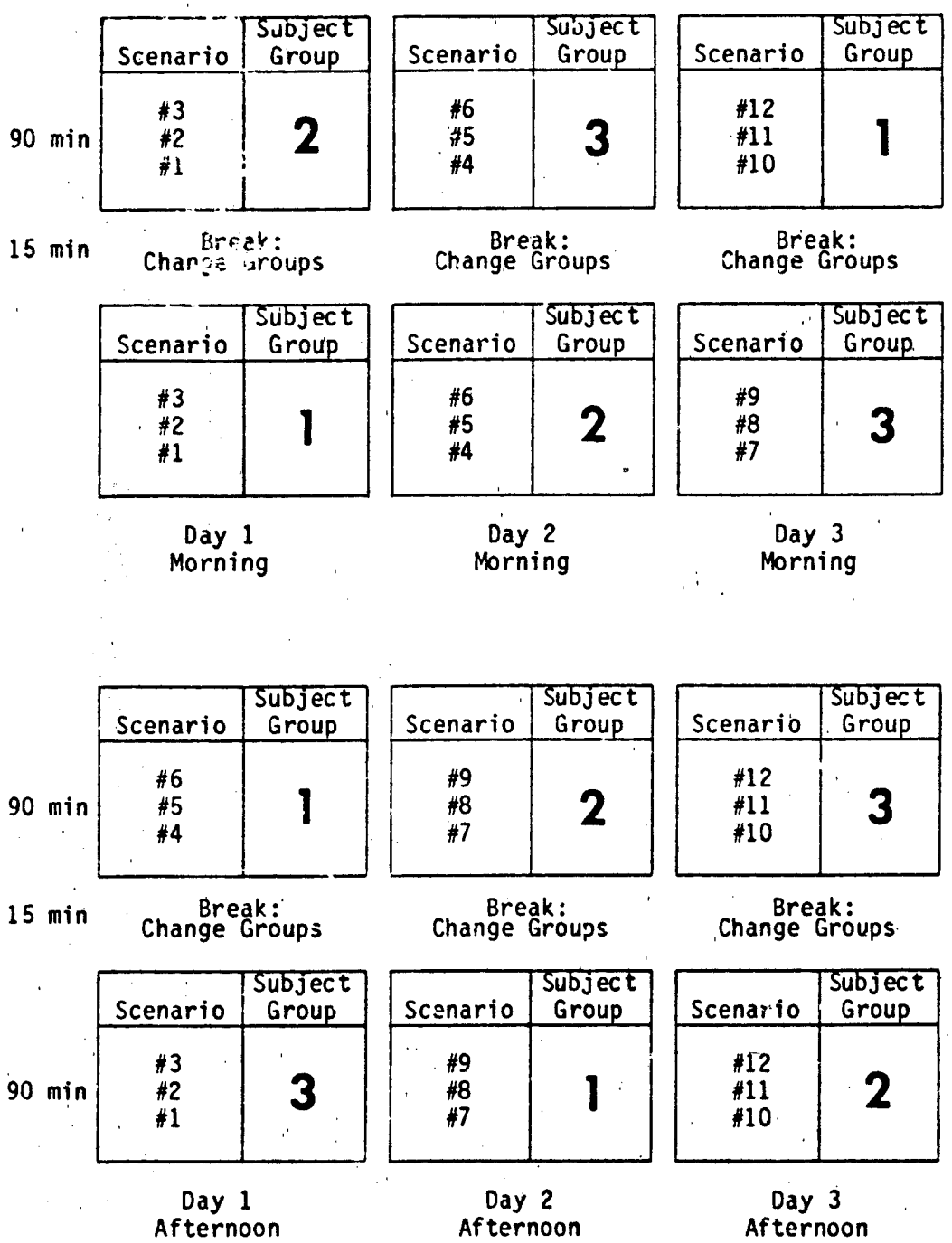


Figure 7-9. Scheduling arrangement for twelve training scenarios and three subject groups (I/O and trainee).

example, data collected will pertain to factors such as operator error, system input time requirements, system time delays, and adequacy of trainee performance information as they relate to the active instructional process and associated I/O task requirements. Thus, analysis of data for this area of concern is relatively straightforward.

The first step is to summarize, for each training scenario, data acquired from the subject groups on their various respective trials. This is accomplished by counting the number of "acceptable" responses in each rating area. For example, if for I/O task X there were 2 "acceptable" ratings and 1 "unacceptable" rating on I/O input time, the total would be 2/3, i.e., 2 acceptable ratings out of the 3 total ratings. This tabularization may be readily done by recording the various counts on a blank training scenario data collection form (one that shows the training scenario, but which has not been used to collect data). Also, the mean workload rating should be recorded. Figure 7-10 depicts a data collection form as used for this purpose.

The next step is to bring together the completed detailed comment forms for each scenario which contain the more specific details and descriptions as to the nature of any operational deficiencies found. For each I/O task evaluated in the scenario, extract consistencies among the data and prepare a summary list of specific comments relating to that task. Each specific comment should contain a clear description of the operational deficiency encountered, the specific IOS components in question (hardware components, display page, etc.), and, if possible, a suggested solution to the deficiency. In addition, the judged priority of the problem (high, medium, low) relative to the overall training utility of the device should be indicated. Figure 7-11 shows a suggested format for displaying this detailed information relating to IOS operational deficiencies.

In interpreting these data, it should be remembered that an operational deficiency may be more a function of inadequate I/O training in the use of the system than a function of inherent IOS design or function characteristics. In some cases, an improvement in I/O training may alleviate the operational deficiency, while in other instances increased training and experience with the device may have no significant beneficial effects; e.g., the basic design of the IOS in a particular task area may go against the logic and sequence of operation in other task areas, may be overly complex, or may be contrary to previous experiences. A determination must be made, therefore, in analyzing and interpreting these operational deficiencies, as to the possible confounding effects of the training administered to the subject I/Os prior to the IOS evaluation.

In documenting test results in this area, a number of alternate formats may be considered than that suggested above. However, the

IOS EVALUATION: DATA SUMMARY

Page 1 of 2

DATE:	DEVICE:	SCENARIO:	TRIAL: (SUMMARY)
00-00-00	ATDX	Instruments #1	1,2,3

AIRCREW:	I/O:	EVALUATOR:
1,2,3	1,2,3	George A. Doe

TRAINEE TASK	I/O TASK	X	OPER. ERROR	INPUT TIME	SYSTEM DELAY	PERF. INFO.	WORK-LOAD
	ATD Set UP						
Execute Tacan Hold	Initialize at 10K Tacan Hold						
	Store Conditions	X	3/3	3/3	2/3	3/3	2
	Insert Malfunc-tion to ECS						
Maintain Hold and execute EP for ECS	Monitor Procedures						
failure							
	Monitor Time/Position	X	3/3	3/3	3/3	2/3	3
	Evaluate Hold-ing Pattern and						
	ECS Recovery						
Execute Tacan Approach	Role Play Ap-proach Control						
	Modify Environ. (lower ceiling)	X	2/3	2/3	3/3	3/3	5
	Monitor approach Procedures						
Execute Missed Approach	Monitor Sit. when reach DH						
	Reset to Hold-ing Pattern	X	2/3	2/3	3/3	3/3	4
Execute Tacan Hold	Repeat as necessary						

Figure 7-10. IOS evaluation data summary format.

SCENARIO:	I/O TASK(X)	DESCRIPTION OF OPERATIONAL DEFICIENCY:
IOS COMPONENT(s) IN QUESTION:		
SCENARIO:	I/O TASK(X)	DESCRIPTION OF OPERATIONAL DEFICIENCY:
IOS COMPONENT(s) IN QUESTION:		
SCENARIO:	I/O TASK(X)	DESCRIPTION OF OPERATIONAL DEFICIENCY:
IOS COMPONENT(s) IN QUESTION:		

Figure 7-11. IOS operational deficiency summary form.

format must consider the guidance contained in the approved test plan and must be responsive to the test objectives identified. Results may be documented also in a separate "IOS human factors test report," or they may be integral to the basic OT&E test report. Within these constraints, results may be organized in a variety of ways. One method would be to report equipment related deficiencies separately from training deficiencies. Then, under each, list findings in order of relative priority, with highest priority items listed first. Results of the IOS evaluation in this area may be of use in effecting modifications to improve the training utility of the tested device as well as that of future production versions. Documenting results in this way can enhance their use for that purpose. Items which are corrected/modified during the OT&E (e.g., minor software problems) should also be included in the report along with a brief description of the modification.

IOS TRAINING SUPPORT CAPABILITIES: The second area of interest has to do with the capabilities of the IOS for supporting effective training across the range of aircrew tasks to be trained in the device. Analysis and interpretation of the data with regard to training support capabilities provided by the IOS must be largely rational in nature, i.e., analyses based upon a logical examination of what occurred during the evaluation with respect to the training impact of IOS functions and design characteristics. The functional IOS concerns identified explicitly in Table 7-1 are of primary interest in this regard, as well as are a number of implied questions that should be considered. The kinds of questions that should be considered in this regard are such as listed below:

- Do IOS control actions promote uninterrupted training activity?
- Are repeated attempts required to accomplish a desired control task?
- How frequently does one control error lead to another?
- Are errors more likely during certain operations or when using certain types of displays?
- Are the procedures consistent across controls, displays and tasks?
- Are system-inherent time delays so long that they discourage use of certain instructional features?
- Is there sufficient information about trainee performance to give the trainee instructional guidance and feedback?

- Is trainee performance information presented in a usable format?

- Must trainee performance information be integrated from an excessive number of sources?

These questions are indicative of the type which should be addressed with regard to IOS training support capabilities. They suggest ways of examining the data collected so that meaningful conclusions can be drawn and stated. As discussed earlier, the training scenario approach to IOS evaluation stresses the assessment of I/O task performance in the context of specific aircrew training tasks. Having data collected in this way thereby enables the desired analyses and interpretation of results to be accomplished readily. A suggested procedure and format for doing so is discussed below.

Each training scenario executed during the IOS evaluation should have been directed to a logical area or grouping of aircrew tasks (e.g., instruments, low-level navigation, air-to-ground weapons delivery), and thus provides a built-in hierarchical structure relative to the evaluation of IOS training support capabilities. That is, by summarizing the IOS training support capabilities as they relate to the specific aircrew tasks within a particular scenario, an "overall" IOS support capability may be derived for the aircrew training domain represented in that scenario. Thus, two levels of IOS training support capability may be addressed--the first level relating to specific aircrew tasks, and the second level relating to the more general training scenarios.

One method for accomplishing this analysis involves making percentage (%) estimates of the training support capability provided. Percentage estimates are attractive for a number of reasons: They can be easily interpreted relative to the objectives of the evaluation, i.e., people are accustomed to dealing in percentiles; they represent ratio level data[1] and, therefore, may be further summarized and analyzed with measures of central tendency (e.g., mean, median, mode) and variability (e.g., average deviation, standard deviation); and, they can be readily compared and contrasted to trainee station training capability data that will have been obtained by either rating scale or transfer of training approaches. Of course, the assignment of a specific percentage estimate for an evaluated I/O task will require the relative importance of various factors to be taken into account.

---

[1]See Appendix A for an explanation of the various types of measures and appropriate descriptive statistics.

Endpoints (i.e., 0%/100%) of the percentage estimates, however, can be based upon some relatively clear guidelines. A 100% value would be assigned if no operational deficiencies were found that impacted performance of the I/O task in question. A 0% value would be assigned if the I/O task could not be performed. The baseline data for these estimates are the previously discussed IOS Operational Deficiencies. In addition to the baseline IOS deficiency information, the implied questions listed earlier should also be taken into consideration.

It is likely that most I/O tasks will fall somewhere between the 0% and 100% endpoints. In these cases, it will be necessary to base the estimate of training support capability upon the relative importance of any operational deficiencies found, and reduce the percentage accordingly. Figure 7-12 illustrates a suggested format for displaying IOS training support capability estimates.

For the second level of IOS training support capability estimates, the procedure is as follows. The intent here is to provide an index relating to the capability of the IOS for supporting training in the training scenarios as a whole. This value is generated by taking the mean of the values for each I/O task within the scenario in question. For example, suppose an air-to-ground training scenario included five I/O tasks with training support capability ratings of 60%, 75%, 55%, 87%, and 80%. The training scenario value could be the mean of these ratings, or 71% (rounded). By generating training support capability ratings across the various training scenarios in this way, it may be possible to make comparisons and draw meaningful conclusions regarding the most effective training use of the ATD.

Alternate Approaches

The mock training scenario approach described above, properly planned and executed, should allow a comprehensive and efficient evaluation of the IOS. Alternate approaches to IOS evaluation may be developed during OT&E that can provide some useful information, but which are not as systematic or sophisticated in approach.

One alternate approach would be the use of some type of human factors checklist procedure. This approach is frequently used in traditional human factors evaluations. An example of a human factors checklist is shown in Figure 7-13 [12]. While far less sophisticated than the training scenario approach, the use of a human factors checklist, when properly employed, can allow identification of certain major IOS design deficiencies.

Human factors checklists can be found in a number of reference sources. For the most part, however, such existing checklists, including the one in Figure 7-13, have been developed to be used in

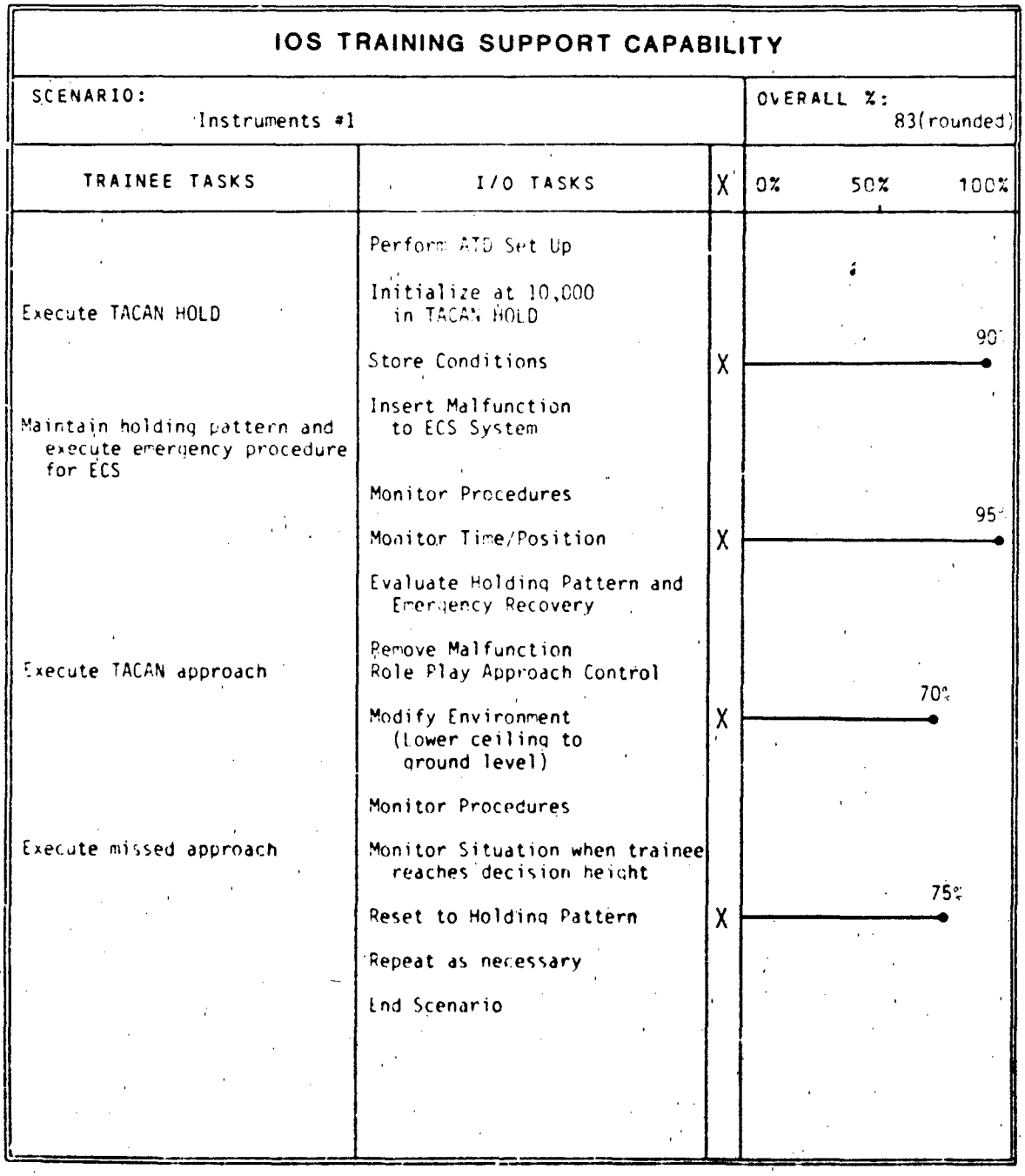


Figure 7-12. Format for IOS training support capabilities summary.

Figure 7-13, page 1 of 4.

Response
(S, U, NA)

DESIGN AREA 1 -- HOUSING ARRANGEMENTS

_____ Size and shape of student and instructor areas. Sufficient space should be allocated to provide for variation in human body sizes. Arrangements should reflect the need for movement between and within areas, the size and method of operation of equipment, and communication requirements.

_____ Traffic flow. Utilization of floor areas should follow from traffic requirements. Sufficient aisle and corridor space should be provided, and entrances, stairs, and ladders should be designed according to accepted standards.

_____ Maintenance. Arrangement of men and equipment should consider the need for and location of maintenance areas. Sufficient aisle and corridor space should be provided, and entrances, stairs, and ladders should be designed according to accepted standards.

_____ Safety. Facilities should be arranged with consideration for the location and movement of men relative to potentially dangerous equipments (high voltage, dynamic) and housing features (steps, ladders, etc.).

(Continued)

Figure 7-13. Human factors design checklist.

Figure 7-13, page 2 of 4.

Response
(S, U, NA)

DESIGN AREA 2 -- WORKSTATION LAYOUT

_____ Principles and criteria of good workplace design. Arrangement of equipment and associated controls and displays should take the following into account: frequency of operation, sequences of operation, functional relationships, the importance of presentation of and practice on the task problem.

_____ Workplace dimensions. Displays should be located within the optimum viewing envelope and manual and foot controls should be within the optimum reach envelope. Seated versus standing operation should be considered with sufficient space allotted for the selected mode of operation.

_____ Location of displays. Visual displays should be located to promote the speed and accuracy of seeing. Considerations should be given to the positioning of two or more operators who can share a display so as to facilitate multiple seeing. The viewing requirements of a mobile observer should also be considered.

_____ Location of controls. Controls should be located to promote speed and accuracy of operation and adjustment. Consideration should be given to the location of controls shared by two or more operators to facilitate multiple use.

(Continued)

Figure 7-13, page 3 of 4.

Response
(S, U, NA)

DESIGN AREA 3 -- ENVIRONMENTAL CONTROLS

_____ Lighting quantity and quality. Lighting quantity and quality should be consistent with general standards. General illumination should be considered with respect to: size of detail to be discriminated, brightness contrast and ratio, time available for viewing, and glare effects.

_____ Acoustics and noise. Noise levels should be within recommended levels. The effects of noise of performance, and noise suppression devices, such as baffles and partitions, should be considered.

_____ Temperature, humidity, and air flow. Provision should be made to maintain temperature and humidity within recommended tolerances for the range of operating conditions as well as for adequate ventilation.

_____ Vibration and acceleration. Provision should be made to hold direction, frequency, and amplitude of mechanical vibration and forces of acceleration within recommended tolerances. The need for simulation of vibration and acceleration should be considered.

(Continued)

Figure 7-13, page 4 of 4.

Response
(S, U, NA)

DESIGN AREA 4 -- EQUIPMENT

_____ Displays and indicators. Symbolic and pictorial displays should promote interpretation of information and increase reading speed and accuracy. Displays should contain appropriate direction-of-movement relationships, be compatible with direction and amount of control movement, and be consistent with other displays. The following display elements should be considered: display scales, zone markings, labeling (position and color coding), and design of alphanumerics. Warning and caution devices for gaining operator attention and indicating the nature of malfunction should also be considered.

_____ Controls. Selection of controls should consider the characteristics of each control type and their suitability for given applications. The following elements of controls should be considered: size, placement, direction-of-movement relationships, compatibility with display movement, control coding requirements, and methods of preventing accidental operation.

design applications. That is, they are to be used by a human factors specialist as an aid in reviewing and evaluating design drawings and engineering development mock-ups. Their use in operational testing requires a very different orientation on the part of the evaluator, since he must now be concerned principally with evaluation of the device with respect to how it is used, rather than with basic human factors design principles only. These types of evaluation procedures, therefore, should only be used in ATD OT&E when other more sophisticated procedures are not possible to effect in the time available.

Trainee station concerns. Most ATDs incorporate a limited degree of training-support capabilities at the device trainee station. Such capabilities at the trainee station include Freeze, Store/Reset, Replay, and Auto-demo controls, and a few devices may contain additional such features. Controls for these functions are typically located together and in an inconspicuous place so as not to interfere with the total simulation (e.g., at the bottom of a radio stack). The limited nature of training-support equipment typically available at the trainee station usually results in relatively little time and effort being required for evaluation of that equipment.

Assessments of the trainee station with regard to these features should be oriented toward their instructional utility. Are those capabilities appropriate for the anticipated training use of the ATD? In the context of specific training tasks, are the training-support features easy to use? If not, why (time delays too long, complicated controls, etc.)? Are they adapted to self-instructional use of the ATD? As in the IOS evaluation, these considerations should be addressed during OT&E in terms of specific trainee tasks, because it is likely that whatever training activities are concurrently taking place would affect use (and evaluation) of training-support features. The basic elements of consideration would be similar to those identified earlier for IOS evaluation (see Tables 7-1 and 7-2).

The training scenario approach described for IOS evaluation may be modified for use in trainee station evaluation. The types of modification necessary have to do more with the content of approach rather than its basic structure; i.e., test objectives and task requirements must now be trainee oriented rather than I/O oriented. This approach would focus upon the characteristics of training-support equipment in the context of device self-instructional utilization; i.e., when device training control is allocated principally to the trainee station.

I/QOT&E vs. FOT&E implications. Evaluations conducted during a follow-on OT&E (FOT&E) can be more comprehensive than those employed in earlier QOT&E or IOT&E. The basic test objectives would remain largely the same with regard to the IOS, but during FOT&E they can be much more detailed and comprehensive in the types of information obtained. The basic evaluation approach also would be similar, but the I/O tasks of interest would be addressed in greater detail during the FOT&E given the more advantageous test environment. An advantage of the training scenario evaluation approach is that some training capabilities evaluations may occur concurrently with IOS evaluation with a resultant savings in testing time. Such a savings may be of special concern in IOT&E. For example, evaluation of the trainee station during "actual" training with the device may be accomplished. During an FOT&E, of course, this may be exactly what the OT&E test team is interested in evaluating.

REFERENCES

1. Caro, P. W., Pohlmann, L. D., & Isley, R. N. Development of simulator instructional feature design guides (Seville Tech. Rep. TR 79-12). Pensacola, FL: Seville Research Corporation, October 1979.

2. Charles, J. P., Willard, G., & Healy, G. Instructor pilot's role in simulation training (NAVTRAEQUIPCEN 75-C-0093-1). Orlando, FL: Naval Training Equipment Center, March 1976.

3. Charles, J. P. Instructor pilot's role in simulator training (Phase II) (NAVTRAEQUIPCEN 76-C-0034-1). Orlando, FL: Naval Training Equipment Center, August 1977.

4. Charles, J. P. Instructor pilot's role in simulator training (Phase III) (NAVTRAEQUIPCEN 76-C-0034-2). Orlando, FL: Naval Training Equipment Center, June 1978.

5. Hughes, R. G. Advanced training features: Bridging the gap between inflight and simulator-based models of flying training (AFHRL-TR-78-96). Brooks AFB, TX: Air Force Human Resources Laboratory, March 1979. (AD-A068 142).

6. Semple, C. A., Cotton, J. C., & Sullivan, D. J. Aircrew training device instructional support features (AFHRL-TR-80-58). Brooks AFB, TX: Air Force Human Resources Laboratory, 1980. (AD-A096 234)

7. Department of Defense. Human engineering design criteria for military systems, equipment, and facilities (MIL-STD-1472B). Washington, DC: Author, 31 December 1974.

Department of Defense. Human engineering design criteria for military systems, equipment, and facilities (MIL-STD-1472B, Change Notice 1). Washington, DC: Author, 10 May 1976.

8. Van Cott, H. P., & Kincade, R. G. (Eds.). Human engineering guide to equipment design (Rev. ed). Washington, DC: U.S. Government Printing Office, 1972.

9. AFR 80-14, Paragraph 17., 19 July 1976.

10. Tactical Air Command. Final report: Evaluation of the simulator for air-to-air combat (SAAC) FOT&E. Eglin AFB, FL: USAF Tactical Air Warfare Center, February 1979.

11. AFTEC. F5E IFS (PEACE WREN), IOT&E Test Plan, April 1978.

12. Smode, A. F., Gruber, A., & Ely, J. H. Human factors technology in the design of simulators for operator training (NAVTRADEVCEN 1103-1). Port Washington, NY: Naval Training Device Center, December 1963.

APPENDIX A: STATISTICAL PROCEDURES

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF TABLES

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

ACKNOWLEDGEMENTS

Permission to reproduce Tables A-4 and A-7 has been granted by the publishers of Annals of Mathematical Statistics.

Permission to reproduce Table A-5 has been granted by the publishers of the Bulletin of the Institute of Educational Resources at Indiana University.

Tables A-3 and A-8 are taken from Tables IV and VII of Fisher and Yates: Statistical Tables for Biological, Agricultural and Medical Research published by Longman Group Ltd. London, (previously published by Oliver and Boyd Ltd. Edinburgh) and by permission of the authors and publishers.

APPENDIX A

STATISTICAL PROCEDURES

INTRODUCTION

This appendix provides the test director with specific instructions on how to perform those descriptive, inferential, and correlational statistical procedures he is likely to need during ATD OT&E. It also provides him with guidance for assessing the reliability and validity of the measurement instruments used to collect the data of interest. As is shown in Table A-1, at least one descriptive, inferential, or correlational statistic is provided for each data type--nominal, ordinal, interval, and ratio. (See Chapter 5 for a discussion of these four data types.)

Even experts in experimental design and statistical analysis sometimes have difficulty in selecting the most appropriate statistical approaches for analyzing specific data sets. For this reason, unless the test director is reasonably familiar with statistical concepts and their underlying assumptions, and unless he feels confident in his ability to discern the implications for statistical analysis of the distinctions among nominal, ordinal, interval, and ratio data, he should seek assistance in deciding which of these analytic techniques are most appropriate for his situation.

Sections A, B, and C of this appendix address, in turn, descriptive, inferential, and correlational statistics. Each section not only defines each statistic, surfaces some of the advantages and/or restrictions to its use, but also provides detailed instructions for its calculation. Section D provides guidance on how to evaluate the reliability and validity of the measurement instruments used to collect questionnaire, rating scale, and performance data.

TABLE A-1. SPECIFIC DESCRIPTIVE, INFERENTIAL, AND CORRELATIONAL STATISTICAL PROCEDURES TO BE USED WITH NOMINAL, ORDINAL, INTERVAL, AND RATIO LEVEL DATA

	Descriptive	Inferential	Correlational
NOMINAL	Mode	Chi Square	Contingency Coefficient
ORDINAL	Median Range	MANN-WHITNEY U	Spearman Rank-Order Correlation Coefficient
INTERVAL/ RATIO	Mean Average Deviation Standard Deviation	ANOVA	Pearson Product-Moment Correlation

A. DESCRIPTIVE STATISTICS

INTRODUCTION

The descriptive function of statistics is to summarize or describe sets of data in clear and convenient ways. For example, if 20 evaluators assessed the capability of an ATD to train a landing task, one would not want to interpret all 20 scores individually. Instead, one would want a method for summarizing those 20 scores so that the information they convey could be communicated easily.

There are two general types of measures that can be used to describe sets of scores: measures of central tendency and measures of dispersion or variability.

Measures of Central Tendency

A measure of central tendency is a single score or number that describes the general "location" of a set of scores. One example would include the average age of Air Force navigators. There are three commonly used measures of central tendency:

The mode. The mode is the score that occurs most often. A set of scores can have more than one mode if two or more scores occur equally often.

The median. The median is the middle score in an ordered series. It represents the mid-point of a set of scores; half of the scores are above the median, and the remaining half are below.

The mean. The mean (i.e., arithmetic mean) is the common "average" of a set of scores. It is the most frequently used measure of central tendency.

The following example shows how the mode, median, and mean are calculated for ratings of eight IPs (Raters A-H) concerning the ability of an ATD to train a landing task (in this example, the ratings could have ranged from 1 to 7).

RATER:	A	B	C	D	E	F	G	H
SCORE:	7	7	6	5	4	4	4	3

The mode of these scores is 4 because it is the score that occurs most often; it is the most frequently observed score.

The median, or midpoint, is 4.5 because, as is shown below, it is exactly in the middle of the distribution of scores.

7
7
6
5
4
4
4
3

The mean, or average, of these scores is 5. This is calculated by dividing the sum of the scores by the number of scores.

$$\frac{7 + 7 + 6 + 5 + 4 + 4 + 4 + 3}{8} = \frac{40}{8} = 5$$

Choice of central tendency measures. The choice of central tendency measures depends upon two factors: the level of measurement represented by the data; and the nature of the distribution of scores. Only the mode can be used to summarize nominal level data, whereas both the mode and median can be used to summarize ordinal level data. All three measures of central tendency can be used with either interval or ratio level data.

Although the mean is the usual measure of choice, it can be unduly influenced by extreme scores. For example, consider the following two sets of scores:

A: 7, 7, 6, 5, 4, 4, 4, 3 Sum of A = 40

B: 47, 7, 6, 5, 4, 4, 4, 3 Sum of B = 80

The mean of the A scores is 5 and the mean of the B scores is 10, even though A and B scores are identical with one exception. The mean of distribution B gives a misleading picture of the distribution of scores. On the other hand, the median for both the A and B sets of scores is 4.5. An advantage of the median is that it is not unduly influenced by extreme scores. The median, therefore, should be used as the measure of central tendency when the highest score(s) are either much further from the mean than the lowest scores(s) (as in the distribution of B scores discussed above), or when the highest score(s) are much closer to the mean than the lowest score(s). In other cases, the mean should be used because it is typically the most stable measure of central tendency and because it is usable as a datum in further statistical analyses as discussed below.

Measures of Variability

The minimum variability among individual scores which can possibly occur is zero. Zero variability occurs only when all the scores in the distribution are the same. There is, however, no limit on how large the variability among individual subject scores can be. This will depend on the actual value of each of the individual scores in a particular distribution. Finally, there is no such thing as negative variability. The variability between individual scores can never be negative, only zero or positive. To illustrate, consider the six distributions involving a small number of scores shown below.

Distribution 1:	7.0	7.0	7.0	7.0	7.0	7.0	7.0	
Distribution 2:	7.0	7.1	7.1	7.1	7.2	7.0		
Distribution 3:	40.0	40.1	40.1	40.1	40.2	40.0		
Distribution 4:	7.0	7.0	6.0	5.0	4.0	4.0	4.0	3.0
Distribution 5:	10.0	10.0	9.0	7.0	5.0	4.0	3.0	2.0
	0.0	0.0						
Distribution 6:	97.8	88.5	83.4	76.2	69.9	67.3	58.4	44.7

Distribution 1 is an example of a distribution of individual scores which has zero variability. This is because all the scores have the same value. Distribution 2 shows a small amount of variability because there are only small differences among individual scores. The variability in Distribution 3 is exactly the same amount as that in Distribution 2, even though the actual values (40 vs. 7) of Distribution 3 scores are much larger than those of Distribution 2. This occurs because the variability of a distribution of scores is not dependent on the numerical values of the scores in the distribution, but instead on the differences among individual scores. For this reason, each of the remaining three distributions of scores (4, 5, and 6) contains more variability than the one that precedes it.

Choice of variability measures. The measure of variability appropriate for a particular set of scores will depend on the level of measurement those scores represent. There is no precise or standard way of describing the variability of either nominal or ordinal level data except in terms of the proportion of cases which fall in the modal category, or in terms of the range of scores in the latter. The range is determined by simply subtracting the largest score from the smallest score.

In the case of interval or ratio level data, two measures of variability are commonly employed: average deviation and standard deviation.

Average Deviation

The average deviation (AD) of a distribution indicates how much distance there is, on the average, between the individual scores in the distribution and the mean of that distribution.

The formula for the average deviation is as follows:

$$AD = \frac{\sum | X_i - M |}{N}$$

where X_i is the individual score, M is the mean, N is the number of scores and $\sum$ indicates the sum of all absolute $| X_i - M |$ values. The AD is calculated as follows:

Step 1. Calculate the mean score.

Step 2. Subtract the mean from each individual score to obtain the absolute differences.

(Note: To obtain the absolute value of all the differences, change any of the differences between the mean and individual scores that are negative to positive values.)

Step 3. Add together all of the absolute differences between the mean and individual scores.

Step 4. Divide the sum by the number of scores.

The following distribution of scores and the calculation of its average deviation are provided to illustrate this procedure.

X_i (Score)	$X_i - M$ (Score - Mean)	$\| X_i - M \|$ (Absolute Value)
7	7 - 5 = 2	2
7	7 - 5 = 2	2
6	6 - 5 = 1	1
5	5 - 5 = 0	0
4	4 - 5 = -1	1
4	4 - 5 = -1	1
4	4 - 5 = -1	1
3	3 - 5 = -2	2
Mean = 5		Sum = 10

$$AD = \frac{10}{8} = 1.25$$

Standard Deviation

The most frequently employed descriptive measure of distribution score variability is the standard deviation (SD). The standard deviation, like the average deviation, provides information about the extent to which individual scores vary around the mean. However, the standard deviation contains more specific information. Approximately 68% of all the scores are within +1 and -1 standard deviations of the mean; approximately 95% of all the scores are with +2 and -2 standard deviations of the mean, and approximately 99.7% of all scores are within +3 and -3 standard deviations of the mean.

The equation for the standard deviation is:

$$SD^1 = \sqrt{\frac{\sum(X_i - M)^2}{N}}$$

where X_i is the individual score, M is the mean, N is the number of scores, "$\sum$" is the sum (of all $(X_i - M)^2$ values), and "$\sqrt{\ }$" is the square root. The SD is calculated as follows:

Step 1. Calculate the mean score.

Step 2. Subtract the mean from each individual score.

Step 3. Square each difference.

Step 4. Sum all of the squared differences.

Step 5. Divide the sum by the number of scores.

Step 6. Take the square root of the average of the summed differences.

Applying this procedure to the following distribution of scores, calculation of the standard deviation is as follows:

---

[1]There are two symbols for the standard deviation: SD and σ. Either one is appropriate. The square of the standard deviation (σ^2) is referred to as the "variance." The variance will be used in the portion of the next section of this chapter that deals with the "analysis of variance" techniques.

X_i (Score)	$X_i - M$ (Score - Mean)	(Squared difference)
7	7 - 5 = 2	2 x 2 = 4
7	7 - 5 = 2	2 x 2 = 4
6	6 - 5 = 1	1 x 1 = 1
5	5 - 5 = 0	0 x 0 = 0
4	4 - 5 = -1	-1 x 1 = 1
4	4 - 5 = -1	-1 x 1 = 1
4	4 - 5 = -1	-1 x 1 = 1
3	3 - 5 = -2	-2 x 2 = 4
Mean = 5		Sum = 16

$$SD = \sqrt{\frac{16}{8}} = \sqrt{2} = 1.4$$

B. INFERENTIAL STATISTICS

INTRODUCTION

Inferential statistics allow inferences to be drawn about an entire population based on data obtained from a small group of individuals (sample) from that population. The inferences to be made by a test director usually involve decisions concerning whether one set of data is the same as or different from some other set of data; or whether the obtained data are the same as or different from expected values. For example, a test director may be faced with deciding whether the performance scores obtained from a group of trainees who received ATD training are meaningfully different from the performance scores obtained from a group of trainees who did not receive ATD training.

There is always a risk that the findings may be influenced by factors that are unrelated to the variables of interest for a given situation. These factors are considered to be non-systematic in their function (i.e., chance). Inferential statistics reflect that chance risk in terms of a probability statement, or p level. Each inferential statistic that is derived is "tested" to determine the likelihood that the obtained results are real or reliable--i.e., would occur again if the evaluation procedures were repeated on another sample from the population of interest.

THE P LEVEL: STATISTICAL AND PRACTICAL SIGNIFICANCE

The p level refers to the probability that obtained differences in performance data between evaluation groups are due to differences in ATD training rather than to chance variation. For example, a p level of .05 means that there are only 5 chances in 100 that the obtained performance difference between the evaluation groups are due to factors other than ATD training. This, in turn, means that there are 95 chances in 100 that the obtained performance difference is indeed due to the differences in training for the two groups, i.e., ATD training.

If a difference in performance is found to exist between the evaluation groups, then it can be inferred (at the specified level of probability) that a similar difference will be obtained for any similar trainee groups which undergo the same training. The confidence that can be placed in such an inference depends on the numerical value of the inferential statistic obtained and its associated p level. If the p level is low (e.g., a p level of .01), and therefore the probability of an obtained performance difference being due to chance alone is quite small, it can be predicted with confidence that a similar difference would be obtained again under comparable circumstances. On

the other hand, if a high p level is obtained (e.g., a p level of .20), the chances of a performance score difference being due to chance alone is also relatively high, and it cannot be predicted confidently that a similar difference will be obtained again.

A level of .05 is usually regarded as the level below which differences are deemed to be statistically significant and above which they are deemed to be statistically nonsignificant. Although this "convention" is appropriate for most research purposes, such a rigid cutoff point may not be useful for describing data from, or making decisions relevant to, operational settings. Therefore, an obtained performance difference with a p level which is higher than .05 may sometimes be adequate for making decisions on the future role of a specific ATD in a particular training program. The practical significance or nonsignificance of an obtained performance difference is another matter. It is a function of the size of a difference and its importance or practical meaning in the context in which it is obtained and used. For example, a reduction of 1.5 knots in airspeed error might be statistically significant (i.e., reliable) in a given situation, but it might be judged to be of no practical significance or consequence in that same situation.

Possible Errors When Using Inferential Statistics

Inferential statistics always involve drawing inferences about an entire population of individuals based on data obtained from a small group (sample) drawn from that population. The inference made by the test director usually involves the statement that one set of data is the same as or different from some other set of data, or that the obtained data are the same as or different from expected values of the obtained data. In reality, the data sets for the populations are either the same or they are different. Since the inference the test director makes always has a chance of being wrong (defined by the p level determined by the inferential test), four possible outcomes of a test director's decision or inference are possible. These are illustrated in Table A-2.

TABLE A-2. POSSIBLE ERRORS WHEN USING INFERENTIAL STATISTICS

		IN REALITY: No Difference Exists	IN REALITY: A Real Difference Exists
Test Director's Inference About Two Sets of Data	No Difference Exists	CORRECT INFERENCE	TYPE II ERROR Test director concludes that no difference exists when, in reality, it does
	A Difference Exists	TYPE I ERROR Test director concludes that a difference exists when, in reality, it does not	CORRECT INFERENCE

From this table, it can be seen that the test director can make two kinds of errors when drawing inferences about an entire population based on data from one or more small samples drawn from that population. To illustrate the first type of error (Type I error), assume the test director concludes from his sample data that crewmen who are trained with a particular ATD reach criterion performance in the aircraft more quickly than do crewmen who receive all of their training in the aircraft itself, when, in reality, the difference obtained between the trainee samples is not representative of that between the two trainee populations as a whole. A Type I error was made by the test director, because he concluded that the ATD would facilitate the training process for all crewmen from that population, when, in fact, it would not. A difference was obtained for the trainees used in the evaluation, but due to sampling error or some other factor, the same difference would not be found if all the potentially available trainees (the population) were tested under the same circumstances.

A test director commits a Type II error when he concludes that no difference exists between two sets of sample data, based on the outcome of a statistical test, when, in reality, a difference does exist, and that such a difference would be found if the entire population of trainees were tested. For example, a Type II error would be committed by a test director if he concluded that an ATD is not an effective training device, based on a TOT evaluation involving samples, when, in reality, the ATD is effective and would be shown to be so if the entire population were used instead.

The key to understanding Type I errors is the p level obtained during data analysis. Recall that an obtained difference with a p level of .05 (regardless of the specific analysis procedure employed) indicates that there are 95 chances in 100 that a difference would be obtained if the entire population of trainees were used. Conversely, there are 5 chances in 100 that the difference obtained for the sample is due to sampling bias error or some other random error. Thus, the probability of making a Type I error in this case is equal to the p level, or .05. If a test director obtains a p level of .05, he should infer from his sample that a real difference exists. However, the risk involved in doing so is 5 chances in 100 that his inference about the entire population is wrong and therefore he does have a 5% chance of making a Type I error.

The probability of making a Type II error is harder to determine because it does not directly correspond to the p level or another numerical index associated with an inferential test calculation. However, it can be said that as the probability of making a Type I error increases, the probability of making a Type II error decreases, and vice versa.

The specific procedures for conducting inferential statistical tests that should be used when analyzing different levels of data (i.e., nominal, ordinal, and interval/ratio) are provided in the following discussion. Should the test director have any difficulty with these materials, he should not hesitate to seek additional guidance.

CHI SQUARE TEST AND NOMINAL DATA

The Chi Square Test is appropriate for use with nominal level data found in two or more independent samples. It is used when the data are expressed in terms of the number (frequency) of scores, people, or things in each of several categories. The Chi Square Test is used to determine whether the frequencies of items in each category are distributed as expected, or whether they are distributed in a manner not expected. For a more detailed discussion of the Chi Square Test and related procedures, see Siegel [1].

As an example, assume that a new ATD has been developed for use in a training program which formerly employed only aircraft experience. An examination of training program records reveals that, in the past, 30% of the trainees entering the program failed to perform satisfactorily and attrited during training. This means that normally 7 out of 10, or 70%, of the trainees entering the program could be expected to graduate successfully from it. The OT&E task is to determine whether or not the addition of the new ATD to the training

program has significantly changed (hopefully improved) the number of trainees successfully completing the course.

During the OT&E, a group of 30 trainees has been trained using the new ATD in addition to aircraft experience. Based on the above cited past experience, 30%, or 9, of these trainees would be expected to attrite during the training program, leaving 70%, or 21 trainees, to graduate if the new ATD has not had an impact on the training program.

Upon completion of the new training program, however, only three trainees (10%) had been eliminated. The question confronting the test director is this: Is the number of trainees attriting from the program, in reality, different from the number being eliminated before the new program was instituted, or is the decrease just a chance and nonsignificant deviation from the previously experienced values? More specifically, did the ATD training reduce attrition?

Performing Chi Square Tests on these data will aid the test director in making this judgment.

Chi Square Test Procedure

The formula for computing a chi square (χ^2) is as follows:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where:

O = the observed number of cases in each category;

E = the expected number of cases in each category; and

$\sum$ = directs one to sum over all categories.

The procedure for conducting a Chi Square Test (χ^2) is outlined step by step below. The hypothetical data from the above example are used to illustrate this procedure.

Step 1. Create a table like the one below which contains the expected and obtained numbers of trainees in each category. (The values are taken from the above OT&E scenario.)

CHI SQUARE (χ^2) DATA TABLE

	Trainees successfully completing program	Trainees attriting during program
Expected	21	9
Obtained	27	3

Step 2. For each trainee category, subtract the expected value from the obtained value, square this difference, and then divide this number by the expected value. Doing this for the hypothetical data, we get:

Category 1: $\frac{(27-21)^2}{21} = \frac{(6)^2}{21} = \frac{36}{21} = 1.71$

Category 2: $\frac{(3-9)^2}{9} = \frac{(-6)^2}{9} = \frac{36}{9} = 4.00$

Step 3. Calculate χ^2 by summing the numerical results of the above calculations for each category. Doing this for these hypothetical data we get.

$$\chi^2 = 1.71 + 4.00 = 5.71$$

Step 4. Calculate a value referred to as "degrees of freedom (df)" by subtracting 1 from the number of rows and 1 from the number of columns in the table, and then multiplying these two numbers together.

df = (number of rows -1) x (number of columns -1)

Doing this for these hypothetical data, we get:

Number of rows = 2; 2-1 = 1

Number of columns = 2; 2-1 = 1

df = 1 x 1 = 1

Step 5. Refer to Table A-3. In the case of these hypothetical data, df = 1 and χ^2 = 5.71. Look up the df value of 1 in the left-hand column. Move across this row towards the right until a χ^2 value is encountered that is larger (i.e., 6.63) than the obtained χ^2 value, of 5.71. Move back (left) one value and refer to the number at the top of this column (i.e., .05). This is the p level for your data. This means that there are 95 chances out of 100 that the obtained numbers are really different from the expected values and that there are only 5 chances in 100 that the obtained values are not different.

Such a significance level would usually be considered adequate reason for the test director to infer that the new ATD does indeed decrease the attrition in this particular training program and that it should continue to be used. There is adequate reason for doing so, because there is only a 5% chance that the test director's inference would be wrong given the Chi Square value that was obtained.

THE MANN-WHITNEY U TEST AND ORDINAL DATA

When at least ordinal level data have been collected, the Mann-Whitney U test may be used to determine whether or not differences exist between measures obtained from two groups of subjects. In applying the Mann-Whitney U procedure, measures obtained from two different groups of trainees are combined and rank ordered in a list from lowest to highest. If the two sets of scores are equal, there should be an equal number of scores from both groups in the low portion of the list, as well as an equal number of scores from both groups in the middle and high portions of the list. If, however, the two sets of scores are not equal, then it would be expected that scores from each group would not be equally represented in the low, middle, and high portions of the rank ordered list. To illustrate, assume that the following sets of performance scores have been collected:

Group 1: 3, 4, 5

Group 2: 8, 9, 10

If these scores are combined and rank ordered, the following list is obtained:

Group 1	Group 2
3, 4, 5	8, 9, 10

As can be seen, the scores from Group 1 are all clustered at the low end of the combined list, while the scores from Group 2 are all

TABLE A-3. CRITICAL VALUES OF CHI SQUARE*

	.20	.10	.05	.01
1	1.64	2.71	3.84	6.63
2	3.22	4.61	5.99	9.21
3	4.64	6.25	7.82	11.34
4	5.99	7.78	9.49	13.28
5	7.29	9.24	11.07	15.09
6	8.56	10.64	12.59	16.81
7	9.80	12.02	14.07	18.48
8	11.03	13.36	15.51	20.09
9	12.24	14.68	16.92	21.67
10	13.44	15.99	18.31	23.21
11	14.63	17.28	19.68	24.72
12	15.81	18.55	21.03	26.22
13	16.98	19.81	22.36	27.69
14	18.15	21.06	23.68	29.14
15	19.31	22.31	25.00	30.58
16	20.46	23.54	26.30	32.00
17	21.62	24.77	27.59	33.41
18	22.76	25.99	28.87	34.81
19	23.90	27.20	30.14	36.19
20	25.04	28.41	31.41	37.57
21	26.17	29.62	32.67	38.93
22	27.30	30.81	33.92	40.29
23	28.43	32.01	35.17	41.64
24	29.55	33.20	36.42	42.98
25	30.68	34.38	37.65	44.31
26	31.80	35.56	38.89	45.64
27	32.91	36.74	40.11	46.96
28	34.03	37.92	41.34	48.28
29	35.14	39.09	42.56	49.59
30	36.25	40.26	43.77	50.89

For df values larger than 30, refer to an outside statistical reference.

*Table is abridged from Table IV of: Fisher, R. A., & Yates, F. Statistical tables for biological, agricultural, and medical research (6th edition). Edinburgh: Oliver and Boyd Ltd., 1963.

clustered at the high end of list. In this example it is easy to see that the scores contained in the two group lists are different. However, differences in real data are not always as easy to identify, so the Mann-Whitney U test is used to help make the proper inferential judgment. The Mann-Whitney U test can be employed whether the numbers of scores in each group are equal or unequal.

Mann-Whitney U Test Methods

The Mann-Whitney U test is calculated in one of two ways, depending on how many scores are in each sample. If each group contains eight scores or less, Method 1 is used. If one or both groups has 9-20 scores in it, Method 2 can be used. When one group has more than 20 scores in it, a special version of the Mann-Whitney U statistic is required (Siegel, [1]).

METHOD 1.

Consider the following two groups of hypothetical performance scores. Notice that neither group has more than eight scores in it.

ATD Group (ATD): 9, 11, 15

Control Group (CON): 6, 8, 10, 13

Step 1. Rank the scores in order from lowest to highest, being careful to record which score came from which group. Doing this for the hypothetical data:

Score:	6	8	9	10	11	13	15
Group:	CON	CON	ATD	CON	ATD	CON	ATD

Step 2. Next, count the number of ATD Group scores that precede each score in the Control Group. For the Control Group score of 6, no ATD score precedes it. This is also true for the Control Group score of 8. However, for the Control Group score of 10, one ATD Group score precedes it, that of 9, while for the final Control Group score, two ATD Group scores precede it, that of 9 and 11. U is equal to the sum of ATD Group scores that precede each Control Group score. Thus:

$$U = 0 + 0 + 1 + 2 = 3$$

Step 3. Refer to Table A-4 to determine the p level of U = 3 in this example. As can be seen, Table A-4 is actually six separate subtables. The appropriate subtable to use depends on

TABLE A-4. MANN-WHITNEY U TEST--METHOD 1

Table of Probabilities Associated with Values as Small as Observed Values of U in the Mann-Whitney Test*

$n_2 = 3$

U \ n_1	1	2	3
0	.250	.100	.050
1	.500	.200	.100
2	.750	.400	.200
3		.600	.350
4			.500
5			.650

$n_2 = 4$

U \ n_1	1	2	3	4
0	.200	.067	.028	.014
1	.400	.133	.057	.029
2	.600	.267	.114	.057
3		.400	.200	.100
4		.600	.314	.171
5			.429	.243
6			.571	.343
7				.443
8				.557

$n_2 = 5$

U \ n_1	1	2	3	4	5
0	.167	.047	.018	.008	.004
1	.333	.095	.036	.016	.008
2	.500	.190	.071	.032	.016
3	.667	.286	.125	.056	.028
4		.429	.196	.095	.048
5		.571	.286	.143	.075
6			.393	.206	.111
7			.500	.278	.155
8			.607	.365	.210
9				.452	.274
10				.548	.345
11					.421
12					.500
13					.579

$n_2 = 6$

U \ n_1	1	2	3	4	5	6
0	.143	.036	.012	.005	.002	.001
1	.286	.071	.024	.010	.004	.002
2	.428	.143	.048	.019	.009	.004
3	.571	.214	.083	.033	.015	.008
4		.321	.131	.057	.026	.013
5		.429	.190	.086	.041	.021
6		.571	.274	.129	.063	.032
7			.357	.176	.089	.047
8			.452	.238	.123	.066
9			.548	.305	.165	.090
10				.381	.214	.120
11				.457	.268	.155
12				.545	.331	.197
13					.396	.242
14					.465	.294
15					.535	.350
16						.409
17						.469
18						.531

*Reproduced from: Mann, H. B., & Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics, 1947, 18, 52-54.

TABLE A-4. (Continued)

Table of Probabilities Associated with Values as Small as Observed Values of U in the Mann-Whitney Test*

$n_2 = 7$

U \ n_1	1	2	3	4	5	6	7
0	.125	.028	.008	.003	.001	.001	.000
1	.250	.056	.017	.006	.003	.001	.001
2	.375	.111	.033	.012	.005	.002	.001
3	.500	.167	.058	.021	.009	.004	.002
4	.625	.250	.092	.036	.015	.007	.003
5		.333	.133	.055	.024	.011	.006
6		.444	.192	.082	.037	.017	.009
7		.556	.258	.115	.053	.026	.013
8			.333	.158	.074	.037	.019
9			.417	.206	.101	.051	.027
10			.500	.264	.134	.069	.036
11			.583	.324	.172	.090	.049
12				.394	.216	.117	.064
13				.464	.265	.147	.082
14				.538	.319	.183	.104
15					.378	.223	.130
16					.438	.267	.159
17					.500	.314	.191
18					.562	.365	.228
19						.418	.267
20						.473	.310
21						.527	.355
22							.402
23							.451
24							.500
25							.549

*Reproduced from: Mann, H. B., & Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics, 1947, 18, 52-54.

TABLE A-4. (Continued)

Table of Probabilities Associated with Values as Small as Observed Values of U in the Mann-Whitney Test*

$n_2 = 8$

U \ n_1	1	2	3	4	5	6	7	8
0	.111	.022	.006	.002	.001	.000	.000	.000
1	.222	.044	.012	.004	.002	.001	.000	.000
2	.333	.089	.024	.008	.003	.001	.001	.000
3	.444	.133	.042	.014	.005	.002	.001	.001
4	.556	.200	.067	.024	.009	.004	.002	.001
5		.267	.097	.036	.015	.006	.003	.001
6		.356	.139	.055	.023	.010	.005	.002
7		.444	.188	.077	.033	.015	.007	.003
8		.556	.248	.107	.047	.021	.010	.005
9			.315	.141	.064	.030	.014	.007
10			.387	.184	.085	.041	.020	.010
11			.461	.230	.111	.054	.027	.014
12			.539	.285	.142	.071	.036	.019
13				.341	.177	.091	.047	.025
14				.404	.217	.114	.060	.032
15				.467	.262	.141	.076	.041
16				.533	.311	.172	.095	.052
17					.362	.207	.116	.065
18					.416	.245	.140	.080
19					.472	.286	.168	.097
20					.528	.331	.198	.117
21						.377	.232	.139
22						.426	.268	.164
23						.475	.306	.191
24						.525	.347	.221
25							.389	.253
26							.433	.287
27							.478	.323
28							.522	.360
29								.399
30								.439
31								.480
32								.520

*Reproduced from: Mann, H. B., & Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics, 1947, 18, 52-54.

how many scores (up to 8) are in the larger sample. In our example, the number of scores in the smaller group (n_1) is equal to 3 and the number of scores in the larger group (n_2) is equal to 4. Remember that U = 3. Referring to Table A-4, locate the subtable for n_2 = 4. Next locate n_1 along the top of the table. Finally, follow down the column until the value across from U = 3 is reached. For this example, this value is .200, i.e., p = .20. In other words, there are 20 chances out of 100 that the scores contained in the ATD and Control Groups are the same, and 80 out of 100 chances that the scores in the two groups are different.

METHOD 2.

If one or both of the group sizes is larger than 8, but less than 20, Method 2 should be used. To illustrate the computation of U using Method 2, consider the following hypothetical performance scores:

ATD Group: 9, 10, 11, 14, 15, 16, 20

Control Group: 4, 5, 5, 6, 7, 8, 12, 13, 17

Note that the Control Group has 9 scores in it, which means that Method 2 must be used.

Step 1. Rank order and list the scores from both groups in the same way that was done for Method 1:

Score:	4,	5,	5,	6,	7,	8,	9,	10,	11,	12,
Group:	CON	CON	CON	CON	CON	CON	ATD	ATD	ATD	CON

Score:	13,	14,	15,	16,	17,	20
Group:	CON	ATD	ATD	ATD	CON	ATD

Step 2. Assign a rank of 1 to the lowest score, a rank of 2 to the next lowest score, etc.:

Rank:	1	2	3	4	5	6	7	8	9	10	11
Score:	4,	5,	5,	6,	7,	8,	9,	10,	11,	12,	13,
Group:	CON	CON	CON	CON	CON	CON	ATD	ATD	ATD	CON	CON

Rank:	12	13	14	15	16
Score:	14,	15,	16,	17,	20
Group:	ATD	ATD	ATD	CON	ATD

Step 3. Sum the assigned rankings for the ATD Group scores:

Sum of Rankings = 7 + 8 + 9 + 12 + 13 + 14 + 16 = 79

Step 4. Compute U, using the following formula:

$$U = n_1 \cdot n_2 + \frac{n_1 (n_1 - 1)}{2} - R$$

where:

- R = sum of ATD rankings
- n_1 = number of scores in ATD Group
- n_2 = number of scores in Control Group
- $n_1 n_2$ = the number of scores in the ATD Group multiplied by the number of scores in the Control Group.

Putting the appropriate values into the formula results in:

$$U = (7)(9) + \frac{7(7+1)}{2} - 79$$

$$U = 63 + \frac{7(8)}{2} - 79$$

$$U = 63 + \frac{56}{2} - 79$$

$$U = 63 + 28 - 79$$

$$U = 91 - 79$$

$$U = 12$$

Step 5. Next, the obtained U value must be subtracted from n_1n_2. This new value is called U':

$$U' = n_1n_2 - U$$

$$U' = 63 - 12$$

$$U' = 51$$

Now values of U and U' have been calculated. The smaller of the two should be used to determine the p levels. Because U is the smaller of the two (U = 12 and U' = 51), U will be used to calculate the p level.

Step 6. To determine the p level, refer to Table A-5. Again, Table A-5 is a group of subtables. This time, however, the subtables are separated out by p level. Two p levels are represented; .05 and .10. (At this point, the test director should decide which p level he wishes to use.) Finally, the test director should locate the appropriate value of n_1 (number of scores in the smaller group) along the lefthand column of the subtable, and n_2 (number of scores in the larger group) along the top row of the subtable. If U is equal to or less than the number listed at the intersection of n_1 and n_2, then the test director may infer that there is a difference in performance scores between the ATD and Control Groups with only 5 or 10 chances out of 100 of being wrong and 95 or 90 chances out of 100 of being right, depending on which table is chosen. If the U value is greater than the value listed at the intersection of n_1 and n_2 in either table, the test director should seek more

TABLE A-5. MANN-WHITNEY U TEST--METHOD 2

Table of Critical Values of U in the Mann-Whitney Test*

Test at p = .05

n_1 \ n_2	9	10	11	12	13	14	15	16	17	18	19	20
1												
2	0	0	0	1	1	1	1	1	2	2	2	2
3	2	3	3	4	4	5	5	6	6	7	7	8
4	4	5	6	7	8	9	10	11	11	12	13	13
5	7	8	9	11	12	13	14	15	17	18	19	20
6	10	11	13	14	16	17	19	21	22	24	25	27
7	12	14	16	18	20	22	24	26	28	30	32	34
8	15	17	19	22	24	26	29	31	34	36	38	41
9	17	20	23	26	28	31	34	37	39	42	45	48
10	20	23	26	29	33	36	39	42	45	48	52	55
11	23	26	30	33	37	40	44	47	51	55	58	62
12	26	29	33	37	41	45	49	53	57	61	65	69
13	28	33	37	41	45	50	54	59	63	67	72	76
14	31	36	40	45	50	55	59	64	67	74	78	83
15	34	39	44	49	54	59	64	70	75	80	85	90
16	37	42	47	53	59	64	70	75	81	86	92	98
17	39	45	51	57	63	67	75	81	87	93	99	105
18	42	48	55	61	67	74	80	86	93	99	106	112
19	45	52	58	65	72	78	85	92	99	106	113	119
20	48	55	62	69	76	83	90	98	105	112	119	127

*Adapted and abridged from Tables 1, 3, 5, and 7 of: Auble, D. Extended tables for the Mann-Whitney statistic. Bulletin of the Institute of Educational Resources at Indiana University, 1953, I(2).

TABLE A-5. (Continued)

Table of Critical Values of U in the Mann-Whitney Test*

Test at p = .10

n_1 \ n_2	9	10	11	12	13	14	15	16	17	18	19	20
1											0	0
2	1	1	1	2	2	2	3	3	3	4	4	4
3	3	4	5	5	6	7	7	8	9	9	10	11
4	6	7	8	9	10	11	12	14	15	16	17	18
5	9	11	12	13	15	16	18	19	20	22	23	25
6	12	14	16	17	19	21	23	25	26	28	30	32
7	15	17	19	21	24	26	28	30	33	35	37	39
8	18	20	23	26	28	31	33	36	39	41	44	47
9	21	24	27	30	33	36	39	42	45	48	51	54
10	24	27	31	34	37	41	44	48	51	55	58	62
11	27	31	34	38	42	46	50	54	57	61	65	69
12	30	34	38	42	47	51	55	60	64	68	72	77
13	33	37	42	47	51	56	61	65	70	75	80	84
14	36	41	46	51	56	61	66	71	77	82	87	92
15	39	44	50	55	61	66	72	77	83	88	94	100
16	42	48	54	60	65	71	77	83	89	95	101	107
17	45	51	57	64	70	77	83	89	96	102	109	115
18	48	55	61	68	75	82	88	95	102	109	116	123
19	51	58	65	72	80	87	94	101	109	116	123	130
20	54	62	69	77	84	92	100	107	115	123	130.	138

*Adapted and abridged from Tables 1, 3, 5, and 7 of: Auble, D. Extended tables for the Mann-Whitney statistic. Bulletin of the Institute of Educational Resources at Indiana University, 1953, 1(2).

complete U Tables from an appropriate statistical reference such as Siegel [1] to identify the p level he has actually obtained. With this additional information, he can then be more certain of the confidence he should have in the inference he eventually makes. Since the tabled value of U for p = .05, $n_1 = 7$, $n_2 = 9$ is equal to the value we computed in our example (U = 12), the test director could conclude (at the .05 level) that the ATD Group did perform better than the Control Group.

ANALYSIS OF VARIANCE AND INTERVAL OR RATIO DATA

Analysis of Variance, or simply, ANOVA, is a procedure for determining the likelihood that differences between two (or more) groups of interval or ratio level scores are statistically different.

The idea behind the ANOVA procedure is as follows. When a sample is randomly drawn from a population, the variance (see Page A-7) of the sample will approximate, or be an estimate of, the variance of the overall population from which it was drawn. By the same token, if two samples are randomly drawn from the same population, the variances of the two samples will not only approximate the variance of the population, but they will also approximate each other. In other words, the variance of samples randomly chosen from the same population should be approximately equal to each other, because each sample variance is an estimate of the same number--the population variance.

Thus, if a ratio is formed from these two estimates of population variance, the resulting quotient should be fairly close to 1.00 if the two estimates are of the same population variance. If the quotient is not close to 1.00, it is likely that the two sample variances have come from different populations.

The ANOVA procedure is a method for obtaining two mathematically independent estimates of population variance from the sample data. The ratio of the two variance estimates will be close to 1.00 if ATD training has had no effect on aircrew training, because there is no real difference between performance scores for the ATD and Control Groups. If, however, the two estimates of population variance calculated from the sample data differ, dividing the larger one by the smaller one will yield a ratio larger than 1.00, and this would be an indication (at some level of probability) that ATD training does influence the performance scores of the ATD Group. This quotient or ratio of variances is referred to as the F ratio.

Analysis of variance procedures. The following step-by-step procedure may be followed to analyze data from a Two-Group TOT evaluation. This procedure may be employed whether the numbers of trainees in the evaluation groups are equal or unequal. It should also be

noted that the ANOVA procedure outlined below is just one of several. ANOVA is actually a flexible, powerful family of statistical procedures. The ANOVA procedure outlined below has been selected because of its general applicability to IOT methods, but the test director may wish to familiarize himself with the other procedures as well. For detailed guidance on how to apply the entire range of ANOVA procedures, see Keppel [2].

It should be noted that the ANOVA procedure is designed to evaluate the performance of two or more groups of subjects on a single measure only. Note that performance measures are often composites made up of two or more measures and that simultaneous analysis of multiple measures will require more complex analysis procedures which are beyond the scope of this appendix. Such procedures are referred to as multivariate analysis of variance (MANOVA). The test director should seek outside statistical guidance and computer support if his data and/or test objectives require this type of analysis.

Procedure for computing ANOVA. Hypothetical trials-to-criterion scores for Task Z in the aircraft for an ATD Group and a Control Group are shown below. In the following discussion, this set of hypothetical data will be analyzed to illustrate the computational procedure.

ANOVA DATA TABLE

Individual Trials to Criterion Scores on Task Z for Trainees in an ATD Group and a Control Group

Control Group		ATD Group	
Trainee	Score	Trainee	Score
1	11	1	10
2	13	2	9
3	10	3	7
4	9	4	6
5	10	5	11
6	15	6	5
7	16		
8	19		
9	11		
10	13		

Step 1. Arrange individual scores by group in the manner shown in the table above. Note that there are different numbers of trainees in each group.

Step 2. Square each score from both groups. Add all the squared scores together. Designate this sum [A].

$$[A] = (11)^2 + (13)^2 + (10)^2 + (9)^2 + (10)^2 + (15)^2 + (16)^2 + (19)^2 + (11)^2 + (13)^2 + (10)^2 + (9)^2 + (7)^2 + (6)^2 + (11)^2 + (5)^2$$

$$[A] = 121 + 169 + 100 + 81 + 100 + 225 + 256 + 361 + 121 + 169 + 100 + 81 + 49 + 36 + 121 + 25$$

$$[A] = 2115$$

Step 3. Sum all individual scores together, square this sum, and divide by the total number of individual scores. Designate this value [B]:

$$\text{Sum of all scores} = 11 + 13 + 10 + 9 + 10 + 15 + 16 + 19 + 11 + 13 + 10 + 9 + 7 + 6 + 11 + 5$$

$$= 175$$

$$(175)^2 = 30625$$

$$[B] = \frac{30625}{16}$$

$$[B] = 1914.06$$

Step 4. For each group, sum the individual scores in that group and square that sum. Divide this squared sum by the number of scores in that group. Sum this result for both groups. Designate this number [C]. For the Control Group (CG):

$$CG = 11 + 13 + 10 + 9 + 10 + 15 + 16 + 19 + 11 + 13$$

$$CG = 127$$

$$(CG)^2 = (127)^2 = 16129$$

noted that the ANOVA procedure outlined below is just one of several. ANOVA is actually a flexible, powerful family of statistical procedures. The ANOVA procedure outlined below has been selected because of its general applicability to TOT methods, but the test director may wish to familiarize himself with the other procedures as well. For detailed guidance on how to apply the entire range of ANOVA procedures, see Keppel [2].

It should be noted that the ANOVA procedure is designed to evaluate the performance of two or more groups of subjects on a single measure only. Note that performance measures are often composites made up of two or more measures and that simultaneous analysis of multiple measures will require more complex analysis procedures which are beyond the scope of this appendix. Such procedures are referred to as multivariate analysis of variance (MANOVA). The test director should seek outside statistical guidance and computer support if his data and/or test objectives require this type of analysis.

Procedure for computing ANOVA. Hypothetical trials-to-criterion scores for Task Z in the aircraft for an ATD Group and a Control Group are shown below. In the following discussion, this set of hypothetical data will be analyzed to illustrate the computational procedure.

ANOVA DATA TABLE

Individual Trials to Criterion
Scores on Task Z for Trainees in an
ATD Group and a Control Group

Control Group		ATD Group	
Trainee	Score	Trainee	Score
1	11	1	10
2	13	2	9
3	10	3	7
4	9	4	6
5	10	5	11
6	15	6	5
7	16		
8	19		
9	11		
10	13		

Step 1. Arrange individual scores by group in the manner shown in the table above. Note that there are different numbers of trainees in each group.

Step 2. Square each score from both groups. Add all the squared scores together. Designate this sum [A].

$$[A] = (11)^2 + (13)^2 + (10)^2 + (9)^2 + (10)^2 + (15)^2$$
$$+ (16)^2 + (19)^2 + (11)^2 + (13)^2 + (10)^2$$
$$+ (9)^2 + (7)^2 + (6)^2 + (11)^2 + (5)^2$$

$$[A] = 121 + 169 + 100 + 81 + 100 + 225$$
$$+ 256 + 361 + 121 + 169 + 100$$
$$+ 81 + 49 + 36 + 121 + 25$$

$$[A] = 2115$$

Step 3. Sum all individual scores together, square this sum, and divide by the total number of individual scores. Designate this value [B]:

$$\text{Sum of all scores} = 11 + 13 + 10 + 9 + 10 + 15 + 16 + 19 + 11 + 13 + 10 + 9 + 7 + 6 + 11 + 5$$

$$= 175$$

$$(175)^2 = 30625$$

$$[B] = \frac{30625}{16}$$

$$[B] = 1914.06$$

Step 4. For each group, sum the individual scores in that group and square that sum. Divide this squared sum by the number of scores in that group. Sum this result for both groups. Designate this number [C]. For the Control Group (CG):

$$CG = 11 + 13 + 10 + 9 + 10 + 15 + 16 + 19 + 11 + 13$$

$$CG = 127$$

$$(CG)^2 = (127)^2 = 16129$$

The number of scores in the Control Group (N) is 10. Dividing 10 into 16129 yields:

$$\frac{(CG)^2}{N} = \frac{16129}{10} = 1612.9$$

For the ATD Group (ATD):

$$ATD = 10 + 9 + 7 + 6 + 11 + 5$$

$$ATD = 48$$

$$(ATD)^2 = (48)^2 = 2304$$

The number of scores in the ATD Group (N) is 6. Dividing 6 into 2304 yields:

$$\frac{(ATD)^2}{N} = \frac{2304}{6} = 384$$

These individual values derived from the two groups are now added together and designated [C].

$$[C] = 1612.9 + 384$$

$$[C] = 1996.9$$

The [A], [B], and [C] values are now used to calculate another value which will indicate the degree to which there are reliable performance score differences between the Control Group and the ATD Group on Task Z. This new value is referred to as the F statistic. The following three values for the hypothetical scores have now been calculated:

$$[A] = 2115$$

$$[B] = 1914.06$$

$$[C] = 1996.9$$

Step 5. Subtract [B] from [C].

$$[C] - [B] = 1996.9 - 1914.06 = 82.84$$

Step 6. Subtract [C] from [A].

[A] - [C] = 2115 - 1996.9 = 118.1

Step 7. Determine the total number of scores contained in all evaluation groups and subtract the number of evaluation groups from it. The resultant value is referred to as [df]:

[df] = Total number of scores - number of groups = 16 - 2 = 14

Step 8. Divide the number obtained in Step 6 by the [df] value obtained in Step 7.

$$\frac{118.1}{14} = 8.44.$$

Step 9. Divide the number obtained in Step 5 by the number calculated in Step 8. This value is referred to as the F statistic:

$$F = \frac{82.84}{8.44} = 9.82$$

Step 10. Turn to Table A-6. Locate df (calculated in Step 7) on the table. Read across to the three probability levels (p) and F values shown. If your calculated F value is less than the F value shown for p = .10, there is less than a 90% chance that a difference in performance exists for the two groups on Task Z. If your calculated F value is greater than the p = .10 value shown in Table A-6, but smaller than the p = .05 value, (.05), there is a 90% probability that a difference in performance exists for the two groups on Task Z and only a 10% probability that there is no difference. If the calculated F value is greater than the p = .05 value, but smaller than the p = .01 value, there is a 95% probability that a difference between the two groups exists and only a 5% chance that it does not. If your calculated F value is greater than the p = .01 F value shown, there is a 99% probability that a difference in performance exists for the two groups on Task Z and only a 1% probability that there is no difference. In our example, the F value of 9.82 is larger than that shown in Table A-6 for df = 14 and p = .01, so it can be concluded that there is less than 1 chance in 100 that there is no difference between the two groups, i.e., the ATD Group took significantly fewer times to reach criterion in the aircraft than did the Control Group.

TABLE A-6. CRITICAL VALUES OF THE F STATISTIC

[df]	p	F
3	.10	5.54
	.05	10.10
	.01	34.10
4	.10	4.54
	.05	7.71
	.01	21.2
5	.10	4.06
	.05	6.61
	.01	16.30
6	.10	3.78
	.05	5.99
	.01	13.80
7	.10	3.59
	.05	5.59
	.01	12.20
8	.10	3.46
	.05	5.32
	.01	11.30
9	.10	3.36
	.05	5.12
	.01	10.60
10	.10	3.29
	.05	4.96
	.01	10.00
11	.10	3.23
	.05	4.84
	.01	9.65
12	.10	3.18
	.05	4.75
	.01	9.33
13	.10	3.14
	.05	4.67
	.01	9.07
14	.10	3.10
	.05	4.60
	.01	8.86
15	.10	3.07
	.05	4.54
	.01	8.86
16	.10	3.05
	.05	4.49
	.01	8.53
17	.10	3.03
	.05	4.45
	.01	8.40
18	.10	3.01
	.05	4.41
	.01	8.29
19	.10	2.99
	.05	4.38
	.01	8.18
20	.10	2.97
	.05	4.35
	.01	8.10
22	.10	2.95
	.05	4.30
	.01	7.95
24	.10	2.93
	.05	4.26
	.01	7.82
26	.10	2.91
	.05	4.23
	.01	7.72
28	.10	2.89
	.05	4.20
	.01	7.64
30	.10	2.88
	.05	4.17
	.01	7.56
40	.10	2.84
	.05	4.08
	.01	7.31

Use of calculators for ANOVA calculation. The typical TOT evaluation involves the measurement of performance on more than one task. The ANOVA procedure described above will have to be repeated for every set of scores which needs to be analyzed. Doing each procedure by hand will take additional time and increases the opportunity for computational errors. In such cases where large amounts of data must be analyzed, a hand-held programmable calculator may be used to advantage. A number of such calculators are capable of having the ANOVA computational procedure "programmed" into them and thus allowing fast, accurate, and easy repetition of the procedure.

C. CORRELATIONAL STATISTICS

INTRODUCTION

Correlational statistics are used to indicate the degree to which two performance measures are related or similar. When such a relationship is shown to exist, the two measures are said to be co-related or correlated. In some instances, the relationship between two measures turns out to be strong and positive. This means that a high score on one measure is usually accompanied by a high score on the other measure. For example, there is a strong, positive correlation between the total number of hours a pilot has spent flying a particular aircraft and the proficiency with which he flies that aircraft. A strong, positive correlation exists in this example because pilots with a lot of flying experience tend to be very proficient, while pilots with little flying experience tend to be markedly less proficient. In other instances, the relationship between two measures is strong and negative. This means that a high score on one measure is usually accompanied by a low score on the other measure. For example, servicemen who are overweight are likely to score low on physical endurance tests. In other cases, relationships may be moderate (either positive or negative), indicating a general, but not overwhelming, tendency for the measures to vary together. Finally, there are instances where there is a weak, or even zero, relationship between two measures. This means that a high score on one measure indicates little or nothing about the individual's score on the other measure.

CORRELATION COEFFICIENTS

The strength of a correlation is indicated by the value of the correlation coefficient (r). The correlation coefficient can range in value from +1.00 through -1.00. A coefficient of 0.00 indicates that there is no correlation, or relationship, between two measures. A coefficient of 1.00 indicates that there is a perfect relationship between two measures. The positive or negative value of the coefficient indicates whether the two measures tend to be alike (i.e., high with high and low with low), indicating positive (+) correlation, or whether they tend to differ (i.e., high with low and low with high), indicating negative (-) correlation.

The concept of correlation can be illustrated using an X-Y coordinate graph. This is done by plotting an individual's score on one measure along the X axis of the graph and plotting the same individual's score on the second measure along the Y axis. Figure A-1 shows how this would be done for a single individual who has a score of 7 on measure A (X axis) and a score of 5 on measure B (Y axis). The

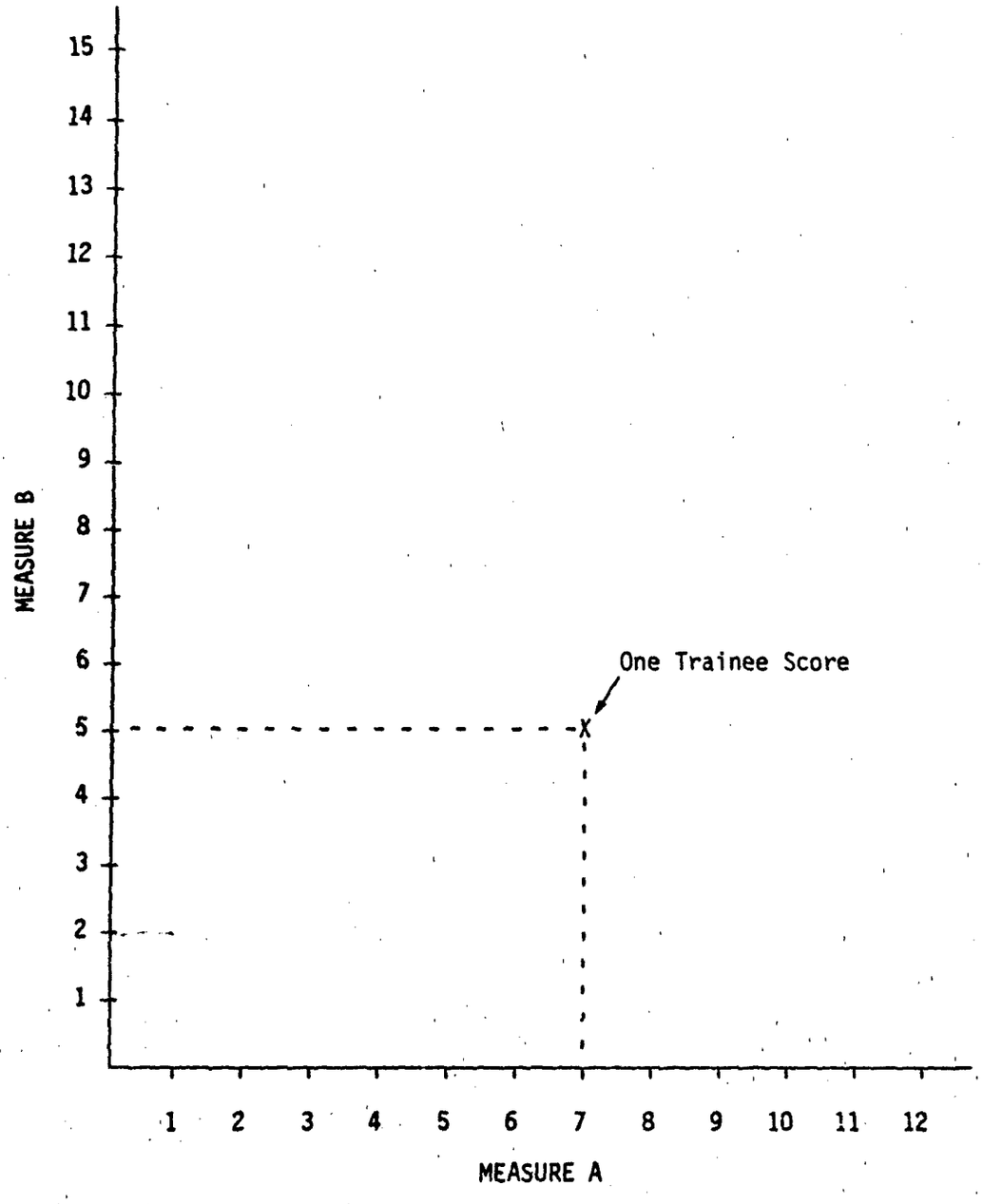


Figure A-1. Sample X-Y plot for a single trainee.

intersection of the dashed lines at the X thus represents that individual's scores on both measures. Repeating this plotting procedure for a number of individuals with scores on the two measures results in the creation of a "scatter plot" as shown in Figure A-2.

The strength of the relationship existing between two measures can be shown graphically by simply enclosing all the X's with a single line, an ellipse, or a circle as required. If a single line will do the job, the correlation is perfect ($\pm$1.00). If the ellipse is long and narrow, the correlation is very strong. As the ellipse approaches being a circle, the strength of the correlation becomes weaker. A perfect circle depicts no relationship at all. Graphic examples of some correlational relationships are depicted in Figure A-3.

CORRELATION VS. CAUSAL RELATIONSHIP

Two measures may be highly correlated for one or more of three reasons: (1) X causes Y, (2) Y causes X, or (3) both X and Y are related to or caused by some third variable.

Thus, just because two measures are highly correlated does not mean necessarily that there is a causal relationship between them. Further, even if a causal relationship does exist between two measures, such cannot be inferred on the basis of a strong correlation alone. Another basis for attributing the causal relationship must be found.

TYPES OF CORRELATION STATISTICS

Several types of correlational statistics exist. The appropriate one to use in a particular context will depend on the level of data which is to be analyzed. The test director should not use correlational procedures with nominal level data because of the numerous restrictions on their use, and because of the difficulties in interpreting correlation coefficients derived from such data. If it is absolutely necessary to show the relationship between two sets of nominal level data, a contingency coefficient may be used. (Refer to Siegel [1] for instructions on how to calculate a contingency coefficient.)

Ordinal, interval, and ratio level data can be analyzed with correlational techniques, and two common methods are discussed below. For guidance on the use of other correlational procedures, the test director is referred to Siegel [1] and Nunally [3].

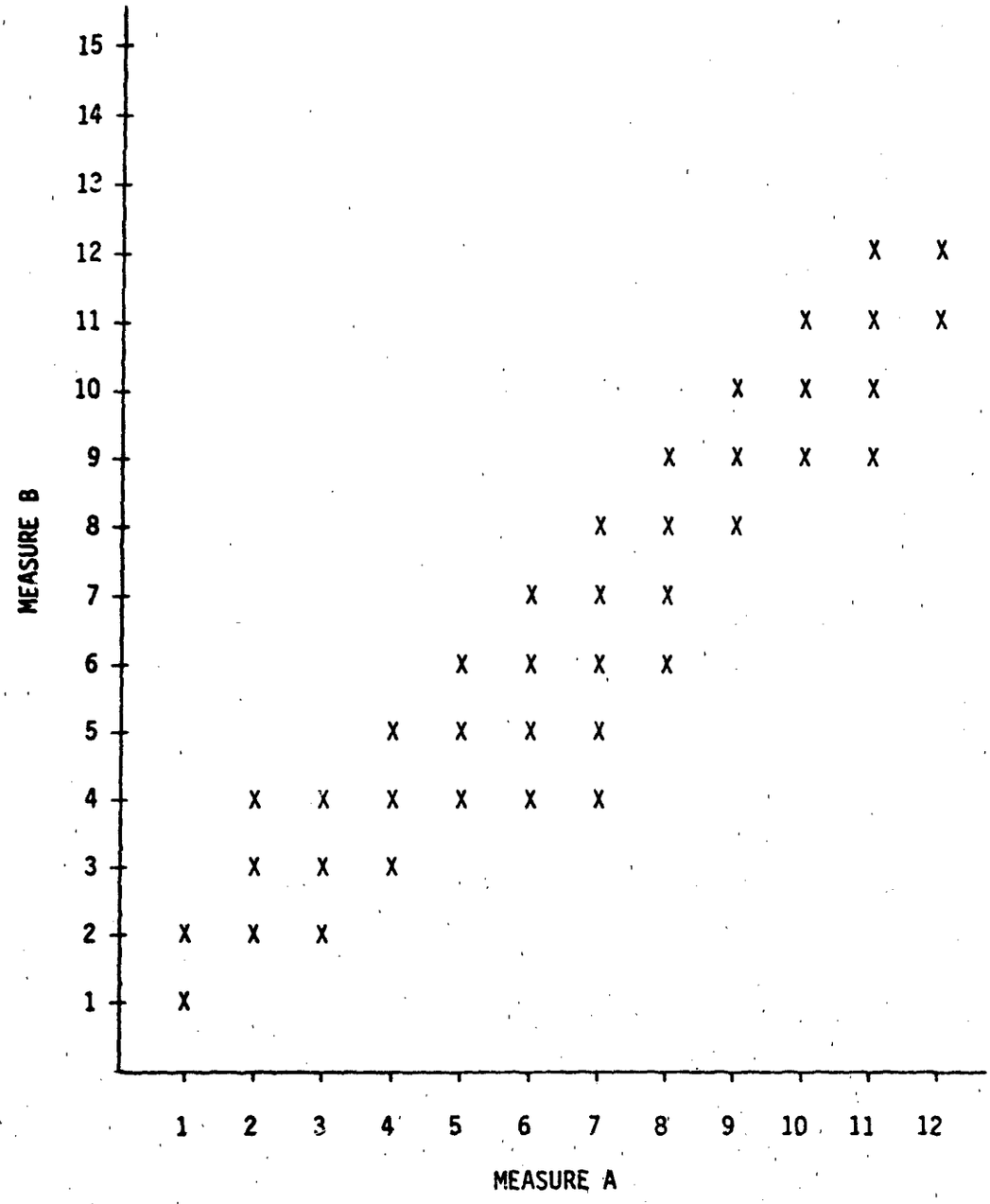


Figure A-2. Sample scatter plot for multiple trainees.

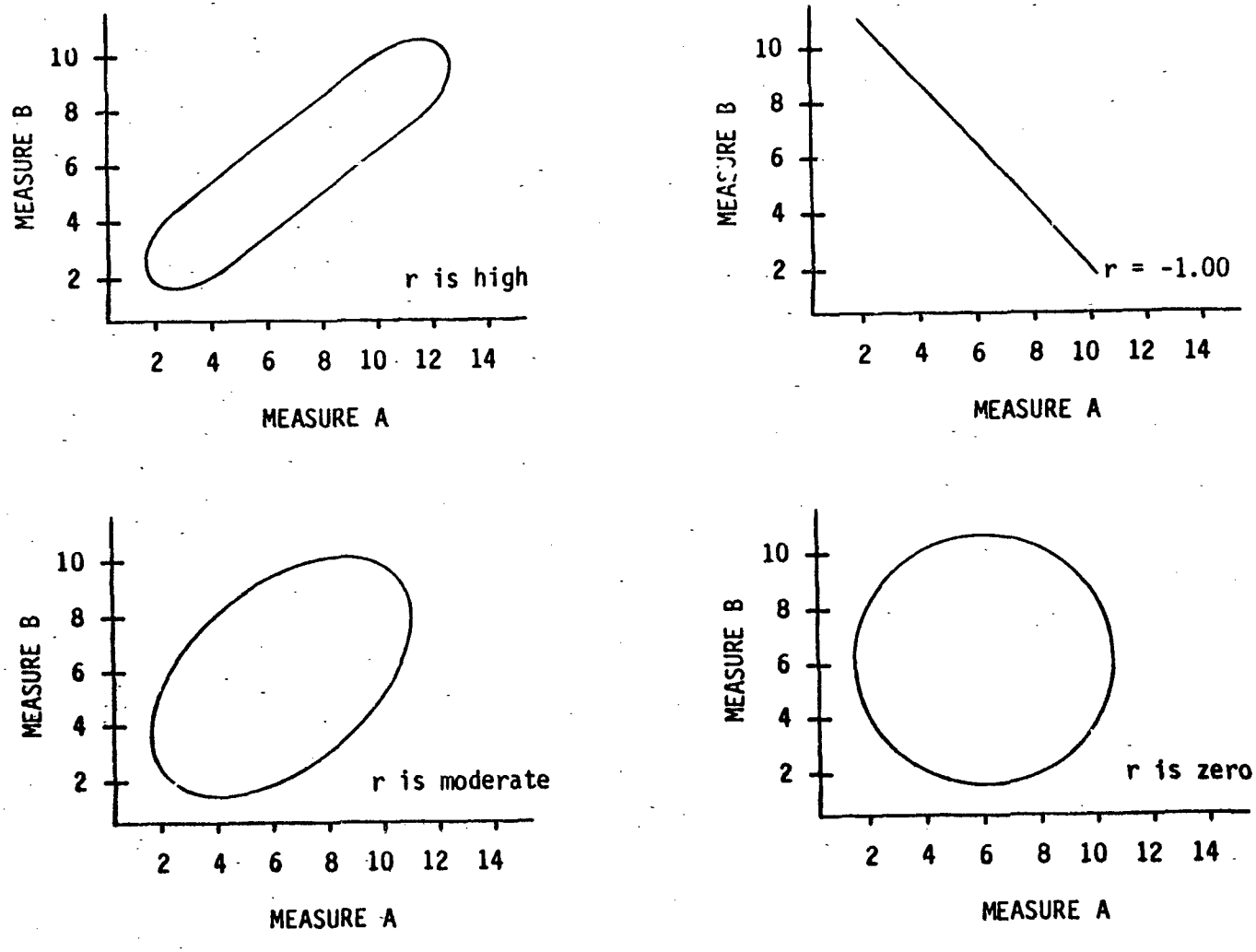


Figure A-3. Sample scatter plots for various correlations.

THE SPEARMAN RANK-ORDER CORRELATION COEFFICIENT AND ORDINAL LEVEL DATA

Spearman Rank-Order Correlation procedures are used to determine the degree to which two sets of rankings are related or similar. For example, suppose that two IPs independently rank order a group of ten trainees from lowest to highest in terms of proficiency on a given maneuver. The rankings are based on their separate observation of each trainee performing that maneuver. The question confronting a test director might be whether the two IPs have ranked the ten trainees from lowest to highest in a similar manner, or whether they have differed greatly in how they ranked them. The Spearman Rank-Order Correlation can be used to determine how the rankings of the two IPs actually are related. Recall that a strong positive correlation indicates that the two sets of rankings are very similar, whereas a weak or zero correlation indicates that there is little or no similarity in how the IPs ranked the trainees. Finally, a strong negative correlation would indicate that the IPs ranked the trainees in opposite ways.

Spearman Rank-Order Correlation Procedures

Assume that the two IPs from the above example have separately rank-ordered the ten trainees from lowest (1) to highest (10) on how well they performed Maneuver X in a particular ATD. These rankings are listed below:

Trainee	Ranks given by IP #1	Ranks given by IP #2
A	3	4
B	2	1
C	5	6
D	9	7
E	1	3
F	10	10
G	8	9
H	4	2
I	7	5
J	6	8

The Spearman Rank-Order Correlation Coefficient (r_s) can be computed for these rankings by using the following formula:

$$r_s = 1 - \frac{6 \sum D^2}{N (N^2-1)}$$

where D = the difference between two rankings of the same performance, event, individual, etc., and N = the number of pairs of rankings.

Step 1. Calculate D for each pair of rankings. In the above example, this involves subtracting the ranking of IP #2 from the ranking of IP #1:

Trainee	Ranks given by IP #1	Ranks given by IP #2	D
A	3	4	-1
B	2	1	1
C	5	6	-1
D	9	7	2
E	1	3	-2
F	10	10	0
G	8	9	-1
H	4	2	2
I	7	5	2
J	6	8	-2

Step 2. Multiply each D value by itself (square it) and then sum these D values:

Trainee	Ranks given by IP #1	Ranks given by IP #2	D	D^2
A	3	4	-1	1
B	2	1	1	1
C	5	6	-1	1
D	9	7	2	4
E	1	3	-2	4
F	10	10	0	0
G	8	9	-1	1
H	4	2	2	4
I	7	5	2	4
J	6	8	-2	4
				$\sum D^2 = 24$

Step 3. Calculate r_s with the formula previously given. In the example, $\sum D^2 = 24$, and N = 10 since ten pairs of rankings are involved:

$$r_s = 1 - \frac{(6)(24)}{10\,(100-1)}$$

$$r_s = 1 - \frac{144}{990}$$

$$r_s = 1 - .15 = .85$$

Step 4. Determine the "p" level of the obtained r_s by referring to Table A-7. Locate the appropriate N value in the left-hand column. Then read across the row to locate the r_s values associated with the various p levels. Doing this for r_s = .85, it can be seen that this value is greater than the .794 value listed for a p level of .01. This means that there is less than 1 chance out of 100 that the rankings of the two IPs are not similar. Therefore, it can confidently be inferred that the two IPs have rank ordered the trainees in a similar manner. Note that not all possible values of N are listed in Table A-7. If the exact N value being used is not listed, refer to the next lower N value to determine the p levels associated with the obtained r_s value.

TABLE A-7. CRITICAL VALUES OF r_s
(SPEARMAN RANK-ORDER CORRELATION COEFFICIENT)*

No. of pairs (N)	.10	.05	.02	.01
5	.900	1.000	1.000	--
6	.829	.886	.943	1.000
7	.714	.786	.893	.929
8	.643	.738	.833	.881
9	.600	.683	.783	.833
10	.564	.648	.746	.794
12	.506	.591	.712	.777
14	.456	.544	.645	.715
16	.425	.506	.601	.665
18	.399	.475	.564	.625
20	.377	.450	.534	.591
22	.359	.428	.508	.562
24	.343	.409	.485	.537
26	.329	.392	.465	.515
28	.317	.377	.448	.496
30	.306	.364	.432	.478

*From: Olds, E. G. Annals of Mathematical Statistics, 1938, 9; and 1949, 20.

THE PEARSON PRODUCT-MOMENT CORRELATION COEFFICIENT AND INTERVAL OR RATIO LEVEL DATA

The Pearson Product-Moment Correlation (r) is used when interval or ratio level data, such as actual performance scores, have been obtained. The formula for the Pearson r is:

$$r = \frac{N\sum AB - (\sum A)(\sum B)}{\sqrt{[N\sum A^2 - (\sum A)^2][N\sum B^2 - (\sum B)^2]}}$$

where:

- A = performance scores from Measure A
- B = performance scores from Measure B
- $\sum AB$ = the sum of all products of each person's score from Measure A multiplied by his score for Measure B.
- $\sum A$ = the sum of all performance scores from Measure A only.
- $\sum B$ = the sum of all performance scores from Measure B only.
- $\sum A^2$ = the sum of all squared performance measure scores from Measure A (after each has been individually squared).
- $\sum B^2$ = the sum of all squared performance measure scores from Measure B (after each has been individually squared).
- $(\sum A)^2$ = the square of the sum of all performance scores from Measure A.
- $(\sum B)^2$ = the square of the sum of all performance scores from Measure B.
- N = the total number of pairs of performance scores.

Although formidable looking, the basic computation is easily accomplished with a four-function calculator.

Assume that the following performance scores have been obtained for ten trainees on performance Measures A and B.

Trainee	Measure A	Measure B
1	28	9
2	27	10
3	31	13
4	18	5
5	25	8
6	32	15
7	32	16
8	27	13
9	17	6
10	26	8

Step 1. Create a new table with the format used below:

List the scores for each trainee on performance measure A and performance measure B where indicated. Caution: Do not change the order in which the scores are arranged within columns. Each trainee's scores for measure A and measure B must be directly across from each other when this table is constructed.

Trainee	A	A^2	B	B^2	AB
1	28		9		
2	27		10		
3	31		13		
4	18		5		
5	25		8		
6	32		15		
7	32		16		
8	27		13		
9	17		6		
10	26		8		

Step 2. First, square each A score and each B score individually and list them under the A^2 and B^2 column heading, respectively. Multiply each individual's A score and B score together and list this product in the AB column.

Finally, sum the numbers in each column:

A	A^2	B	B^2	AB
28	784	9	81	252
27	729	10	100	270
31	961	13	169	403
18	324	5	25	90
25	625	8	64	200
32	1024	15	225	480
32	1024	16	256	512
27	729	13	169	351
17	289	6	36	102
26	676	8	64	208

$\sum A = 263 \quad \sum A^2 = 7165 \quad \sum B = 103 \quad \sum B^2 = 1189 \quad \sum AB = 2868$

Step 3. All the values which must be used in the computational formula have now been calculated. For these hypothetical data, N = 10, the number of score pairs. Substitute these values into the formula:

$$r = \frac{(10)(2868) - (263)(103)}{\sqrt{[(10)(7165) - (263)^2]\ [(10)(1189) - (103)^2]}}$$

Step 4. Calculate r:

$$r = \frac{28{,}680 - 27{,}089}{\sqrt{[(71{,}650 - 69{,}169)]\ [(11{,}890 - 10{,}609)]}}$$

$$r = \frac{1591}{\sqrt{(2481)(1281)}}$$

$$r = \frac{1591}{\sqrt{3{,}178{,}161}}$$

$$r = \frac{1591}{1782.7} = .89$$

Step 5. Subtract 2 (because there are 2 scores in each pair) from N. This value is df. In the example:

df = 10 - 2 = 8.

Step 6. Refer to Table A-8. Find the appropriate df in the left-most column. Read across this row. The row contains four r values. The left-most r value represents the minimum r value which must be obtained when comparing only ten pairs of scores to ensure that there are only 10 chances in 100 that the obtained r value is greater than zero. If the obtained r value is less than this value, there is a greater than 10% chance that the obtained r value does not reflect a true relationship between the two performance measures that is greater than zero. The next r value in the table is the minimum value needed to suggest that there are only 5 chances in 100 that the true relationship between these measures is zero. Finally, the last two r values are the minimum values needed to say that there are only 2 in 100 and 1 in 100 chances, respectively, that the obtained r value does not differ from zero.

From Table A-8, it can be seen that there is less than 1 chance in 100 that there is no correlation between performance measures A and B in our example. The relationship between the A and the B measures in the example (r = +.89) is strong and positive. This means that, in general, subjects high in A will also tend to be high in B, and that trainee's low in A will tend to score low in B. Since the relationship is not perfect--that is, r is not 1.00--there will be some instances in which a high score in A will not be accompanied by a correspondingly high score in B. To a large extent, however, (when r = +.89) high scores on one measure will tend to be accompanied by high scores in another, and low scores on one measure will tend to go with low scores in the second measure.

Correlation statistics can be helpful, in that once a relationship has been determined between two measures, it becomes possible to make predictions about how a group of trainees will perform in a second situation if it is known what they have done in the first. The relationship would have to be nearly 1.00, or at least very strong, to allow very precise prediction in the individual case, since even a moderately strong relationship (e.g., r = +.75) would allow many exceptions to the rule that high scores go with high and that low scores go with low. However, in a group situation one could accurately predict how many trainees could be expected to perform at a given level in a second situation if one were provided the information about performance in the first and the degree of relationship existing between the measures in question.

TABLE A-8. CRITICAL VALUES OF THE PEARSON r*

df = N - 2; N = number of pairs	Probability Levels			
	.10	.05	.02	.01
1	.988	.997	.9995	.9999
2	.900	.950	.980	.990
3	.805	.878	.934	.959
4	.729	.811	.882	.917
5	.669	.754	.833	.874
6	.622	.707	.789	.834
7	.582	.666	.750	.798
8	.549	.632	.716	.765
9	.521	.602	.685	.735
10	.497	.576	.658	.708
11	.476	.553	.634	.684
12	.458	.532	.612	.661
13	.441	.514	.592	.641
14	.426	.497	.574	.623
15	.412	.482	.558	.606
16	.400	.468	.542	.590
17	.389	.456	.528	.575
18	.378	.444	.516	.561
19	.369	.433	.503	.549
20	.360	.423	.492	.537
21	.352	.413	.482	.526
22	.344	.404	.472	.515
23	.337	.396	.462	.505
24	.330	.388	.453	.496
25	.323	.381	.445	.487
26	.317	.374	.437	.479
27	.311	.367	.430	.471
28	.306	.361	.423	.463
29	.301	.355	.416	.456
30	.296	.349	.409	.449
35	.275	.325	.381	.418
40	.257	.304	.358	.393
45	.243	.288	.338	.372
50	.231	.273	.322	.354
60	.211	.250	.295	.325
70	.195	.232	.274	.302
80	.183	.217	.256	.283
90	.173	.205	.242	.267
100	.164	.195	.230	.254

*Table is abridged from Table IV of. Fisher, R. A., & Yates, F. Statistical tables for biological, agricultural and medical research (6th edition). Edinburgh: Oliver and Boyd, 1963.

D. ASSESSING EVALUATION INSTRUMENT RELIABILITY AND VALIDITY

INTRODUCTION

Regardless of the format of a measurement instrument, whether it be a scale rating or questionnaire, there are certain minimum technical requirements which must be met if the instrument is to provide data useful to an OT&E effort. Some of these requirements, such as question loading or biasing, have been discussed in Chapters 5 and 6. However, two central technical considerations have not been treated: measurement reliability and validity.

Pretesting the Questionnaire

Reliability and validity are always of concern in the development and use of evaluation instruments. The general and specific guidelines offered concerning construction of questionnaires and rating scales will aid in the production of instruments that are reliable and valid. However, there are statistical techniques for assessing reliability and validity with which the test director should be familiar should it become necessary that he demonstrate the reliability or validity of his measures. These techniques are based on the Pearson correlational statistic described in Section C of this appendix. The following discussion describes these techniques as they might be used with a rating scale or questionnaire to produce valid and reliable instruments to use to measure various attributes such as combat readiness, ATD fidelity, etc.

A number of procedures for estimating the reliability and validity of measures are described. However, there are no absolute rules which will indicate to the test director that he should use one procedure rather than another. The choice of any procedures for estimating reliability and validity depends on the purposes of the OT&E and the type of measure evaluation considered most relevant.

Reliability and validity can be assessed at various times within the evaluation process. For example, the content validity of a measure should be assessed before it is ever administered. If at all possible, a questionnaire should be pretested to gain an initial estimate of its reliability, and then it should be amended as necessary.

Finally, it should be noted that no measurement instrument developed by the test director will ever be completely reliable or valid because of restrictions on time, resources, and testing environment flexibility. However, this is true for any measurement instrument regardless of the context involved. The job of the test director is to assure that the measurement instruments involved are adequate and

appropriate for the use to which they are to be put. This will, of course, involve the procurement or development of reliable and valid measurement instruments.

Reliability

Reliability is a measure of consistency. Reliability depends on the amount of error in a measure. In this context, error does not refer to right or wrong, but to variability or discrepancies in obtained scores that cannot be attributed to a specific, identifiable, systematic cause. These discrepancies are referred to as unsystematic error. As the amount of unsystematic error in scores obtained using a given measurement instrument increases, the reliability of that instrument and of those scores decreases. Thus, a measurement instrument is reliable to the extent that: (1) it produces the same results when measuring a given thing on different occasions; (2) different scorers using the instrument arrive at the same score; and (3) different (equivalent) forms of the measure produce the same results.

Consider the following hypothetical example. A simple, inexpensive, self-assessment questionnaire of pilot combat readiness is developed for use with F-15 fighter pilots. The test is to be administered once a month to all mission ready pilots. If the monthly questionnaire scores obtained for each pilot vary widely when nothing about the pilot or his flying skills has actually changed, then there is substantial reason to believe that this new test is not reliable. That is to say, there is wide variability in the obtained scores which cannot be attributed to changes in the pilot or his skills. Thus, a significant amount of unsystematic error would be present in the scores obtained using the instrument. Given this outcome, the questionnaire should be revised. On the other hand, if the obtained scores of a group of pilots remain consistent over a time period when factors affecting their combat readiness remain constant, then the test can be said to be reliable. In the same vein, if different instructors or test directors cannot obtain the same results when administering the questionnaire to the same group of subjects, or if different forms of the questionnaire give different results from the same group of trainees, then the reliability of the questionnaire is suspect.

The reliability of a questionnaire or rating scale can be determined in three ways: (1) determining the degree of agreement (correlation) between the scores obtained on two separate administrations of a questionnaire (test-retest reliability); (2) determining the internal reliability of a questionnaire by correlating scores obtained on one half of the instrument with scores obtained on the other half of the instrument (split-half reliability); or (3) determining the correlation between alternate forms of the same questionnaire (alternate-form reliability). Each of these methods is

discussed in detail below. Normally, the test director will only be interested in Test-Retest and alternate forms reliability measures during the normal OT&E process. However, there may be instances when a measure of internal reliability such as split-half reliability may also be useful. For this reason, guidance on its use is also included below.

Test-retest reliability. The most direct way of estimating the reliability of a questionnaire or rating scale is to administer it more than once to the same group of subjects. By doing this, a test-retest reliability coefficient may be computed by correlating the two sets of scores obtained from the separate administrations of the instrument. If there is a strong, positive correlation between the two sets of scores, then the instrument is probably reliable. If, however, there is a weak or negative correlation between the two sets of scores, the instrument should be considered unreliable.

As an example, assume that the reliability of a questionnaire is to be determined by employing the test-retest method. This can be accomplished by obtaining a set of scores from an initial administration of the questionnaire to a group of subjects. Assume that any subject can receive an overall score between 0 and 100 based on his responses to all questionnaire items, and that 10 subjects were included in the subject group. The same questionnaire is readministered to the same group of subjects after some period of time has passed. This time period may vary, depending on the application of the questionnaire. Assume that the two administrations of the questionnaire yielded scores as follows:

Subject	First administration	Second administration
A	86	81
B	55	53
C	19	21
D	92	92
E	16	19
F	56	61
G	57	55
H	36	37
I	72	71
J	81	80

The degree of correspondence between these two sets of scores is then determined by calculating a Pearson correlation coefficient (r) using the procedure shown in Section C of this appendix. Following this procedure for the two sets of scores listed above produces a correlation coefficient, r, of + 0.99. This means that subject scores obtained on the two administrations of the questionnaire are very similar (the strongest possible value for r is 1.00).

This result shows that the questionnaire is reliable, because it yielded similar results when administered at two separate points in time. An indication that the questionnaire was not reliable would have been obtained if the scores on the second administration were not similar, i.e., bore little or no relationship to those obtained on the first, as would have been indicated if a weak or near zero r value had been obtained.

There are several disadvantages to using the test-retest method. First, a person's performance on the retest may be influenced by the first administration of the instrument. For example, he may remember his previous responses and seek to reproduce them rather than reflect his current feelings. In this case, we would have a good measure of the person's memory, but not of the reliability of the instrument. Second, even if the sets of overall scores are highly correlated, the instrument still can be unreliable if a person's responses to individual items on the questionnaire or rating scale are inconsistent from one occasion to the next.

Split-half reliability. The internal consistency of an instrument is a product of the overall agreement among all the items that make up the measure. The simplest method of determining this "internal consistency" aspect of a questionnaire, for example, is to divide the items that make up the instrument into two halves, and to calculate the correlation between scores obtained from the two halves. The items can be halved either by dividing the measure into "first" and "second" halves of the test, or by taking the odd-numbered items and placing them in one half, and the even-numbered items and placing them in the other half. Once the scores are divided into halves, regardless of the method, the two set of scores can then be correlated. As above, a strong positive correlation is again indicative of a reliable (in this case internally reliable) questionnaire, whereas weak or zero correlation coefficients indicate an internally unreliable questionnaire.

For example, assume that ten subjects receive the following overall, first half, and second half scores on a questionnaire. Overall scores can range from 0 to 100. Each subject's overall score is based on his responses to all of the questionnaire items. To determine the split-half reliability of the questionnaire, the subjects' responses on the first half of the questionnaire can be correlated with their responses to the second half.

Subject	Overall score obtained from entire questionnaire	Overall score on first half of questionnaire	Overall score on second half of questionnaire
A	88	45	43
B	56	25	31
C	74	40	34
D	83	40	43
E	39	19	20
F	40	22	18
G	96	45	51
H	18	9	9
I	57	28	29
J	64	33	31

Correlating the two sets of half scores yields a correlation coefficient of +.95. This indicates that the questionnaire is internally consistent because subjects respond to the two halves of the questionnaire in a similar fashion. If the correlation coefficient had been weak, for example +.27, then it would be obvious that subjects were responding differently to the two halves of the questionnaire and that it does not have strong internal consistency.

Alternate forms reliability. Sometimes it is necessary to construct and administer more than one version of a questionnaire. Alternate forms of the same questionnaire can be used when subjects must provide the same type of information at different times, and when it is important that responses made at an earlier administration not influence responses at a later administration. For example, alternate forms of a questionnaire might be used to assess changes in attitudes toward simulation before and after experience with a given ATD.

In the alternate forms procedure, two questionnaires that are equivalent are administered to the same group of subjects, and the scores obtained for each subject on the two questionnaires are then correlated.

As an example of how to assess the reliability of alternate forms of a questionnaire, assume that a test director wants to determine the effect of experience with a new state-of-the-art simulator on training effectiveness evaluations of the simulator. To assess accurately any changes in their evaluation of the ATD, the IPs must be questioned before and after each has had experience with the new simulator. Thus, the test director will want to administer different (but equivalent) questionnaire forms to the IPs before and after ATD experience. However, he must first be certain that the alternate forms of the questionnaire are equivalent, so that he can be sure that any differences obtained are indeed due to exposure to the new simulator and not to differences in the questionnaire forms themselves.

To assess the alternate-forms reliability of the questionnaires, (i.e., Form A and Form B), the test director could administer both forms of the questionnaire to a group of IPs before ATD experience, and both forms to this same group after ATD experience. If the two forms are equivalent or reliable, Forms A and B should produce similar responses from the IPs before they have received experience with the ATD, and, by the same token, they should produce similar responses from the IPs after they have become experienced with the ATD. However, because exposure to the ATD will likely alter IP estimates of ATD effectiveness, the overall group of "before" scores obtained from Forms A and B may well differ from the overall group of "after" scores obtained from the same two forms.

To continue the above example, assume that the following overall effectiveness scores (which can vary from 1 to 10) were obtained:

	Subject	Form A score	Form B score
Before ATD Experience	A	6	7
	B	3	5
	C	7	6
	D	1	2
	E	4	4
	F	3	2
	G	1	1
	H	7	7
	I	5	4
	J	4	5

	Subject	Form A score	Form B score
After ATD Experience	A	8	9
	B	10	10
	C	8	8
	D	5	6
	E	7	7
	F	8	7
	G	7	8
	H	9	9
	I	10	10
	J	9	10

The above data show that, overall, the "before" scores from Forms A and B are lower than the overall "after" scores obtained from the same two forms. This indicates that IP estimates of ATD effectiveness, in general, become more positive after experience with the new

simulator. However, to assess alternate forms reliability between Forms A and B, the correlation between the two sets of "before" scores, and the correlation between the two sets of "after" scores must be determined and evaluated separately. The important issue in determining form equivalence or reliability is whether or not the two forms measure IP training effectiveness estimates of the ATD in the same way, first before, and then after, they have been given experience with the new simulator.

The correlation coefficient between Form A and Form B scores before new simulator experience is +.89. The correlation coefficient between Form A and Form B scores after new simulator experience is +.90. Both scores are strong and positive, indicating that Form A and Form B are similar regardless of the context. This means that, insofar as reliability is concerned, the test director can use these two equivalent forms interchangeably in the future.

Which of these three methods a test director chooses to employ to assess the reliability of a questionnaire or rating scale will depend on the time and resources available to him in a particular testing environment. For example, the test-retest method will take more calendar time and resources to use than will the split-half method because of the need to administer the questionnaire or rating scale twice.

Validity

Validity refers to the degree to which the instrument measures what it is intended to measure. This is the most essential characteristic of a measurement instrument--it should fulfill the purpose for which it was designed. This definition can be somewhat misleading, in that it implies that there is only one type of validity associated with a measurement instrument. Actually, there are several types. Each has a specific purpose, use, and assessment methodology associated with it.

Unlike reliability, which is affected only by unsystematic errors of measurement, the validity of a test is affected by both unsystematic and systematic errors as defined in the previous section. This implies that a measurement instrument may be reliable without being valid, in the sense that it can consistently measure something, but not what it was intended or designed to measure. However, an instrument can rever be valid unless it has some degree of reliability. In other words, reliability is a necessary, but not sufficient, condition for validity.

Predictive validity. The predictive validity of a measurement instrument refers to the degree to which scores obtained on the

instrument allow the prediction of performance on some criterion behavior. As above, the strength of this predictive relationship is determined by correlating the two sets of scores.

For example, the ratings enlisted men receive on the Airman Performance Report (APR) might be used to predict their performance on a job proficiency test. In this case, the criterion, job proficiency, provides a direct measure of the relevant behavior. The extent to which APR ratings predict job proficiency is a measure of the predictive validity of the APR as a job proficiency prediction instrument.

If there is a high correlation between APR ratings and job proficiency, then this measure exhibits high predictive validity. On the other hand, if there is a low correlation between APR ratings and job proficiency, the measure exhibits low predictive validity.

Content validity. Content validity is the extent to which the statements contained in a questionnaire or rating scale address all relevant components that contribute to a particular subject's response. It is possible to devise an instrument that does not measure the complete range of issues in which the researcher is interested. For example, a researcher might construct a measure of potential flying ability which, in fact, contained only statements relating to the physical fitness of the individual. The result would be a measure of physical vigor, not a measure of all the abilities which contribute to the operation of an aircraft.

Content validity is an important check during the crucial first stage of any measurement process. In this stage the researcher chooses the statements (items or questions) from which he will develop his measurement scale. As a necessary first step he needs to decide how specific, or how general, he wants his measures to be. Next, a list of all the components or elements that contribute to the issue in question must be constructed. For example, if a measure of potential flying ability were being constructed, general categories of abilities could be listed, such as physical vigor, psychomotor skills, decision making ability, and ability to cope with stress. Each general category can then be further subdivided.

Once the "content" listing is complete, actual items or statements can be developed in accord with the guidance previously given. The purpose of the content listing is to ensure that the range of statements corresponds to the range of components that comprise the attribute to be measured.

The greatest shortcoming of content validity is that the adequacy of the original content list is limited by the foresight and ability of the individual preparing the list. Hence, content validity is assessed ultimately only by the researcher's own judgment. Despite this limitation, attention to content validation should be part of every measure development.

REFERENCES

1. Siegel, S. Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill, 1956.

2. Keppel, G. Design and analysis: A researcher's handbook. Englewood Cliffs, NJ: Prentice-Hall, 1973.

3. Nunally, J. Psychometric theory. New York: McGraw-Hill, 1967.

☆U.S. GOVERNMENT PRINTING OFFICE:1982- 569-054/2