DOCUMENT RESUME

ED 224 960 CE 034 751

AUTHOR	Sorensen, Philip H.; Pennell, Roger
TITLE	Technical Training: Development of Instructional Treatment Alternatives. Final Report.
INSTITUTION	SRI International, Menlo Park, Calif.
SPONS AGENCY	Air Force Human Resources Lab., Lowry AFB, Colo. Technical Training Div.
REPORT NO	AFHRL-TR-82-32
PUB DATE	Nov 82
CONTRACT	MDA903-79-C-0393
NOTE	160p.
PUB TYPE	Reports - Research/Technical (143) -- Guides - Non-Classroom Use (055)
EDRS PRICE	MF01/PC07 Plus Postage.
DESCRIPTORS	Adults; *Computer Managed Instruction; Criterion Referenced Tests; *Evaluation Methods; Independent Study; *Learning Modalities; Military Training; Models; *Predictive Measurement; Predictive Validity; *Programed Instructional Materials; *Teaching Methods; Test Items; Test Validity
IDENTIFIERS	Air Force

ABSTRACT

This report was written to provide guidance in the development and evaluation of alternative instructional approaches that hold promise of improving instructional effectiveness. The main focus of the report is on how to identify and test interactive relationships between individual differences among learners and instructional conditions or treatments. The report directs considerable attention to problems of measurement that are basic to diagnosis and evaluation. The report rests on the conviction that no acceptable substitute for careful empirical experimentation exists; an approach must be tried, often several times, before evidence sufficient for credible evaluation is available. Underlying this conviction is commitment to principles of measurement. Detecting and quantifying the effects of instruction require dependable measurement. Thus, a portion of this report concerns (1) the development of measures of learner aptitudes (e.g., the learner's repertoire of knowledge, skills, and abilities) as a basis for assignment to instructional treatment and (2) learner achievement as a function of instructional treatment. Particular emphasis is given to the importance of homogeneity among items that comprise a test of achievement because a reliable test cannot be composed of nonhomogeneous items. Also emphasized is that an unreliable test cannot be useful for either diagnosis of evaluation. (Author/KC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

AFHRL-TR-82-32

AIR FORCE HUMAN RESOURCES LABORATORY

TECHNICAL TRAINING: DEVELOPMENT OF INSTRUCTIONAL TREATMENT ALTERNATIVES

By

Philip H. Sorensen

SRI International
333 Ravenswood Avenue
Menlo Park, California 94025

Roger Pennell

LOGISTICS AND TECHNICAL TRAINING DIVISION
Technical Training Branch
Lowry Air Force Base, Colorado 80230

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official NIE position or policy

November 1982

Final Report

Approved for public release; distribution unlimited.

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235

ED224960
CE034751

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.

ROGER PENNELL
Contract Monitor

JOSEPH A. BIRT, Lt Col, USAF
Technical Director, Logistics and Technical 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

1. REPORT NUMBER
AFHRL-TR-82-32

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE *(and Subtitle)*
TECHNICAL TRAINING: DEVELOPMENT OF INSTRUCTIONAL TREATMENT ALTERNATIVES

5. TYPE OF REPORT & PERIOD COVERED
Final

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR*(s)*
Philip H. Sorensen
Roger Pennell

8. CONTRACT OR GRANT NUMBER*(s)*
MDA903-79-C-0393

9. PERFORMING ORGANIZATION NAME AND ADDRESS
SRI International
333 Ravenswood Avenue
Menlo Park, California 94025

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
61102F
2313T212

11. CONTROLLING OFFICE NAME AND ADDRESS
HQ Air Force Human Resources Laboratory (AFSC)
Brooks Air Force Base, Texas 78235

12. REPORT DATE
November 1982

13. NUMBER OF PAGES
158

14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*
Logistics and Technical Training Division
Technical Training Branch
Air Force Human Resources Laboratory
Lowry Air Force Base, Colorado 80230

15. SECURITY CLASS *(of this report)*
Unclassified

15.a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

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

18. SUPPLEMENTARY NOTES

This contract was partially funded by the Defense Advanced Research Projects Agency.

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

alternative instructional approaches
evaluation
individual differences
instruction
Instructional System Development (ISD)
instructional treatment
learner aptitudes
measurement
statistical methodology
testing
training

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

The purpose of this report is to provide guidance in the development and evaluation of alternative instructional approaches that hold promise of improving instructional effectiveness. The main focus of the report is on how to identify and test interactive relationships between individual differences among learners and instructional conditions or treatments. The report directs considerable attention to problems of measurement that are basic to diagnosis and

DD Form 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

Item 20 (*Continued*)

evaluation. The report rests on the conviction that there is no acceptable substitute for careful empirical experimentation; an approach must be tried, often several times, before evidence sufficient for credible evaluation is available. Underlying this conviction is commitment to principles of measurement. Identifying and defining learner aptitudes require dependable measurement. Detecting and quantifying the effects of instruction requires dependable measurement. Thus, a portion of this report concerns the development of measures of (a) learner aptitudes (e.g., the learner's repertoire of knowledge, skills, and abilities) as a basis for assignment to instructional treatment and (b) learner achievement as a function of instructional treatment. Particular emphasis is given to the importance of homogeneity among items that comprise a test of achievement because a reliable test cannot be composed of nonhomogeneous items, and an unreliable test cannot be useful for either diagnosis or evaluation.

SUMMARY

Objective

The objective is to explicate a methodology to assess the need for, and to assist in the development, implementation, and evaluation of alternative instructional treatments especially applicable to self-paced, computer-managed instructional settings.

Background/Rationale

If a computer-managed instructional program were divided into alternative instructional modules or lessons, the instructional manager would want its use optimized to lead students from lesson to lesson as efficiently as possible. The manager would want the various lesson alternatives selected to be those that are most appropriate to the particular characteristics of individual learners. Whereas some lesson approaches might be acceptable to most learners, different approaches might be better for other learners. Methods exist for deciding which lesson approaches are most appropriate for which learners. These methods need to be explicated in a manner that can be used by Air Force instructional managers, especially those working in self-paced, computer-managed instructional settings.

Approach

A suitable methodology should be able (a) to identify lesson approaches suitable for most students and the student characteristics that seem to be related to lesson success, (b) to suggest more-suitable instructional approaches for students who have different characteristics, and (c) to implement and evaluate the effectiveness of the alternative approaches. Accordingly, emphasis is placed on how to identify and test interactive relations between individual-learner characteristics and instructional conditions or treatments. Considerable attention is devoted to problems of measurement that are basic to instructional diagnosis and evaluation, as well as to the development of measures of learner aptitudes and achievement. Finally, statistical methods are outlined for use in the design and analysis of experiments to evaluate alternative instructional treatments.

Specifics

The report assumes that users are familiar with basic statistical concepts and the rudiments of experimental design. References are cited for those users who may wish to review statistical and measurement concepts, because some understanding is required of measures of central tendency, variance about a central value, and relations between such measures.

The first of three sections deals with basic concepts of evaluation of alternative instructional treatments. It includes a generalized model for evaluation, an example of an instructional evaluation, and techniques for planning experiments and evaluating treatments. The second section deals with tests and test items in a criterion-referenced setting. It includes concepts of measurement using such test items and techniques for selecting and evaluating them. The third section deals with the design and evaluation of alternative instructional treatments. It includes the methodology required to assess the need for an alternative instructional treatment, then to develop, implement, and evaluate it.

Five appendixes provide technical details and an example of the overall methodology using specific data. They cover the following topics: (a) evaluation of candidate test items, (b) development of criterion-referenced tests using cross-sectional samples, (c) regression analysis as applied to the development of experimental treatments, (d) analysis of learner characteristics in the design of treatments, and (e) commonly encountered statistical concepts.

Conclusions/Recommendations

A methodology has been explicated to assist in the development, implementation, and evaluation of alternative instructional treatments. Also, the conceptual framework and overall methodology needed for the improvement of test items and tests, and for the development of treatments in typical Air Force criterion-referenced settings, have been presented. These methods should be used by managers of Air Force instructional programs, especially in self-paced, computer-managed instructional settings, and to this end this report should provide useful guidance.

PREFACE

The purpose of this report is to provide guidance in the development and evaluation of alternative approaches that hold promise of improving instructional effectiveness. The main focus of the report is on how to identify and test interactive relationships between individual differences among learners and instructional conditions or treatments.

It is assumed throughout this report that the instructional setting permits individualized management of instruction. Management support for experimentation with instructional activities also is assumed. Although computer assistance for instruction is not a requisite, many procedures and recommendations in the report would be enhanced if instruction was computer-managed or computer-assisted.

Many of the examples used in this report are based on hypothetical test item response data generated to provide an illustration that is internally consistent. Other concrete examples are drawn from analyses of the Precision Measuring Equipment (PME) course taught at the Lowry Technical Training Center (LTTC) at Lowry AFB.

The report directs considerable attention to problems of measurement that are basic to diagnosis and evaluation. A major premise is that there is no acceptable substitute for careful empirical experimentation -- often an approach must be tried several times before evidence sufficient for credible evaluation is available. Underlying this conviction is commitment to principles of measurement because defining learner aptitudes and quantifying the effects of instruction both require dependable measurement. Thus, a portion of this report concerns the development of measures of learner aptitudes (e.g., the learner's repertoire of knowledge, skills, and abilities) as a basis for assignnent to instructional treatment and also measures of learner achievement as a function of instructional treatment. The importance of homogeneity among items that comprise a test of achievement is emphasized because a reliable test cannot be composed of nonhomogeneous items nor can an unreliable test be useful for either diagnosis or evaluation.

The report assumes that readers are familiar with, and have ready access to, the many useful suggestions contained in Air Force Manual (AFM) 50-2, Interservice Procedures for Instructional Systems Development (ISD). This report is compatible with the ISD model and refers to it for supplemental guidance.

Finally, this report assumes that users are familiar with basic statistical concepts and the rudiments of experimental design. The discussions and examples in the report require at least some understanding of measures of central tendency, variance about a central value, and association between measures. References are cited for those users who may wish to review statistical and measurement concepts.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Figure

LIST OF TABLES

I BASIC CONCEPTS AND PRIORITIES

A Generalized Model for Evaluation

The causal directions of selected influences upon performance in a training course are illustrated in Figure 1. This figure seeks to represent the increasingly rich mixture of factors that influence performance on successive lessons[1] within a course and in a job assignment following course completion.

Figure 1 intentionally ignores the broad class of other influences defined by the environment within which instruction occurs. The boxes labelled as "lessons" in Figure 1 represent lesson content, methods of instruction used, and the immediate environmental setting in which instruction occurs. Other influences on performance may be important but they are implied rather than shown.

Figure 1 may be viewed as a general model for prediction of performance. For example, if "performance" in Lesson 1 is the variable to be predicted, the primary predictor variables are those that define learner aptitudes and experiences prior to exposure to Lesson 1. ("Aptitude," as used here, follows the broad definition by Cronbach and Snow (1977); that is, aptitude is "any characteristic of a person that forecasts his probability of success under a given treatment.") The content, methods, and setting of instruction in Lesson 1 define the "treatment" variables in the equation.

Influences on performance accumulate and merge with each successive lesson in the course, as the horizontal arrows between each lesson and the left-hand box, "cumulative proficiency," are intended to show. Thus, the prediction equation for each successive lesson is incrementally more complex than the preceding one. By the final lesson in the course -- Lesson n in Figure 1 -- the predictor variables have been augmented by the cumulative effects of all instructional experiences in the course to that point.

---

[1] The term, "lesson," is used throughout this report to denote any segment, unit, or block of instruction from which progress to the next segment cannot occur without successful performance on a mastery test. "Successful performance" is defined by the criterion level specified for each test of achievement in "lesson" content.

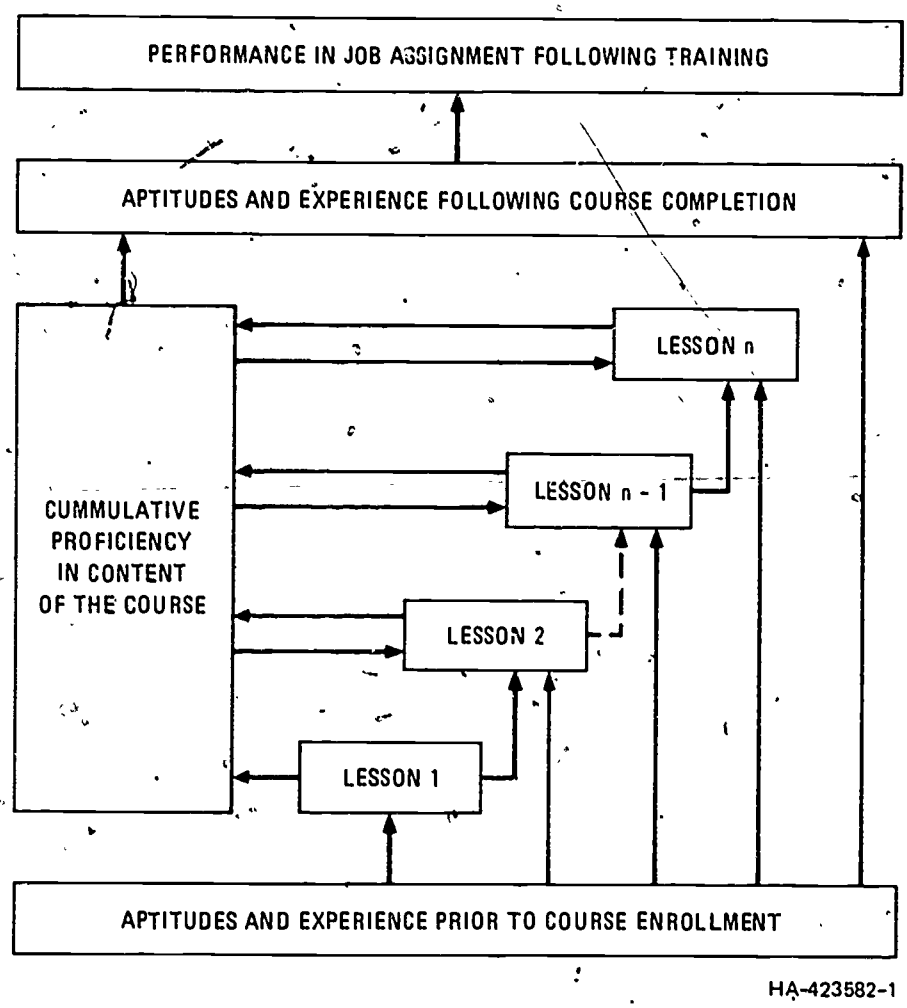

FIGURE 1 RELATIONSHIPS AMONG APTITUDES, EXPERIENCE, LESSON ACHIEVEMENT, COURSE ACHIEVEMENT, AND SUBSEQUENT JOB PERFORMANCE

The aptitudes that describe a person after completion of a course include the cumulative changes acquired during the course (e.g., new skills, increased knowledge, changed attitudes) as well as the surviving characteristics from among those that described the person at the beginning of the course. The "new" profile of aptitudes also will reflect influences that were external to the environment of the course, as noted earlier. All of the differences between an aptitude profile at the beginning and the end of a course cannot be attributed to participation in the course.

Whatever the sources of influence on the aptitude profile, the "new" aptitude profile then becomes the source of predictor variables for on-job performance, as shown at the top of Figure 1.

Variables in the Evaluation Model

The variables involved in evaluating the effects of instructional content and organization on trainee performance may be categorized in various ways. Table 1 shows a classification suitable for evaluating technical training in a computer-managed instructional setting.

Table 1

VARIABLES FOR PREDICTION OF TRAINEE PERFORMANCE
IN A LESSON OR SEGMENT OF INSTRUCTION

Variables	Remarks
Personal Descriptors	
• Preassessment Battery Scores	Variables in this class are constant or fixed for each trainee throughout the duration of a course.
• Other descriptors	
- Sex	
- Age (date of birth)	
- Service branch	
- Length of service	
- Air Force Specialty Code	
- Prior duty assessment	
- Etc.	

Table 1 (concluded)

Variables	Remarks
Treatment Variables	
• Shift assignment • Instructor assignment • Instructional group size • Date of first enrollment • Instructional materials and procedures - Review materials - Practice materials - Self-check test items - Individual coaching - Etc.	These variables are sometimes called "process variables." Some variables in this class will be constant throughout the duration of a course. Other variables may change with each lesson. Comparing trainee performance under different instructional arrangements or "treatments" is the essence of evaluation of instruction.
Achievement Variables	
• Achievement in prior lessons - Measured time to criterion (LTM) - Number of attempts to criterion (NATT)	Achievement indicators from lessons preceding the one being evaluated are predictor variables.
• Achievement in current lesson First attempt measured time - (MTM) - First attempt score (LSC) - Measured time to criterion (LTM) - Number of attempts to criterion (NATT)	MTM and LSC scores are "within lesson" predictors of lesson achievement as indexed by LTM and NATT scores. LTM and NATT may be used as outcome measures ("dependent variables") singly or in combination; important to control for LSC and MTM because MTM and LTM are not independent and LSC predicts both LTM and NATT.

A Hypothetical Example of Instructional Evaluation

Consider an evaluation to assess the effects of certain revisions in a lesson. For example, assume that review of trainee performance suggested that performance might be improved in a segment of instruction if several practice exercises with self-check test items were provided. Suppose, further, that two groups of trainees were given opposing guidance to influence the effort spent by trainees on the exercises: one random half of trainees was strongly urged to attempt all the practice exercises and told not to attempt the criterion test before succeeding on all exercises, and the other random half of trainees was mildly encouraged to go through the practice exercises but also urged to attempt the criterion test as soon as possible.

Assume, for the illustration, that the guidance given trainees did affect the amount of attention given to practice exercises. Strong urging to attend to the practice exercises led most trainees so urged to work the practice exercises whereas only mild encouragement led most of the remaining trainees to skip through the practice exercises.

Suppose the results of the experiment were portrayed as in Figure 2. Here, the two groups (exercises "urged" or "not urged") are further divided into high skill and low skill, where "skill" might be an aptitude measure such as reading comprehension or numerical reasoning. The hypothetical results suggest that the exercises did not have much effect on the high-skill group, as the average number of attempts was about the same for both conditions. For the low-skill group, working the exercises seemed to lower the average number of attempts.

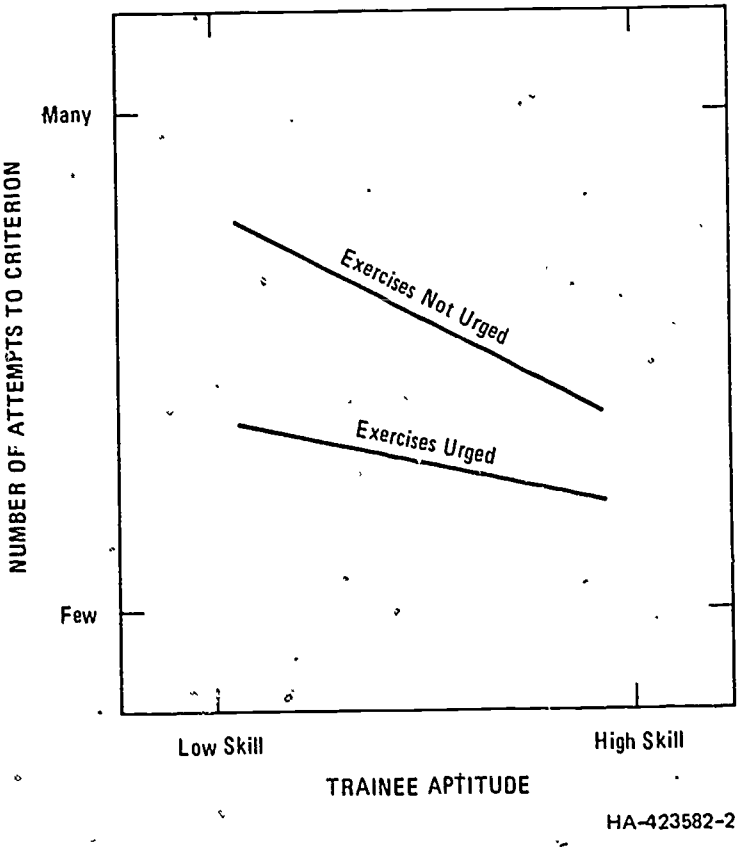

FIGURE 2 EFFECTS OF COMPLETING PRACTICE EXERCISES ON NUMBER OF ATTEMPTS TO CRITERION ACCORDING TO TRAINEE APTITUDE

Results such as those in Figure 2 would provide evidence for an aptitude-by-treatment interaction (ATI) since the effects of the treatment depend on aptitude. Other, more complicated, outcomes are possible depending on how many variables are considered in the experiment. Findings such as these could lead to a revised instructional strategy:

1. More practice exercises would be added to the lesson.

2. All trainees with lower-than-average scores on basic skill measures in the Preassessment Battery would be strongly urged (perhaps required) to complete all the practice exercises before attempting the criterion test.

Which Should Come First -- Good Measures or Good Experiments?

The above heading does not pose a real choice. Although it is possible to have good measures and poor experiments, it is not possible to have good experiments without good measures.

The preceding paragraphs described the results of a hypothetical experiment on the effects of practice exercises. The findings from that experiment were fairly unambiguous and suggested useful implications for changing instructional strategy. However, the findings from the hypothetical experiment assumed the following:

1. The contents of the practice exercises and the criterion test were relevant to the instructional content.

2. The measures of performance were dependable.

3. The measures obtained in the instruction were related to performance in a job assignnent.

Experiments that modify training approaches to achieve a better fit between instruction and trainee aptitude are unlikely to lead to trustworthy conclusions if the data from the experiments are not also trustworthy.

Figure 3 presents a sequence of evaluation questions and decision options for an assessment of curriculun content, means for measuring training and on-job performance, and the relationship between training performance and on-job performance. The questions and decision paths shown in Figure 3 portray a diagnostic evaluation that should precede efforts to improve the payoff from instruction through adaptations of instructional treatments to learner aptitudes.

The arguments for performing the diagnostic steps in Figure 3 before undertaking experiments with alternative instructional treatments are discussed in the following paragraphs. These arguments are consistent with the position maintained throughout the Interservice

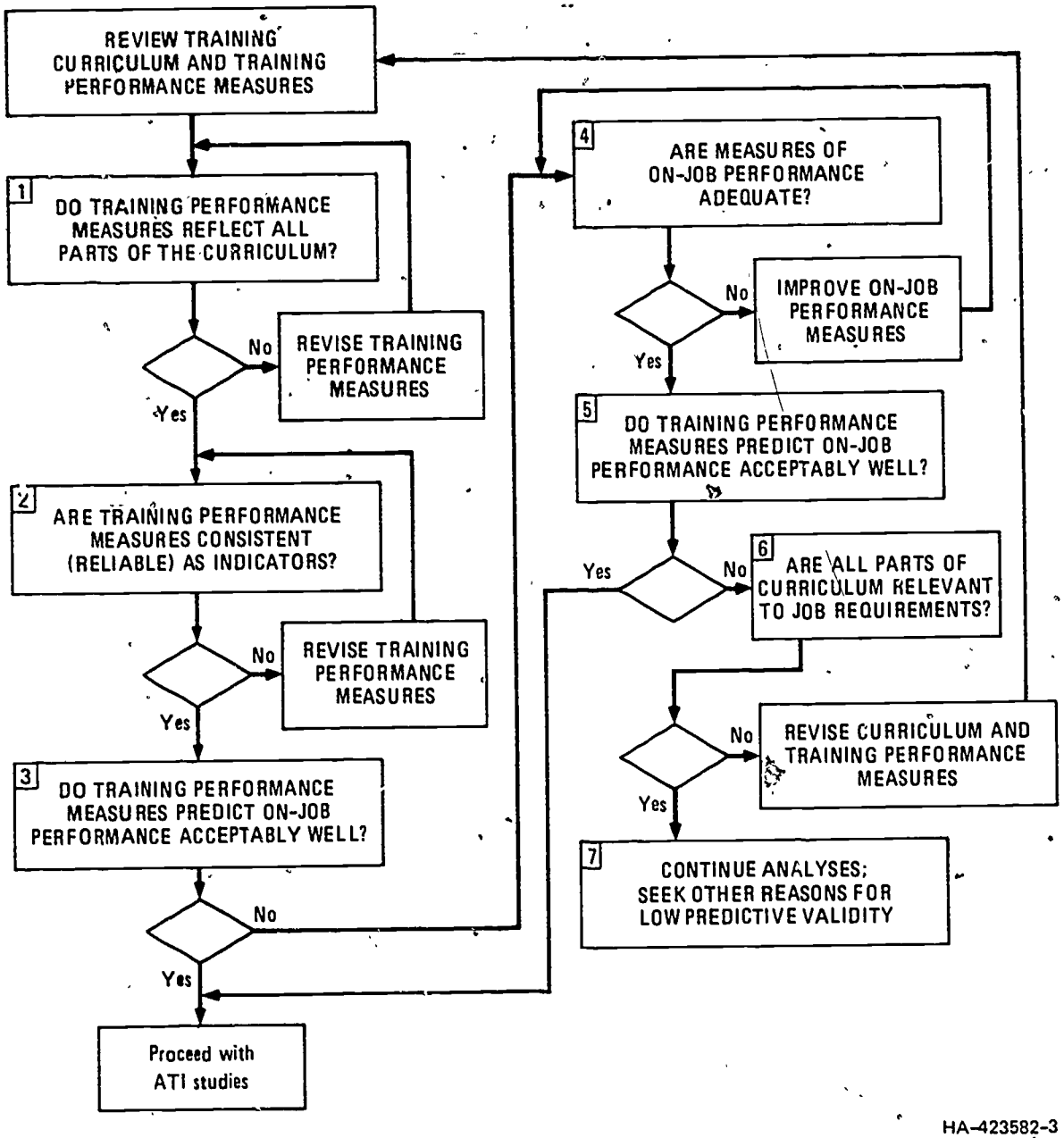

FIGURE 3 STEPS IN EVALUATING CURRICULUM CONTENT AND ADEQUACY OF PERFORMANCE MEASURES

Procedures for Instructional Systems Development (AFM 50-2). The foundation of the ISD model is the dependence of instructional content and procedures on the function and task requirements of jobs. Job relevance is a primary criterion for evaluating instruction.

This report supports the fundamental position of the ISD. This report also emphasizes that effective diagnostic evaluation of existing instructional programs depends on relevant, reliable, and comprehensive measurement of these programs. As every navigator knows, plotting a course means knowing the present position as well as the intended destination. The following discussion of the procedures illustrated in Figure 3 concern the importance of appraising the present program before proceeding in new directions.

- The essential purpose of training is to improve on-job performance. Specific instruction may be narrow (deal with only one or a few tasks or functions required by a job) or broad (address all tasks and functions that define a job). Whatever the coverage, the content of instruction is directed toward qualities that a successful job performer must possess. These qualities may be specific or generalized knowledge, a variety of skills, or attitudes that influence behavior on the job. Regardless of focus, the content of instruction is based on analyses of functional requirements of the job. The ultimate proof of instructional effectiveness in training is improved performance in the job. The sequence shown in Figure 3 assumes that the curriculum and plan of instruction grew from an analysis of job requirements.

- Every instructional objective important enough to be stated implies an associated process for reaching that objective and means for measuring the degree to which the objective has been reached. This assertion is meant to emphasize two complementary points:

  1. Instructional objectives and instructional processes should reinforce one another. For every explicit objective, there should be an identifiable process for achieving it. Furthermore, every instructional process or activity that consumes staff or trainee energy should be justified by an identifiable objective.

  2. Objectives imply measurement. Without measurement relevant to an objective, there is no dependable way to estimate the degree to which the objective has been reached.

- Measurements of student performance during a course of instruction must be acceptably reliable (consistent, dependable, accurate) and fully representative of the content of instruction (i.e., possess curricular or content validity). A central purpose of this report is to suggest ways to devise instructional treatments suited to learner characteristics. This is

analogous to a medical prescription based on current symptoms and other characteristics of the patient. Just as a responsible medical prescription requires dependable measures of the state of the patient relative to a desired healthy state, so too does an instructional prescription require dependable measures of the learner's state relative to the desired one. Training achievement measures are the analogs of medical measures of healthfulness. Evaluation demands the best possible measures of status so that changes in status can be assessed accurately.

The sequence of evaluation questions and decision options shown in Figure 3 concerns the content of training, the appropriateness and dependability of training performance measures, and the relationship of training performance to performance on the job following training.

Question 1

Question 1 in Figure 3 concerns the content validity of measures of student achievement. The essential issue is whether or not the performance measures provide a fair sample of the content of instruction. If material is being taught but its mastery is not being tested in any way, then either the tests should be expanded to cover the material or the appropriateness of the material should be reconsidered.

Question 2

Question 2 concerns the reliability of measures of achievement. Two causes of test unreliability may be detected. One cause is that the items making up a test are not homogeneous; that is, they do not all measure some aspect of the same attribute. The second cause is that a test is not long enough.

Summing scores from nonhomogeneous items to a total test score can lead to confusion rather than to clarity. The meaning of test scores must be clear if achievement tests are to be a trustworthy source of information for decisions about instructional treatments adapted to trainee aptitudes.

Consider, for example, a brief test composed of five items in which Items A and B are homogeneous with one another, and Items C, D, and E are homogeneous with one another but are not homogeneous with with Items A and B. Imagine, further, that criterion performance has been defined simply as "passing three or more items." With five items, there are 16 different response patterns that will yield a total score of three or more. At one extreme, a trainee could fail both items A and B but pass items C, D, and E to meet the minimum criterion. At the other extreme, a trainee could pass both items A and B but fail any two of items C, D, and E and also meet the criterion as defined. If criterion tests are composed of subsets of items that are not homogeneous, then criterion performance should be defined by minimum performance on

each subset. In the above example, a better specification of criterion performence would be to "pass either A or B or both and pass any two or more of C, D, and E."

Another factor that affects test reliability is the number of items in the test. Generally speaking, the longer the test the more reliable it will be. Of course, in an operational setting such as Air Force training, tests cannot always be long enough to obtain excellent reliability; some balance must be maintained between testing time and instructional time.

There is no single answer to the question, How long is long enough? Any performance test is a sample of items from a much larger possible set of items to measure a category or "domain" of behavior. Small samples mean large sampling error. If decisions based on test performance are important, then the sample of items that make up the test should be large enough to provide some redundancy in measurement as a means for reducing sampling error. If decisions based on test performance are not of major importance, then a few items may suffice. Decisions about test length call for assessing the consequences of a wrong decision based on test performance.

Question 3

The third question in Figure 3 concerns the predictive validity of training; that is, the extent to which success in training is associated with satisfactory performance on the job.

The relationships between training performance and job performance are imperfect for a variety of reasons -- the complexity of human behavior, the low reliability of many measures of behavior, failure to measure certain influences on training performance or job performance, and so on. In practice, it is rare to find correlations between training performance and job performance that exceed .35 or so, thus implying that only about 10%-20% of variance in measures of job performance can be attributed to measures of performance in training.

Even discounting for the problem of measurement and the complexity of the relationship, the ultimate justification for training is to improve performance on the job. At the very least, the relationship between a measure of training performance and one of performance on the job should be positive and greater than zero -- training performance should predict job performance better than chance.

Question 4

The fourth question deals with the adequacy of measures of performance on the job. As is shown elsewhere in this report, the upper limit of a measure of association between two measures is defined by the least reliable measure. As a rule, it is easier to achieve reliability in measures of training performance than in measures of performance on the job. When measures of association between training and

on-job performance are low and it has been established that the training measures are as reliable as practical considerations warrant, the first place to search for improvement is in the measures of on-job performance. This question directs attention to that source.

Improving measures of job performance may go beyond the recognized responsibilities of a course designer or manager of instruction, but people responsible for training design and managenent clearly have interest in the problem. Measures of performance on the job, following training, are essential to functions designated in the ISD model as "external evaluation." Clearly, training personnel, course designers, and managers of instruction must help address the issue. Relationships between external and internal evaluation, as the functions are defined in the ISD model, are discussed in more detail in Section III of this report.

Question 5

Question 5 in Figure 3 is Question 3 asked again following efforts to improve measures of on-job performance. If inability to predict on-job performance acceptably well was a function of inadequacies in the measures of on-job performance, then improvements in the on-job measures may be required before proceeding with ATI studies to increase the effectiveness of training. This possibility is denoted by the "yes" path from Question 5 to the box, "proceed with ATI studies."

It also is possible that inadequacies in the training curriculum will be revealed only after the issue of the predictive validity of training performance has been pursued thoroughly. This possibility is identified by the "no" path leading from Question 5.

Question 6

As noted previously, it is possible that inconsistencies between a training curriculum and the requirements of a job will become evident only after completion of a serious effort to improve measures and affirm relationships between them. Question 6 deals with the match between a training curriculum and the requirements of a job. The question directs the curriculum designers to reconsider the curriculum content in light of revised analyses of job requirements. If curriculum revision is necessary, then measures of training performance will need to be revised also.

The ISD model (AFM 50-2) provides methods for job and task analysis and for translating findings from such analyses into specifications for instruction.

Box 7

Box 7 in Figure 3 is the "court of last resort" -- further research is necessary. If the curriculum fits the job, the training performance measures fit the curriculum, the measures of job

performance are as satisfactory as one can make them, but training performance still does not predict job performance, then other reasons for lack of relationship must be sought. If this decision is reached, then the most likely problem is that some unmeasured environmental variables affecting job performance are operating. Examples could include the workplace layout, mode and quality of supervision, fluctuations in work demands, and so on.

Literally, Box 7 invites the reciprocal of Question 6 -- are all job requirements represented in the curriculum? If some important aspects of the job are not represented in the curriculum, then training performance cannot be expected to predict job performance. However, if important aspects of the job are not represented in the curriculum being evaluated, then either that curriculum should be revised to accommodate additional job requirements or supplementary curricula should be developed.

Planned Experiments with Alternative Instructional Treatments

Systematic instructional research and evaluation as part of an operating training program implies a recurring cycle of planned trials with alternative instructional approaches. Figure 4 illustrates a generalized sequence of steps in such a program of research and evaluation. The focus is on relationships among classifications of trainees and of instructional treatments. The search is a continuing one for dependable generalizations about the effectiveness of instructional approaches for trainees characterized by certain patterns of aptitudes and prior performance.

Underlying the formulation shown in Figure 4 is the expectation that dependable generalizations are more likely from repeated sequences of experiments directed toward questions of limited scope than from complex experiments directed toward broad questions. This is not meant to discourage efforts to strive for crucial experiments but to recognize that modest findings are more likely than dramatic ones even with the most carefully planned instructional experiment.

The Simple Inter-Group Comparison

In the classic experiment, the focus is on the effects that independent variables have on dependent variables. The independent variables that the experimenter can manipulate are the experimental variables. The formal research proposition is in "if-then" terms -- "if X under such-and-such conditions, then Y will be observed." In such a formulation, "X" defines the experimental variable, "such-and-such conditions" define the circumstances of the experiment (including the characteristics of the subjects of the experiment), and "Y" defines the measure of outcome or the dependent variable. The essence of experimentation is control over the variables involved and the conditions under which the variables will be observed.

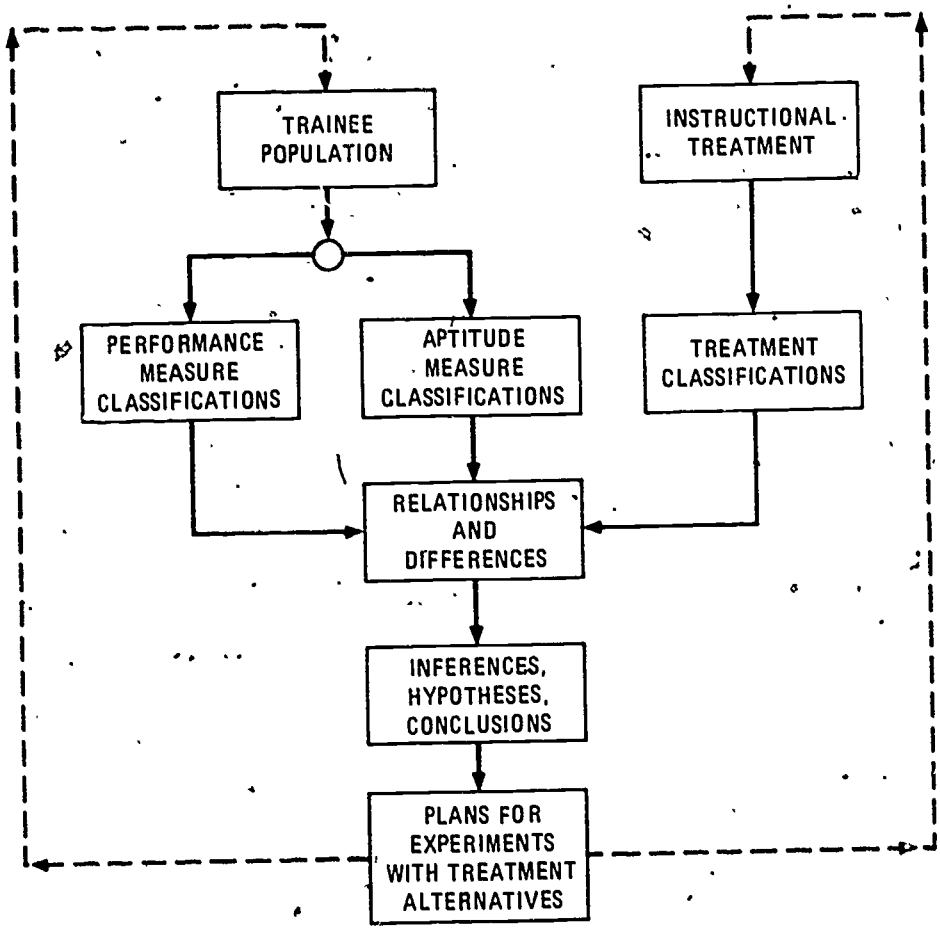

FIGURE 4 GENERALIZED SEQUENCE OF STEPS IN ANALYZING TRAINEE PERFORMANCE AS A FUNCTION OF TRAINEE APTITUDES AND INSTRUCTIONAL TREATMENTS

Consider first a straightforward experiment in which the purpose is to assess the effectiveness of a "new" approach for teaching some segment of a training curriculum. This could be done within the context of normal instruction by assigning one randomly selected group of current trainees to the "new" method while the other randomly selected group of trainees experienced the "old" or current method. Outcome measures of interest (e.g., measured time to criterion) would be obtained routinely and the two distributions of scores -- those from the "new method" group and those from the "old method" group -- could be compared to see if the difference between the two methods was large enough to be attributed to the method of instruction rather than to chance.

Even this straightforward experiment is not quite so simple as the description makes it appear. For example, the number of trainees at the appropriate stage of instruction at any one time might not be large enough to provide two samples of sufficient size to allow an adequate test of the two methods. In this case, samples of adequate size could be accumulated over several successive classes; literally, a small-sample experiment would be replicated several times, thus posing the additional problem of whether data should be pooled over cohorts for a single analysis or whether probabilities should be combined over analyses of several replications.

Another complication might arise if instruction were organized for administrative purposes into shifts or periods defined by time of day. To adjust to this, one could make random assignment of trainees to methods within each shift so that method effects would not become entangled with shift effects. But splitting each shift might make it difficult to insulate trainees experiencing one method from those experiencing the other; a systematically balanced schedule might be worked out so that each shift experienced each method over several classes of trainees. Again, some problems of analysis could arise, especially if the characteristics of trainees were to vary markedly from one incoming class to another.

For convenience of illustration, assume that complications in implementing the alternate methods experiment are worked out and two rival distributions of outcome scores are generated under conditions that are as close to identical as can be managed. Comparing the two score distributions, to decide whether one method was superior, still addresses only the question of which method is best on the average.

Many differences among trainees, both within and between groups, may be substantial -- prior experiences, performance on earlier segments of instruction, skill in reading, proficiency with tools, and so on. Randomization of assignment to treatment protects against bias by making the chances of "unusual" performance equally likely in either group. However, the simple score distribution comparison cannot address important questions about how individual differences among trainees are related to the experimental treatments and influence the outcome measures obtained. Even so, the simple score distribution

comparison between subjects from the same trainee population assigned at random to rival approaches is a legitimate approach when the research or evaluation question is no more than "which treatment is best on the average.". Comparisons need not be restricted to two groups; there may be as many comparison groups as the credible alternatives, available subjects, and the logistics of experimentation will permit.

Analysis of Variance and Multiple Regression Analysis Models for Instructional Treatment Experiments

Designs for experiments cannot be discussed without also considering some statistical analysis issues that are related closely to experimental design decisions. This is particularly the case when an objective of instructional research is to find and establish dependable generalizations about interactions between the characteristics of trainees and the characteristics of instructional treatments.

Data from instructional treatment experiments designed to "pick a winner" -- that is, experiments designed to find the treatment whose average effect is greatest among rival treatments -- are often subjected to the statistical technique called analysis of variance (ANOVA). When the independent variables that define the conditions of the experiment are categorical and also are functionally independent of one another (i.e., are not correlated), then ANOVA may be the most appropriate technique for the statistical analysis.

Briefly, ANOVA provides a means for testing whether the mean differences on the dependent variable measure between two or more independent groups are sufficiently large to be considered nonchance or "statistically significant." This test of statistical significance makes use of ratios of variances, hence the name "analysis of variance." (The variance is an index of spread or dispersion of scores around the mean or arithmetic average of a distribution of scores.) For example, in an experiment comparing three methods of instruction to one another, ANOVA may be used to provide an overall test of the differences among the three means on the criterion test. The key statistic, or F-ratio, is the ratio of the "between groups" variance to the "within groups" variance. When the F-ratio exceeds an expected value by a sufficient amount, the conclusion is that the groups are different. Literally, the method estimates the probability that a predictor variable (such as the method of instruction) could yield results different from simple random selection.

The analysis of variance technique, when extended to two-way classifications or higher, permits identification of interactions between and among variables. To illustrate, assume that a simple Treatment A vs. Treatment B experiment was set up so that participating trainees could be differentiated on more than the single dimension of "member of the A (or B) group." For example, suppose that the training course is one to which persons from two different Air Force Specialty Code (AFSC) backgrounds are assigned. Assume that each AFSC

indexes a qualitatively different experience background so that developing a training curriculum suitable for both simultaneously has led to segments of the curriculum being better suited to one background than to the other. The question that justifies an experiment, then, is whether a revised training segment (Treatment B) is better suited than the current material (Treatment A) for those who have had difficulty with Treament A.

Table 2 shows a two-way experimental design in which treatment (A or B) is crossed with a trainee characteristic (AFSC 1 or AFSC 2). Note that there are now nine score distributions that, collectively, describe the results of the experiment -- four sets of AFSC within Treatment scores, two sets of Treatment scores over AFSC, two sets of AFSC scores over Treatment, and one overall or grand distribution of scores.

The notation used in Table 2 is conventional for denoting means of groups. Implied by the notation is a distribution of several scores in each cell that can be summed in raw form to yield row totals, column totals, and an overall total. It also is implied that the four cells in the body of the table denote independent proportions of the total sample (i.e., $N_{1A} + N_{1B} + N_{2A} + N_{2B} = N_T$). Two-way and higher-order ANOVAs are computed more easily and interpreted more readily if subsamples are equal in size to one another. In fact, two-way (and higher-order) classifications for ANOVA become quite untidy when sub-samples are not equal or, at worst, not proportional. Thus, in the example illustrated by Table 2, equal-sized sub-samples are assumed.[2]

Figure 5 displays a graph of imaginary results that would be highly favorable to resolving the problem that stimulated the experiment; that is, to develop a revised training segment that is better suited to those who have had difficulty with the present material. Figure 5 shows an interaction between trainee "aptitude" (represented by prior experience and training underlying the two AFSC categories) and "treatment." The symbols used in Figure 5 correspond to the notation for cell, row, and

---

[2] The actual composition of a trainee cohort "population" is unlikely to provide the convenience of equal numbers on some desired classification variable, such as AFSC in the Table 2 example. Random samples of equal size can be drawn to match the size of the smallest cross-classification in the cohort population or any acceptable minimum size less than the smallest sub-set. If it is more convenient administratively to draw the samples for analysis after the experimental data have been collected, equal sample sizes for analysis can be created by sampling the data set. If the latter procedure is followed, one must be sure that "group size" is irrelevant to the experimental treatment. If treatment involves some considerations of group size, then groups should be constructed in advance of the experiment so that group size is an explicit factor in the experimental design.

Table 2.

SCHEMATIC OF TWO-WAY CLASSIFICATION EXPERIMENTAL DESIGN DENOTING GROUP MEANS

Trainee AFSC	Treatment A	Treatment B	Total
1	$\bar{X}_{1A}$	$\bar{X}_{1B}$	$\bar{X}_{1.}$
2	$\bar{X}_{2A}$	$\bar{X}_{2B}$	$\bar{X}_{2.}$
Total	$\bar{X}_{.A}$	$\bar{X}_{.B}$	$\bar{X}_{T}$

column means shown in Table 2. A reasonable policy decision, given the evidence summarized in Figure 5, would be to use Treatment A for trainees with AFSC 1 background and use Treatment B for trainees with AFSC 2 background or, if that were not feasible, to replace Treatment A with Treatment B for all trainees.

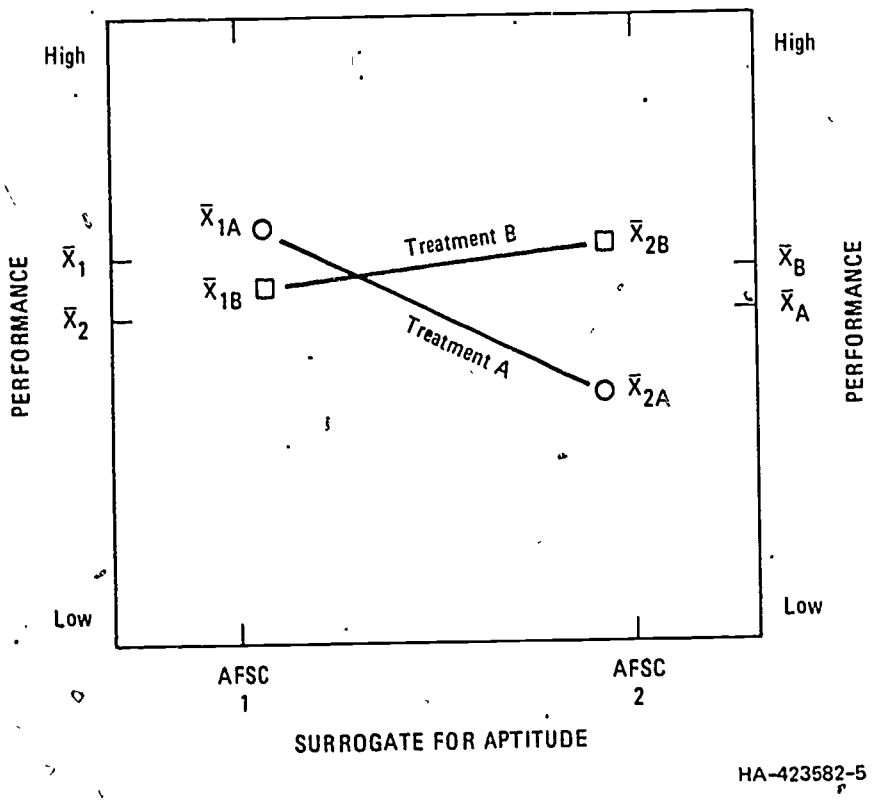

FIGURE 5 IMAGINARY RESULTS OF ALTERNATIVE TREATMENT EXPERIMENT USING AFSC AS A SURROGATE FOR TRAINEE APTITUDE

Figure 5 illustrates that effects for interaction between treatment and aptitude may be present and highly visible (i.e., significant in a statistical sense) with the average difference between treatments being of no consequence. More specifically, the performance difference between Treatment A and Treatment B, averaged over all subjects, is small; the graph is intended to imply that the average difference between treatments is too small to matter. That "no difference" result is the only one that would have been detected in a one-way ANOVA. When the aptitude factor (represented by AFSC) is introduced in a two-way ANOVA, however, the important aptitude-by-treatment interaction becomes evident.

The ANOVA is well-suited to experiments in which the experimenter controls the independent variables and when the independent variables are functionally independent categories. The ANOVA ceases to be the best technique, and may be completely inappropriate, as one or more of the following conditions arise:

1. When independent variables are functionally related to one another.

2. When the independent variable (in a one-way design) is continuous rather than categorical or when the independent variables (in a two-way or higher order design) are a mixture of continuous and categorical variables.

3. When cell frequencies are unequal and also disproportionate.

4. When four or more independent variables are used in the classification of treatment and subjects.

As a general rule, multiple regression analysis is better suited than is ANOVA to data from instructional treatment experiments. This follows primarily from the lack of functional independence among individual differences variables used to represent trainee aptitudes. Personal characteristics (personality traits, abilities, skills, educational levels, etc.) typically are not independent of one another. There are reasons beyond independence among variables that generally favor regression analysis over ANOVA. Figure 6 provides a rough guide for use in choosing between ANOVA and regression analysis when planning an instructional treatment experiment.[3]

An excerpt from McNemar's no-nonsense discussion of analyses involving classification or predictor variables that are not independent of one another provides an appropriate summary of the problem of choosing the appropriate statistical technique.

---

[3] Both ANOVA and regression analysis belong to the class that statisticians call "the general linear model." ANOVA can be shown mathematically to be subsumed under regression analysis.

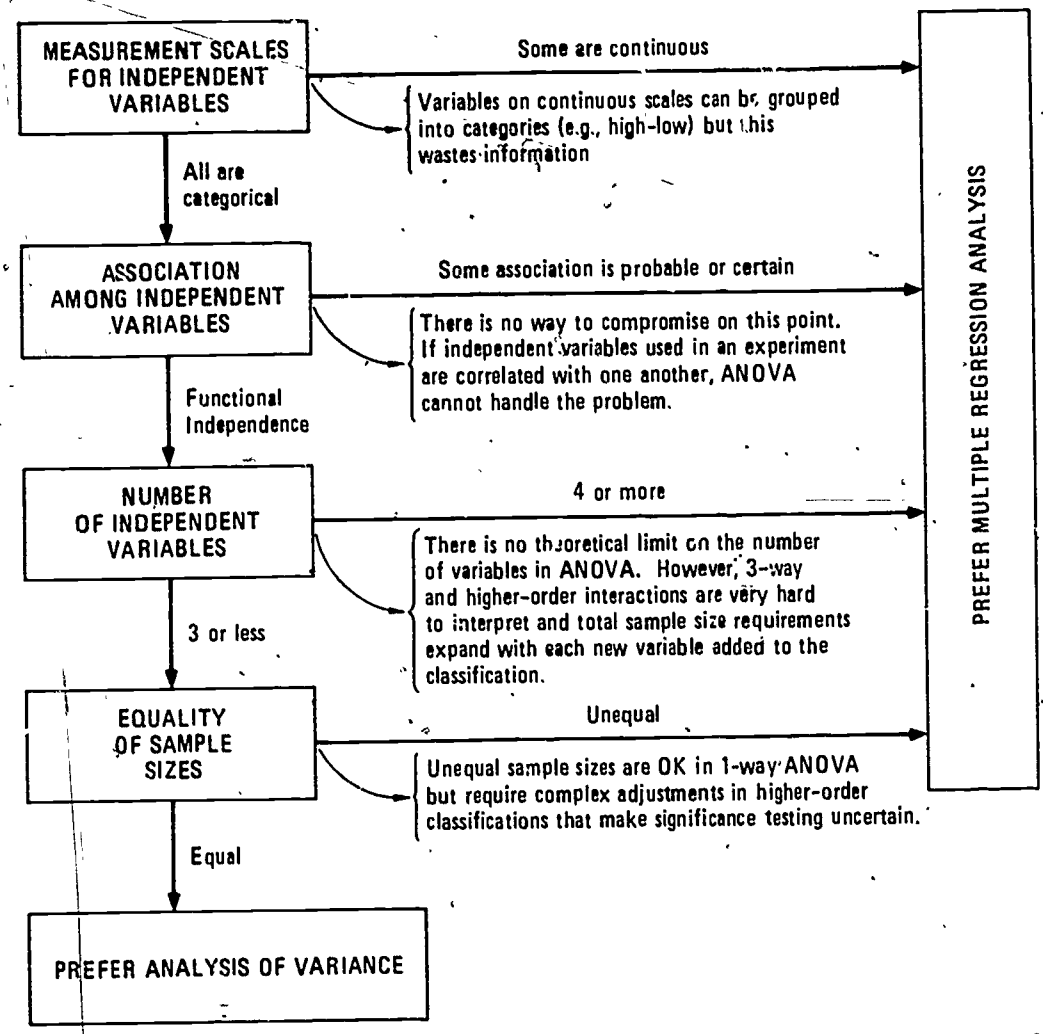

FIGURE 6 DECISION GUIDE FOR CHOOSING BETWEEN ANALYSIS OF VARIANCE AND MULTIPLE REGRESSION ANALYSIS

"... the factorial design approach [the conventional ANOVA arrangement of variables by levels] is inferior to the multiple regression technique as a method for testing the statistical significance of factors that are characteristics of individuals. The analysis of variance of the data obtained by factorial experiments provides tests as to whether factors have produced variation. Multiple regression, in contrast, has traditionally been associated with analyzing natural (not laboratory produced) variation into sources with no requirement that the sources be uncorrelated with one another" (McNemar, 1969, p. 453).

Some Suggestions for Do-It-Yourself Regression Analyses

Multiple regression analysis is a statistical method for analyzing the collective and unique contributions of two or more independent variables, X_1, X_2, . . . X_k, to the variation of a dependent variable, Y. The method is oblivious to the analyst's motives -- it can be used in exploratory "data snooping" when one is trying to get a better idea of what goes with what and it can be used to help test carefully formulated if-then propositions about what one expects to observe under particular conditions.

Appendix C contains an overview of regression analysis, the coding of categorical variables (such as instructional treatments) for use in regression analysis, and the creation of variables to represent aptitude-treatment interactions in regression analysis.

The suggestions that follow are intended to show how multiple regression analysis may be used for either exploratory or explanatory purposes. Section III of this report is directed toward problems of designing and evaluating alternative instructional treatments. Therefore, in this discussion, independent variables (sometimes called predictors) and a dependent variable (sometimes called the criterion) will be used but operational meanings will not be given to each X and Y.

Suggestion 1: Start with a problem whose solution you think you could interpret.

This suggestion is another way of saying that the variables used in the analysis should be ones that make conceptual sense to you. Whatever statistical findings are obtained, sooner or later those findings must be interpreted in words to someone who has less understanding of the analysis than you do.

Suggestion 2: Look first at the pieces before trying to put them together.

The raw ingredients are sets of scores or values that describe people. The raw data matrix has N rows (one row for each person or case) and k columns (one column for each variable). Each variable or column in the data matrix can be described by the number of points (N),

the range of values, its central value (mean), its dispersion (standard deviation), and its shape (skewed or normal, flat or peaked, unimodal or multimodal).

Each pair of columns or variables can be correlated over N cases; thus, each pair of variables also can be described by the number of paired points (N) and the distributional characteristics of the scatter of points on a plane defined by perpendicular axes. This scatter also can be described by the relationship or correlation between the two sets of paired values.

The correlation coefficients (variously called zero-order correlations, simple correlations, and bivariate correlations) are the basic ingredients of multiple regression analysis. If there are k variables, $(k^2 - k)/2$ different correlation coefficients are computed to describe all the two-way relationships represented in the N x k matrix of raw data.

Looking first at each of the two-way scattergrams resulting from the candidate variables will help answer some questions that are important.

1. Are the two-way relationships generally linear? If some relationships appear curvilinear, could they be made more nearly linear by some transformation in the scale of one of the variables?

   Some apparent nonlinearity in two-way scatters may be due to the influence of a third variable; interaction terms added to the regression equation may be helpful. Other curvilinear relationships may be dealt with by transforming a scale through the use of logarithms or exponents. The SPSS manual (see "Special Topics in General Linear Models" in Reference 13) illustrates common transformations that may prove helpful.

2. Do the individual data points in the various distributions appear reasonable? In particular, look for extreme values (outliers) that deviate markedly from the pattern. If these are coding, scoring, or tabulation errors, correct them or drop the case from the data set.

   Extreme values, whether high or low, exert unusual leverage on measures of variation and association. Deviant cases that are not due to scoring or tabulation error are troublesome enough without creating added problems by retaining erroneous data.

   A handy rule of thumb for samples of 50 or more cases is that the standard deviation of the distribution usually will be about 20% of the range of scores (i.e., the difference between highest and lowest scores). If standard deviations substantially different from that are obtained, a closer look is required at the scores in the distribution.

Any data set may contain errors that are never spotted because they fall within the expected range. A basic assumption in all statistical analyses is that such errors are random, rather than systematic, and that they are unrelated.

3. The multiple regression equations that do the best job of accounting for variation in the dependent variable are derived from predictor variables that are not highly correlated with one another but show reasonable correlation with the dependent variable.

Recall that the logic of regression anslysis assumes additivity. One wants combinations of predictors that will add something new, rather than redundancy, to the explanation. If two or more predictors are correlated strongly with one another and also correlatad reasonably strongly with the dependent variable, consider either dropping one of the predictors or combining the predictors into a composite variable. Factor scores often prove helpful in reducing a set of interrelated predictors to a lesser number. This procedure is not without hazard, however, for the derived factor score is an abstraction that is not always easy to interpret.

It is essential to do something about predictor variables that are correlated very highly with one another. Appropriate ways of dealing with what statisticians call the "multicollinearity problem" go beyond a simple visual scan of bivariate (two-way, zero-order) relationships between pairs of independent variables. However, forewarning of likely multicollinearity problems comes from discovering several zero-order correlations of .80 or more (whether positive or negative in sign).

One of the problems of multicollinearity is that the regression coefficients will be unstable from sample to sample. In an extreme case of multicollinearity, the regression solution may be indeterminate. The rule of thumb -- get rid of the redundant measures by dropping them or folding them into constructed composite variables.

Suggestion 3: Expect to go through several trials of "cut-and-try."

One way to state the goal of multiple regression analysis, as noted in Appendix C, is to minimize errors of estimate. This implies that the best solution will be the one that most nearly satisfies the following assumptions:

1. The model has no errors of specification:

a. Relationships are linear (or are made so through scale transformations),

b. All relevant variables are included,

c. All irrelevant variables are excluded.

2. No errors of measurement exist.

3. Errors of estimate are centered at zero, are approximately normally distributed, have similar variance throughout the ranges of X-values, and are unrelated to the independent variables.

Without denying the importance of assumptions regarding the errors of estimate, the most important of these three assunptions are those regarding specification and measurement error. The cut-and-try referred to at the outset involves trying out different combinations of predictors to assure inclusion of relevant ones and exclusion of irrelevant ones. At each step in such trials, a new prediction model is being tested. It is not unusual to go through a half-dozen or so trials before a couple of models settle out as offering essentially equivalent total R^2 values. These two or so "best" models will differ somewhat in the regression coefficients associated with variables common to each model due to differences among the models in the total combination of predictors. Selection of a favored solution, given essentially identical error or residual for each, becomes a matter of judgment and preference. Usually the preferred model will be the one that requires the fewest number of predictors or that can be interpreted most simply or both.

The assumption about errors of measurement can and should be taken seriously. For example, if fallible tests provide data for some of the variables, the efforts directed toward improving the reliability of the tests can be worthwhile as a means of reducing measurement error. Such effort, however, implies a cycle of inquiry rather than mere cut-and-try for the best combination of predictors because a change in a test will also result in a change in the basic data set. Refining measures as a means for improving prediction through regression analysis implies replication over new samples.

Suggestion 4: Consult a statistical analyst whom you trust.

Perhaps this should have been the first suggestion, but that would have inplied even more emphasis on questions of procedure rather than of substance. Every analyst, no matter how experienced, will find occasions when other opinions are helpful. For an analyst who is not particularly experienced with moderately advanced statistical methods or who is unfamiliar with statistical packages that simplify computerized procedures, consultation may be essential rather than merely helpful.

Many variations in the planning and conduct of statistical analyses of multivariate data have not been mentioned in this report (see especially Appendix C). Some may be useful and particularly fitting to the evaluation research questions you wish to tackle. For example, the stepwise procedures in multiple regression analysis can be helpful if one is seeking to specify an efficient prediction equation from a pool of candidate predictor variables. If the order in which predictor variables are entered in a stepwise program is controlled, the several combinations of solutions can be generated to allow the somewhat controversial commonality analysis to be performed (see Mood, 1971; Kerlinger & Pedhazur, 1973; Cooley & Lohnes, 1976). Path analysis, as a method of testing hypothesized causal relationships, has not been mentioned; again, Kerlinger and Pedhazur (1973) or the ubiquitous SPSS manual (Nie, et al., 1975) provide introductions and references to other sources.

Factor scores also have been mentioned as a way to construct composite variables from several related measures. Factor analysis is a sufficiently specialized topic to need a consultant who is familiar with various methods.

Section III of this report concerns more substantive questions of conceiving and evaluating different instructional treatments as alternatives to present ones. The preceding discussion and Appendix C identify some of the methods that can be used to assess the worth of new instructional treatments as alternatives to present ones.

II TEST ITEMS AND TESTS IN CRITERION-REFERENCED MEASUREMENT

Introduction

Guidelines for the development of items and tests for criterion referenced measurement are presented in some detail in the Interservice Procedures for Instructional Systems Development (AFM 50-2), particularly as part of Block II, Phase II of the ISD model. Following ISD procedures will help assure that tests are consistent with learning objectives and training content. The guidelines in this report are intended to supplement guidance in ISD. The following paragraphs describe some practical approaches for assuring the selection of good test items when combining them into tests.

Measurement Assumptions

Measurement of student progress and achievement in any instructional environment is with reference to specified standards or criteria. The assumption underlying instruction is that a student will perform to the criterion specified for the instructional segment, given time, effort, and access to appropriate instructional resources.

Because students differ from one another on such factors as the skills they already possess, their motivation to learn new things, the amount and quality of assistance they seek and obtain, and so on, students also will differ in the time they require to achieve a criterion score. Each student's route to mastery of instructional content may differ from others in the number and kinds of errors they make and the number of trials they require. It is assumed, however, that eventual achievement of a criterion score denotes mastery of the instructional content to which a test applies. Thus, criterion-referenced measurement seeks to specify what a student can do.

An illustration of differences among students is summarized graphically in Figure 7. This graph illustrates the differences in variability among trainees in a self-paced technical training course on two complementary measures of performance: the number of test attempts required to reach criterion and the measured time required to reach criterion. Both number of attempts and time have been scaled to a common artificial scale to simplify comparison.

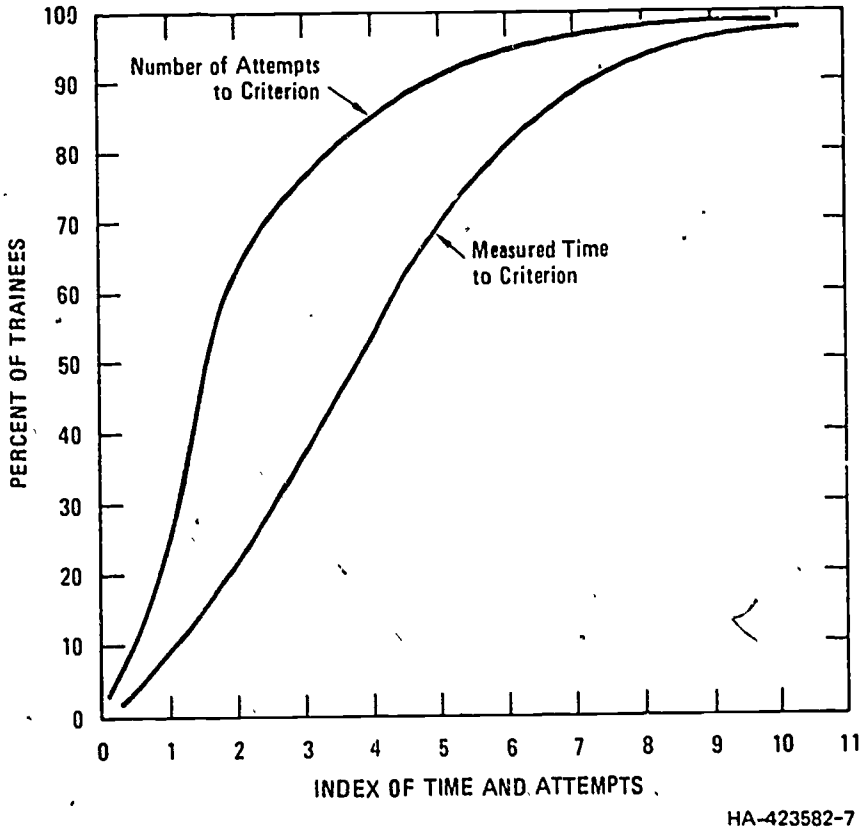

FIGURE 7 VARIABILITY AMONG TRAINEES ON TWO COMPLEMENTARY MEASURES OF PERFORMANCE IN SELF-PACED INSTRUCTION

The graphs are cumulative percentage (or ogive) curves. The curve for "number of attempts" reflects a highly skewed distribution (i.e., most trainees required very few attempts to reach criterion, although a few trainees did require many attempts). By contrast, the curve for "measured time to criterion" is much more nearly normal and symmetrical in shape.

Curves of the same general form as shown in Figure 7 also are characteristic of differences between different tests or, for that matter, between trainees on a common test. For example, if the horizontal axis in Figure 7 were "total test score," a curve like the upper one would illustrate a difficult test -- about 60% of persons attempting it had a score of 2 or less. By contrast, the lower curve would illustrate an easier test -- more than 60% of persons attempting it had a score of 4 or more.

Evaluation and Selection of Test Items

In any measurement effort, it is important that items making up the test be homogeneous; that is, be relevant to the particular instructional content whose mastery the items seek to measure. Assessing the relevance of items and the degree to which a set of items provides adequate coverage of the instructional content to which they apply are largely judgmental decisions for subject matter experts. The process of arriving at such decisions is often referred to as determining the "content validity" of a test.

While judgment must be relied upon to assess content validity, determining the characteristics of items requires data from actual trials of candidate items. It is only by trying items under conditions similar to their intended use that the relative difficulty of items can be determined, as well as the effectiveness with which the items differentiate more able students from less able ones and the consistency with which they supply such information.

Statistical concepts and procedures are an indispensable part of measurement theory and test development. In the sections that follow (and in Appendixes A and B, which include more detail about certain procedures), the use of statistics has been limited to descriptive methods that are generally familiar. Graphic methods have been substituted for equations when possible. For detail regarding more quantitative methods for assessing and describing item and test characteristics, see references 2, 6, 10, and 14.

Evaluating Candidate Test Items Through Trial Use During Instruction

In criterion-referenced measurement under conditions of student self-pacing or essentially unlimited instructional time, it is assumed that all students eventually will perform successfully on all items in the test. Furthermore, the test items are assumed to discriminate

consistently; that is, if a student passes an item at one time, the student will again pass the item if it is administered at a later time. These assumptions are illustrated in Table 3, using "P" to denote pass and "F" to denote fail.

Perfectly consistent patterns, as illustrated in Table 3, are rarely found in practice. Individual differences among students, guessing, ambiguous or otherwise poorly constructed items, and other factors may result in irregular response patterns that deviate from the idealized one. The purpose of item trials is to screen out the poor items (or identify poor items for revision and improvement) so that the eventual test is one that provides both reliable and valid indications of "true" student performance.

Table 3

IDEALIZED PATTERN OF STUDENT PERFORMANCE ON TEST ITEMS

	Time 1				Time 2				Time 3				Time 4				. . . . .
	Item				Item				Item				Item				
Student	a	b	c	d	a	b	c	d	a	b	c	d	a	b	c	d	. . . . .
1	P	P	P	F	P	P	P	P	P	P	P	P	P	P	P	P	
2	P	P	F	F	P	P	P	F	P	P	P	P	P	P	P	P	
3	P	F	F	F	P	P	F	F	P	P	P	F	P	P	P	P	
4	F	F	F	F	P	F	F	F	P	P	F	F	P	P	P	F	
5	F	F	F	F	F	F	F	F	P	F	F	F	P	P	F	F	

. . .

In the illustration shown in Table 3, the time scale denotes points that span the period of instruction: preceding, during, and after. If Time 1 in Table 3 denoted a four-item test given prior to instruction, the results would indicate that Student 1 probably required very little instruction since that student passed three of the four items before receiving any new instruction. By contrast, Students 4 and 5 passed none of the items at Time 1. If Time 4 denoted the same four-item test given upon completion of instruction, the results would show that Students 1, 2, and 3 passed all items, Student 4 passed three of the four items, and Student 5 passed two of the four items. Another inference that could be made from the pattern of successive test administrations is that Item a was the easiest item, Item b was the next easiest item, and so on.

If practical considerations governing the instructional arrangements permit it to be done, a direct approach for assessing and screening items is to administer the trial items to actual students at similarly spaced intervals before, during, and following instruction coincident with development and refinement of test items. Successful performance on test items is assumed to be a function of instruction. Therefore, the expected performance pattern would approximate that shown in Table 3 if the items were good ones.

How many students are needed for such trials and how many trials should be made? Clearly, there must be at least two trials, and three or more trials will yield more dependable evidence. The minimum number of students needed depends partly on the degree of confidence desired in the results and partly on the number of possible response patterns. If approximate estimates are acceptable, a three-administration trial should have at least 16 students, and trials with more than three test administrations should have 30 or more students. If highly accurate estimates of item characteristics are needed, then the number of students needed may be many times that number. As a general rule, however, trials with from 30 to 60 students should yield data that are accurate enough for most practical decisions regarding the quality of items.

Appendix A is an extended discussion of an approach for evaluating candidate test items during the conduct of self-paced instruction. The major part of Appendix A is devoted to a step-by-step example with imaginary data for a three-trial evaluation of candidate items. Following that detailed example, suggestions are given for extending the method to more than three trials.

Evaluating Candidate Test Items Through Trials with Cross-Sectional Samples

Another approach for estimating the characteristics of candidate test items for criterion-referenced measurement is to administer the items to samples of people selected to represent levels of competence in the performance area to which the test items apply. This approach may be an alternative when it is not feasible to administer the candidate items to students at intervals before, during, and following training as described above. The approach also may be used to complement the repeated measures procedure and thereby add to the information available for evaluating and refining items.

In the three-trial approach described in Appendix A, students respond to the full set of candidate test items at approximately equally spaced time intervals as they experience training. With that approach, the functional relationship between training and test item performance is established directly. If the items are well constructed and appropriate to the content of instruction, the mean level of student performance on each test item will increase with each successive administration. If this pattern does not occur for certain items,

then those items are assumed to be flawed in some way and either are revised and tried again or are rejected and replaced with new items for tryout.

The cross-sectional sample approach substitutes known levels of experience, proficiency, or competence for time in training. Item characteristics similar in form to those illustrated earlier should result -- assuming, of course, that the items are well designed and appropriate to the performance domain to which the instruction is directed. These assumptions are illustrated in Figure 8.

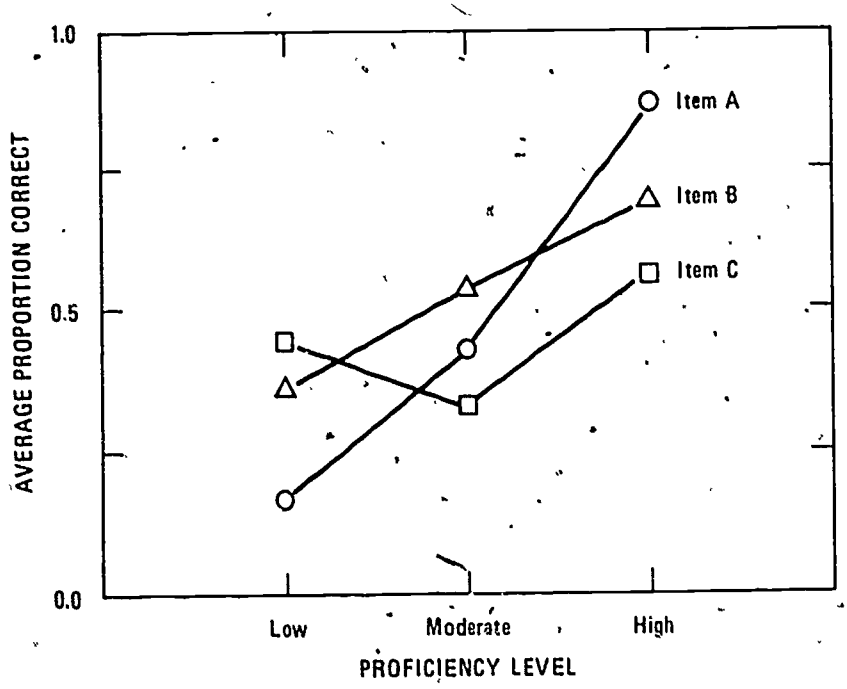

FIGURE 8 TYPICAL ITEM OPERATING CHARACTERISTICS FOR ITEMS ADMINISTERED TO PERSONS OF DIFFERENT LEVELS OF PROFICIENCY

The graph in Figure 8 shows the mean proportion of correct responses to hypothetical items A, B, and C by persons from three broad levels of proficiency in the substantive area to which the items apply. The steep slope for Item A indicates an item answered correctly by very few persons of low proficiency and answered correctly by most persons of high proficiency. By contrast, Item C is one that does not differentiate between persons of low and moderate levels of proficiency and also is not answered correctly by nearly half of the persons with high proficiency. The behavior of Item C signals a problem. Item C may be poorly constructed so that response errors are due to ambiguity, clumsy wording, or other flaws. However, if study of the item does not indicate obvious flaws in wording, instructions, etc., it could be that the instructional content to which the item applies concerns an attribute that is not related to proficiency. If further investigation supported that possibility, then one would reconsider the portion of the curriculum that dealt with the attribute in question. Thus, the screening of test items on samples of persons representing degrees of proficiency in the skills toward which the instruction is directed can help identify curriculum problems as well as provide a way for refining test items.

Sampling Considerations. The sampling objective is to construct a sample of people whose proficiency in the content area of interest spans the range from "not at all proficient" to "very proficient or expert."

1. If the tryout of candidate test items is for a course now being taught but under revision, or for a "new" course that is related closely to already existing courses, then an appropriate proficiency range might be satisfied by selecting (a) persons assigned to the school who have not yet begun instruction, (b) persons at various intermediate stages of instruction, and (c) near-graduates and very recent graduates of the instruction.

2. If the test items are for use in a wholly new course of instruction, identifying people who are very proficient or expert may not be easy. In test development for wholly new courses, one may need to use the same subject matter experts who helped develop the instruction as some of the trial subjects for candidate test items. Finding experts other than those may be difficult. (The problem of finding experts is offset somewhat by the fact that there will be a large pool of people with little or no proficiency and who are prospective students in the new course; for the lower range of the proficiency scale, the problem is simply one of appropriate selection.)

It should be emphasized that the method used for classifying people according to level of proficiency must be based on considerations other than actual test performance. Such descriptors as pay grade, AFSC, or years of experience may be useful in deriving a

working definition of proficiency. Ratings by supervisors of performance on the job also may be helpful. Even self-nomination could be considered. The point is that, even though the classification of proficiency inevitably will be affected by judgment, the bases for classification must be independent of performance on the test items.

A Detailed Example. Appendix B describes an approach for evaluating candidate test items and constructing tests with trials of items over a range of proficiency. The approach is developed through an example of steps to follow in identifying clusters of homogeneous items and then expanding the number of items in each cluster to achieve acceptable measurement reliability.

The importance of homogeneity among test items is emphasized throughout Appendix B, just as the same point is emphasized in Appendix A. Homogeneity is essential to reliability. Furthermore, in criterion-referenced measurement, item homogeneity is vital to one's ability to make diagnostic interpretations of achievement test results. Dependable diagnoses of learning difficulties provide the essential basis for devising alternative instructional approaches to improve the payoff from instruction.

III DESIGNING AND EVALUATING ALTERNATIVE INSTRUCTIONAL TREATMENTS

Introduction and Overview

This section defines, describes, and illustrates an operational approach for a continuing program of research and evaluation directed toward improvement in technical training. The approach builds upon the preceding sections, as well as the appendices, in which the following elements were developed:

1. The relationship of technical training to on-job performance
2. A classification of predictors of performance in training
3. Diagnostic evaluation of both curricula and means for measuring trainee performance to identify areas of desirable improvement
4. Rudiments of experimental design and statistical models appropriate to the analysis of data from such experiments

The principal new ingredient in this section, developed in some detail to illustrate certain steps in the integrated approach, is an example of how one might analyze and evaluate the performance of trainees as a basis for designing alternative instructional treatments to be tried experimentally and evaluated before adoption or rejection.

Figure 9 displays the main functions in a continuing program to evaluate, refine, and improve technical training.

The upper box in Figure 9 identifies the required external evaluation -- the critical exchange of evaluative information between an instructional program and the commands that use its graduates. A central function of those responsible for management of an instructional program is to assure that this information exchange is sustained continuously.

The lower box in Figure 9, labeled internal evaluation, encloses those functions related most closely to the appraisal, revision, and empirical tryout of instructional programs intended to improve the quality of instruction. The continuous exercise of internal evaluation functions is the responsibility of staff of a training facility.

Each of the functions illustrated in Figure 9 is discussed in greater detail in the paragraphs that follow.

Linkages Between External and Internal Evaluation

The ISD model represents both internal and external evaluation as elements of the Control phase of instructional systems development, thus underscoring (a) the importance of systemetic evaluation as a basis for revision and improvement of instruction and (b) program management responsibilities for assuring instructional quality. The ISD model offers many useful suggestions for the planning and conduct of internal and external evaluations, as well as ways in which evaluation data can be used to guide revisions in instructional content or procedures.

The ISD approach generally assumes the existence of an ISD Program and an ISD Program Manager with responsibility for assuring that evaluation functions are appropriately planned, documented, and carried out. In the ISD concept, plans for internal evaluation are developed in parallel with plans for the instructional program. Internal evaluation occurs throughout all stages of instruction. The primary objective of internal evaluation is to determine the degree to which specific instructional objectives are met. A secondary objective of internal evaluation is to ascertain whether the ISD process was successful.

External evaluation, in the ISD model, occurs after students have completed instruction and have been assigned to jobs. The focus of external evaluation is on post-instructional performance. External evaluation assesses the quality of the ISD job analysis toward which instruction was directed and the "fit" between job requirements and the instruction provided. Thus, external evaluation identifies job tasks for which provided instruction was not adequate. In the ISD concept, it is preferable to assign external evaluation functions to agencies that were not involved in the instruction. It is assumed that this detachment enhances objectivity.

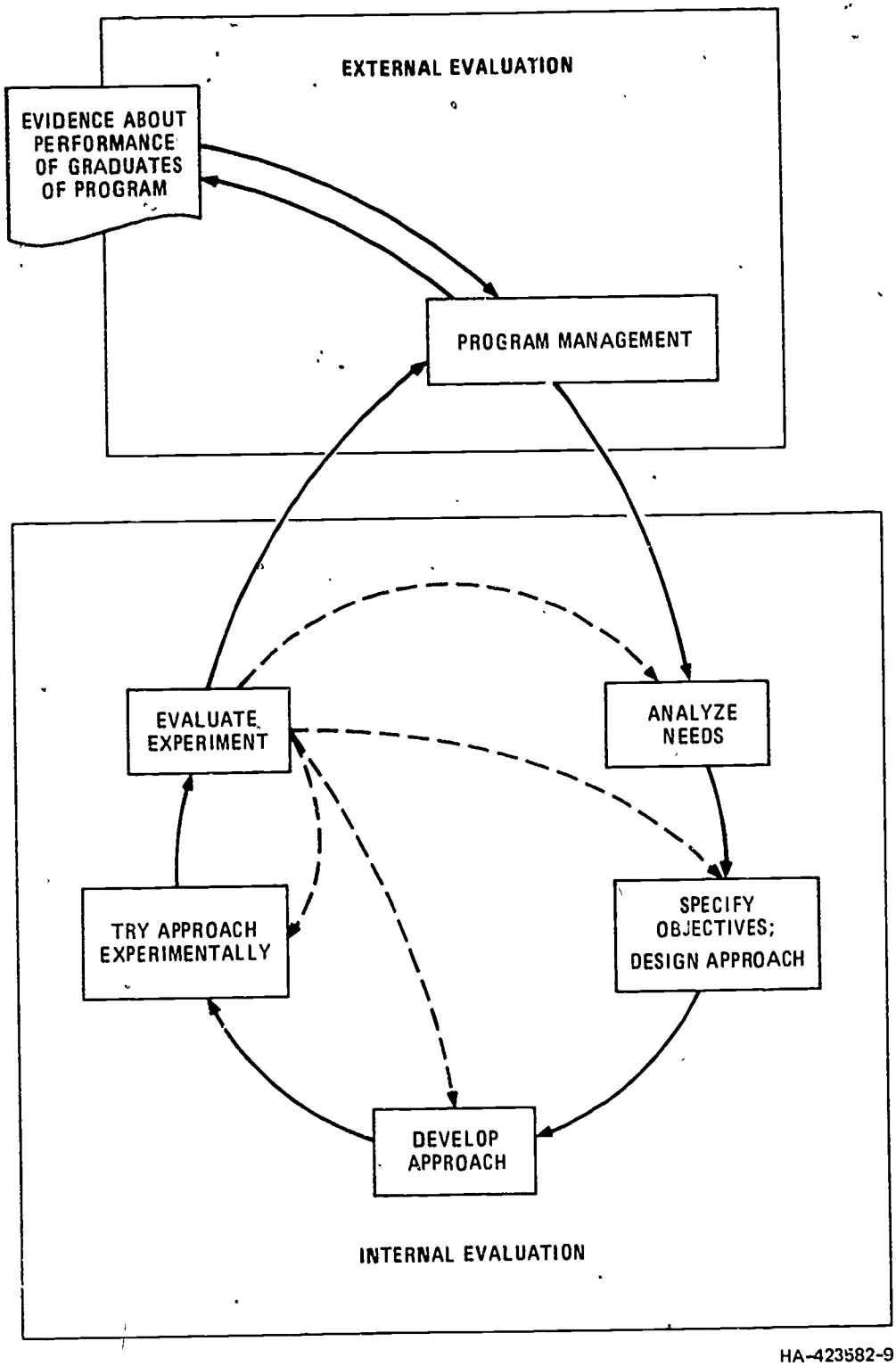

FIGURE 9 AN INTEGRATED MODEL TO GUIDE DESIGN AND EVALUATION OF ALTERNATIVE INSTRUCTIONAL TREATMENTS

The distinctions drawn between internal and external evaluation by the ISD model have many similarities to formative (internal) and summative (external) evaluation (see, for example, Scriven, 1967). Even though summative evaluation may be performed competently by someone also involved in formative evaluation, the issue of credible objectivity remains. It is wise to separate the functions as much as possible.

The ISD concepts of internal and external evaluation are sound, and the suggestions in the ISD manual for conducting both are practical ones. This report endorses the concepts and urges their systematic application.

The model shown earlier in this report as Figure 9 is intended to be in harmony with the ISD model. Figure 9 is drawn to emphasize a point, however, that does not always emerge clearly from the ISD formulation: it is a school management (or command) responsibility to assure that both external and internal evaluation occur. This responsibility holds whether or not it is feasible to establish ISD Program teams or assign external evaluation functions to outsiders.

"Program management," as identified in Figure 9, refers to the management of the instructional program -- the commanding officer, course supervisor, and the staff responsible for the school or training facility. When it is possible to establish an ISD Program staff, accountable in its performance to the school, such an arrangement is desirable and likely to result in more attention being given to evaluation functions. When such arrangements are not possible, the evaluation functions remain to be addressed by the best available alternate means.

This report supports fully the position that the ultimate criterion for judging the adequacy of job-related training is the performance of the school's graduates when assigned to those jobs for which their instruction was intended. This viewpoint was presented first in Figure 1 in Section I and represented again in Figure 3 in Section I. The same viewpoint is focal to the relationships shown above in Figure 9.

An internal evaluation that reveals weaknesses in instruction by identifying objectives that are not being met almost necessarily means that an external evaluation also will identify flaws in the instructional program. ("Almost necessarily" because there is a remote chance that instruction will meet job task requirements but not fully satisfy internal course objectives. If this occurs, it means that the instructional program is over-designed against the criterion of job performance.)

An internal evaluation that shows instructional objectives as being met does not necessarily mean positive findings from an external evaluation. A course may appear effective by internal criteria and yet be found inadequate in some respects when judged by external criteria.

This could be caused by many factors, including: changes in the job, inadequate instruction, poor analysis of the job to be trained, or invalid external evaluation.

Evaluation of instructional content and procedures must refer back to the job requirements that provide the justification for instruction in the first place. It is not sufficient to show, through internal evaluation, that course objectives are being met unless it also can be shown through external evaluation that the objectives are appropriate to the requirements of the jobs for which the instruction is intended.

Assessing Needs for Alternative Instructional Treatments

Needs assessment is the process of identifying and specifying the differences between desired (or acceptable) conditions (often stated as goals) and present conditions. Discrepancies between desired and existing conditions are called needs. End products of a systematic needs assessment are statements of objectives that the instructional program will seek in order to eliminate discrepancies between what is desired and what currently exists -- objectives designed to "meet the needs." The means for achieving these objectives constitute the program plan.

There is no special mystique or hidden trick in needs assessment -- the process amounts to a commonsense audit to create, so to speak, a balance sheet in which goals are rectified against accomplishments to determine needs. Some reminders may be helpful, however:

1. Systematic audits of goals and accomplishments should be undertaken periodically to assure that no important viewpoints or sources of evidence are overlooked.

2. Outcome goals should be stated in language that specifies a measurable attribute (ability, skill, application of knowledge, attitude toward job, etc.). Process goals will identify conditions that trainees or instructional staff will experience or cause to occur. An outcome goal should fit a sentence stem such as the following: "At the end of the program (unit, lesson, segment, block), trainees will demonstrate..." A process goal should fit a sentence stem such as, "During the program, trainees (instructors) will experience (take part in, do, cause, observe)..."

Needs assessment, then, serves both diagnostic and prescriptive purposes by identifying discrepancies that call for correction, and providing information from which to rectify discrepancies. The tools of needs assessment are very much like those of evaluation -- end-goals (products) must be stated in measurable terms, process-goals and implementation objectives must be defined, and measures must be obtained from which to assess the manner in which existing conditions meet standards defined by the program's goals.

Needs assessment calls for both external and internal evidence. External evidence of training needs will come primarily from feedback about the performance of graduates of training programs, and from information about job task requirements.

Internal evidence of training needs will come largely from assessment of the degree to which course objectives are being satisfied -- failure rates are higher than desired, training time is longer than intended, certain trainees have unusual difficulties with some materials, and so on.

The importance of integrating external and internal evaluation may be illustrated by a commonsense question that is not always raised -- what characteristics of performance during training are associated with later performance on the job?

Among staff in a training facility, it is all too easy to assume that the best performers during training also will be the best performers on the job. Conventional instructional wisdom usually views rapid learning as ideal in a self-paced instructional environment, and error-free learning as ideal in a group-paced instructional environment. But it does not necessarily follow that the trainees who perform best in an internal evaluation will be evaluated similarly by their supervisors on the job.

More specifically, consider a cross-classification of types of trainees according to two performance measures in a criterion-referenced self-paced instructional environment:

Time to Criterion	Attempts to Criterion: Few	Attempts to Criterion: Many
Long	B	D
Short	A	C

From a training performance perspective, Type A performers pose the fewest problems (are "best"), and Type D performers pose the most problems (are "worst"). Types B and C represent contrasting styles; if speed is given high weight, then C is better than B but if error-free performance is given high weight, then B is better than C. Left unanswered is the question of how such performance patterns relate to performance on the job. If Type B trainees perform better on the job than do Type C trainees, then efforts by training staff to minimize only training time could be misdirected.

The formulation shown in Figure 9 implies a separation between steps or stages of internal evaluation. Actually, the functions of analyzing needs, specifying objectives, and designing alternative approaches are very closely coupled to one another, and all draw from a common base of evidence. Needs assessment feeds directly into the

specification of objectives for new or revised programs, while the design and development of programs fulfill those objectives.

Readers will recognize similarities between needs assessment functions, as discussed above, and earlier discussion in Section I in conjunction with Figure 3. The earlier formulation illustrated in Figure 3 is a decision-oriented needs assessment that focuses on trainee performance measures, both during and following training, as the prime sources of evidence from which to judge the need for devising alternative instructional treatments. A central point in the earlier arguments was that reliable and valid measures are essential to a diagnosis of program strengths and weaknesses.

Specifying Objectives and Designing Approaches

Instructional objectives and their associated instructional treatments are statements about the expected performance of trainees following exposure to a segment of instruction. As noted above, an explicit performance objective will specify a measurable attribute that a trainee will demonstrate as evidence that the objective has been achieved.

In criterion-referenced measurement, particularly with student self-pacing, the key measures of instructional effectiveness are time to criterion achievement, and number of attempts to successful performance on the criterion test. Subordinate measures also may be obtained, such as the time to the first attempt of the criterion test and test score obtained on the first attempt (or on each attempt). (See Appendix D for an elaboration of these measures and discussion of the ways in which they related to one another in an analysis of actual data from technical training in a computer-managed environment.)

Objectives specific to units or segments of instruction in a criterion referenced, self-paced instructional environment normally will refer to measures of performance such as those noted above. Other program objectives may refer to additional measures of importance, such as the following:

1. Average total time for course completion.

2. Variability among trainees in the total time to course completion.

3. Failure rates.

The task of contriving real alternative treatments to satisfy instructional objectives is more difficult than stating the objectives. Appendix D presents an illustration of steps that can be taken in analyzing trainee characteristics and performance as a basis for developing alternative instructional designs.

Appendix D contains two main parts. The first part concerns relationships among four measures of trainee performance in a computer-managed instructional setting in which instruction was largely self-paced. The four performance measures are (a) measured time to first attempt of criterion test (MTM), (b) score on first attempt of criterion test (LSC), (c) measured time to successful performance on the criterion test (LTM), and (d) number of attempts to criterion (NATT). Among other things, the analysis in Appendix D shows the following:

1. Time to first test attempt (MTM) explains some 86% of the variability in time to criterion (LTM), since MTM is a component of LTM. With the addition of the LSC as a predictor, about 92% of the variability in LTM is accounted for.

2. First attempt score (LSC) accounts for some 66% of the variability in number of attempts to criterion (NATT). Time to first attempt (MTM) adds nothing since the time difference between MTM and LTM is essentially unrelated to MTM. This implies that other factors, such as stylistic differences, must also account for some variance in NATT.

3. LSC and MTM are correlated, but the relationship is not a strong one ($r = -.36$; $r^2 = .13$). A cross-tabulation of LSC and MTM scores provides one way to characterize four broad groupings of trainees according to their response styles: A = fast and accurate, B = slow and accurate, C = fast and inaccurate, and D = slow and inaccurate.

In the last portion of Appendix D, a hypothetical analysis is illustrated with fictitious data to show how certain trainee attributes could be examined in relation to performance measures as a means for interpreting differences in response styles. This hypothetical example shows a "basic skills" factor as the dominant predictor in accounting for LTM. The example also shows that basic skills may interact with other variables in unexpected ways. For example, with the fictitious data, an "anxiety" factor appears to interact with basic skills such that the combination of high skill and high anxiety is better for predicting low or short LTM than is the combination of high skill and low anxiety. Some speculative questions are derived from this hypothetical analysis to illustrate a possible approach to the problem of contriving alternatives to the present instructional treatment as a way of reducing the average time to criterion and reducing the number of errors (and hence the number of attempts before achieving criterion).

To extend the analysis beyond the point reached in Appendix D, imagine that three ideas evolved as potential alternatives and an investigation was needed to determine whether any of these had the desired effect of reducing time or reducing errors or both:

- The first idea is to use a pretest or screening test to estimate whether a trainee should proceed directly to the criterion test rather than spending time on instruction. This idea is illustrated in Figure 10. The evaluation questions implied by Figure 10 are the following:

  1. What is the validity of the pretest as a predictor of scores on the criterion test?

  2. What is the validity of the pretest as a predictor of time to criterion and number of attempts to criterion?

  The idea illustrated in Figure 10 implies a cutting score (a score which is "high" enough) on the pretest but does not define it. The best way to identify an appropriate cutting score would be to ignore the pretest score at the first decision point in Figure 10 and assign trainees at random to one or the other path to the criterion test. This would assure the full range of pretest scores on both tracks shown in Figure 10. Specification of an appropriate cutting score for future use would then be based on actual performance data.

  As an alternative to this, a moderate level cutting score could be defined provisionally with the expectation of adjusting it as performance data accumulated. This approach is not as "pure" from an experimental design viewpoint, but it has the practical appeal of reducing the number of trainees with low pretest score who would by-pass instruction and be almost sure to then fail the criterion test.

  We suggest starting with a fairly low cutting score on the first few trials and gradually moving it higher with successive blocks of trainees. We would increase the cutting score based on considerations of testing time vs. instructional time saved. Any cutting score from imperfectly reliable tests, it must be remembered, will be wrong some of the time.

  The principal gain from the screening test idea illustrated in Figure 10 will come from forcing trainees who do not need the instruction to bypass it. In terms of trainee response styles referred to earlier, some "accurate but slow" trainees will be forced to the fast track and some "inaccurate but fast" trainees may be encouraged to be more deliberate.

- The second idea is a variant of the first one. In this alternative, the length of criterion tests would be increased -- for example, doubled in length. Trainees would be strongly encouraged to attempt the criterion test as soon as possible, but also would be advised that they were not expected to pass it on the first attempt.

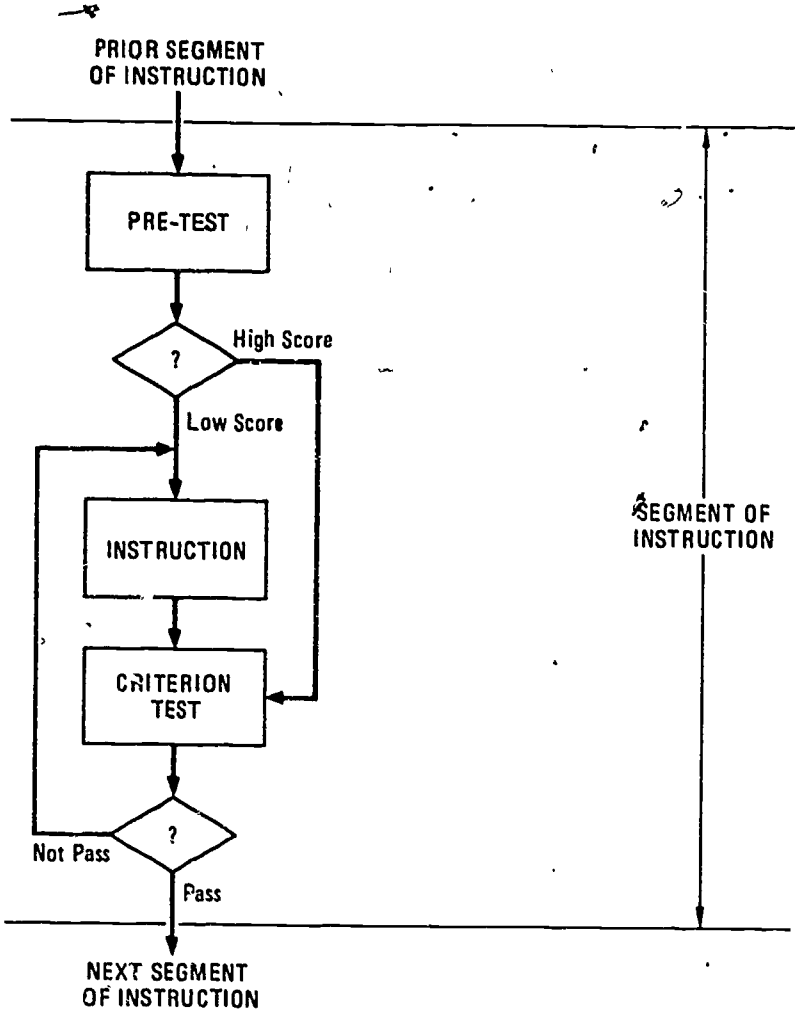

FIGURE 10 EXAMPLE OF USE OF PRE-TEST TO SCREEN TRAINEES FOR CRITERION TEST READINESS

The main argument underlying this alternative is that trainees may learn more from failing a test item than from answering it correctly. The proposition to be tested is that feedback from errors made over several rapidly paced trials will lead to earlier criterion performance than will fewer, more deliberate, trials with few or no errors.

Important additional purposes will be served by increasing the length of the criterion test. First, a test with many items makes it easier to be certain that all important instructional content is represented in the test. (In measurement jargon, the "domain sample" is increased so that the "curricular validity" of the test will be greater.) Second, test reliability is a partial function of test length. The longer the test, the more reliable are the scores obtained from it. (See Appendixes A and B for discussions of test reliability as a function of item homogeneity and the number of items.) The point is that if we are going to pass trainees with less instructions, we want to be quite sure about the correctness of our pass-fail decisions.

- The third idea is somewhat more radical than the preceding two, both of which assume individual self-pacing. The third alternative introduces two notions that depart from the self-pacing mode: group-pacing and peer tutoring.

The rationale for the third idea grows from the range of response style differences among trainees. To repeat the earlier categorization:

Time to Criterion	Attempts to Criterion: Few	Attempts to Criterion: Many
Long	B	D
Short	A	C

The idea is to create tutorial groups by cross-matching opposite types and, also, to define standards of performance that groups must satisfy before individual group members progress to the next instructional segment. Thus, one group would be composed of Type A and Type D trainees and the other group would be made up of Type B and Type C trainees. (Several groups could be organized. For example, if there were 10 Type A trainees and 10 Type D trainees in a class, one might organize from three to five A-D groups so each tutorial group would be fairly small.) The group standard for progress to the next instructional segment might be defined as successful criterion performance by a specified percentage of all trainees in the group; for example, 80% or so. It might be specified further that all trainees must attempt the criterion test within some defined time limit.

The intent of the A-D grouping is to assure that Type D students have able tutors. The imposed requirements of group performance are to prevent Type A trainees from abandoning their slower companions. Similarly, the intent of the B-C grouping is for each to favorably influence the other -- for Type B trainees to help Type C trainees with instructional substance and for Type C trainees to help Type B trainees overcome some of their apparent reluctance to commit an error.

Prior research offers some clues as to what one might expect from a tutorial group experiment. For example, it is often found that those who tutor receive greater benefit than do those who are tutored. If such a finding were to occur in the training experiment, it might suggest that Type A trainees should be excluded from tutorial groups and that tutorial groups be composed of roughly balanced combinations of Types B, C, and D trainees.

The imposition of a group standard of performance could have undesirable effects on trainees' attitudes toward instruction, especially if the group standard meant that too many otherwise qualified trainees were being detained while their slower companions caught up with them. On the other hand, if the experiment worked out, the net effect should be (a) a reduction in the average number of trials, (b) a reduction in performance variability across the aggregate of all trainees, and (c) a reduction in the overall average time to criterion. These net effects would be the result of faster performance and fewer errors by Type D trainees, faster performance by Type B trainees, and fewer errors by Type C trainees. Type A trainees might not show faster or more accurate responses on the criterion test than they do now, but they should benefit from their tutorial roles.

A group tutorial experiment as described above could not be arranged until the first few units of instruction in a course had been completed so that trainee response styles could be reliably established. Means for organizing groups could be considered as an experimental question in itself since the effects of heterogeneous grouping on individual motivation can only be guessed. Some trainees might be more effective as tutors if they knew that was why they were in the group. On the other hand, less able trainees could find that their performance was inhibited, rather than facilitated, by knowledge that they had been paired intentionally with more able companions who were expected to help them. The questions are worthy of research under the heading of "internal evaluation."

Developing Alternative Approaches

The task at this stage is to operationalize the concepts developed in the preceding stages so that an evaluation experiment can be conducted. Discussion of steps in developing an alternative approach suitable for tryout is illustrated by extending the three alternative design ideas just presented.

The Screening Test Design

The first idea was to use a screening test or pretest prior to beginning an instructional segment. Based on pretest performance, trainees would be either routed directly to a first attempt on the criterion test or required to go through the instructional materials. The purpose of the approach is to reduce average total time to criterion by forcing an immediate attempt on the criterion test by trainees with a high probability of successful performance.

The new product needed to evaluate the utility of this approach is a pretest that is parallel to the criterion test. By a strict definition of "parallel," this means a new test that covers the same content areas, is identical in difficulty, and results in a score distribution identical to that of the criterion test if the criterion test also were administered prior to instruction. If literal identity between the two tests could be satisfied, the correlation between them would be +1.0. It is impossible to wholly satisfy the literal requirements of a parallel test. Nevertheless, statistical identity in test parameters (content, mean, variance, number of items) defines the development objective.

Developing a parallel test to be used as a pretest offers an ideal opportunity to reexamine and refine the criterion test as well. Accordingly, the parallel test development should begin with analysis of the existing criterion test. Particular attention in this analysis should be given to item homogeneity, as discussed in Appendixes A and B.

Accumulated records of the performance of earlier trainees on the criterion test may be sufficient for this analysis. If not, an initial step will be to obtain response data on the criterion test items from a sample of persons like those for whom the test is intended. Guidelines for sample size relative to the number of items are discussed in Appendixes A and B.

Development of the parallel form probably is undertaken most conveniently by treating the existing criterion test as an item pool, each item of which is to be paralleled:

1. Be sure that the pool of items (i.e., the existing criterion test) provides content coverage of the instructional segment. A systematic way to make this check is first to prepare a comprehensive list of the objectives and sub-objectives in

the instructional segment and then to identify test items that apply to each objective and sub-objective. If there are no test items for an objective or sub-objective, either additional items are needed or the relevance of the objective should be reconsidered.

2. Determine the homogeneity of the items in the pool (see Appendix A or B). If non-homogeneous items are counted as part of the same test, re-group items into homogeneous subsets. (A subset, at this point, might have only one item.) If criterion performance on the test had been defined earlier as correct responses to a proportion of all items and all items are not homogeneous, re-define criterion performance as correct responses to proportions of each homogeneous subset.

3. Prepare new items that are parallel to items in the existing pool for each content area defined by an instructional objective or sub-objective. Each instructional objective or sub-objective should now have at least two test items; preferably, some multiple of two (i.e., 4, 6, 8, etc.). The number of items per objective should be roughly proportional to the importance of the objective.

4. Consider the original items and the new ones as a single pool. When content coverage, editing, format, and other features appear satisfactory, administer the entire pool of old and new items to a new sample of persons like those for whom the test is intended. Perform an item analysis (Appendix B provides an example of issues to consider in such an analysis). Cluster items into relatively homogeneous subsets as necessary.

5. Construct two parallel forms of the total test so that each form covers the same content and both are as nearly similar in other respects (e.g., item means, test means, test variances) as can be managed. Flip a coin and call one form the "pretest" and the other form the "criterion test." If it proves impossible to achieve very high similarity between the two forms, designate the easier form as the pretest.

The Extended Criterion Test Design

The second idea presented earlier (i.e., encouraging earlier trials on tests of increased length) is based on the notion that testing oneself to get feedback from errors is a constructive strategy for accelerating learning time. Making errors is uncomfortable for many people, however, so trainees must be shown that errors on the first trial or two will help them focus their efforts to learn and are therefore desirable. Criterion tests would be lengthened to increase the likelihood of errors in early trials and also to improve the quality of performance measurement.

The procedure for developing a pretest that parallels the criterion test also is appropriate for lengthening the criterion test if one wished to try out this second idea. The only difference in test-development procedures is that the two forms (pretest and criterion test) simply would be merged into one longer criterion test. Procedural steps for developing new items, however, would be the same as already described.

The Tutorial Groups Design

The tutorial groups idea does not require the development of new instructional materials or additional test items. Arranging conditions for tryout of the tutorial groups idea would call for analysis of trainees' performance in prior segments of instruction so that grouping could be based on established response styles. Working definitions of acceptable group performance also must be specified (e.g., time limits within which criterion tests must be attempted and the proportion of a group that must meet criterion before the group progresses to the next instructional segment).

If this approach is to be evaluated properly, such evaluation data as trainee reactions to the procedure should be obtained. Several possibilities are reasonable, ranging from group or individual interviews to questionnaires with a few rating scales and space for comments. Performance data by group may be collated by aggregating performance information as collected currently from individual trainees; all that this implies is a code in each trainee record to define trainee response style classifications and tutorial group assignments so that data can be grouped appropriately for later analysis.

Evaluating Alternative Instructional Approaches Experimentally

Evaluating the merits of one or more alternative instructional approaches relative to the instructional approach currently being used calls first for developing a plan. The term, design, is more formal than plan, but a design is no more and no less than a plan that specifies conditions and procedures for obtaining and analyzing the data on which to base the evaluation.

In the closing portion of Section I of this report ("Planned Experiments with Alternative Instructional Treatments"), some basic design principles were discussed in relation to models for statistical analysis appropriate for dealing with the data. Also, Appendix C contains an overview of multiple regression analysis since this statistical model is most appropriate to situations in which several variables are of interest, groups may not be equal in size, and the variables may be a mixture of both categorical and continuous scale measurement. Appendix C includes some discussion of ways to code experimental and control groups so that aptitude-by-treatment interactions can be estimated.

A few points on the logic of experimentation, statistical tests of experimental hypotheses, and an illustration of some basic designs will simplify later discussion.

As noted in Section I, experiments are performed to test propositions or hypotheses of the general form, "If X under such-and-such conditions, then Y will be observed." In this formulation, "X" defines the experimental variable that is under experimenter control, "such-and-such conditions" define the circumstances under which the experiment occurs, and "Y" defines the outcome measure or the dependent variable.

For example, suppose that for a particular block of instruction the first lesson was critical for understanding subsequent material in the block. We might hypothesize that students with poorer reading skills would benefit if this lesson was presented in a filmstrip mode with few reading demands. In order to test this hypothesis, we convert the lesson to filmstrip and randomly assign students to either the new or old lesson and observe the effect on block time and score. Here, "X" is which lesson is taken, the conditions are random assignment, and "Y" is block time and score.

Experiments are designed to ensure, insofar as possible, that the outcomes are due to the experimental manipulations, i.e., are direct tests of the experimental hypothesis. In the previous example, we randomly assigned students so that we could, after block completion, identify the poor readers (using previously obtained measures of reading ability) and compare block performance as a function of which lesson was taken. As an aside, this design also permits us to compare the performance of better readers.

The traditional statistical approach to framing the experimental question is to say that performance will not be affected by which lesson was taken. This is called the null hypothesis. We then use block times and scores to reject this hypothesis, i.e., to show that there actually was a difference. If we are successful in rejecting the null hypothesis, the only conclusion we can come to (because of the experimental manipulation) is that the filmstrip lesson caused differential performance.

A numerical difference between experimental conditions sufficient in magnitude for one to conclude that observed differences are not attributable to chance factors is a function of the size of the observed difference divided by the standard error of that difference. The standard error, in turn, is a function of the number of observations (sample size) and the variability of the measures. Commonsense prevails -- small differences with very small standard errors may be statistically significant and large differences with large standard errors may not be statistically significant.

Statistical analysis is based on the theory of probabillty which assumes that samples of cases are drawn randomly from the population to

which one wishes to generalize. Estimates of population values (parameters) are made from sample values (statistics). Any specific sample may be good, bad, or indifferent as a basis for estimating the population value. Without drawing all possible samples, one never knows for sure whether a specific sample provides a close estimate of the population value. Sampling variation is one of the "chance factors" that may lead to rejecting, or not rejecting, a hypothesis.

Inferences about population values based on sample values make use of theoretical distributions (e.g., the bell-shaped, normal curve) as a foundation for probability statements about the likelihood of an obtained value differing fron an expected one or about differences between one or more obtained values. When an analyst says that a value (such as the difference between the mean of an experimental group and the mean of a control group) is "statistically significant," the analyst is using shorthand to say that the probability is so small that a value as large would be obtained by chance, that the null hypothesis of "no difference" has been rejected.

Significance levels are defined in probability terms. For example, values such as $p < .05$ may be attached to an obtained value (such as the difference between two means). This probability statement -- $p < .05$ -- means that the chances are less than 5 in 100 that a value that large would be obtained if the null hypothesis were true.

Specifying a probability level for acceptance or rejection of the null hypothesis is a matter of convention; such probability values as .05 or .01 are commonly used. These levels, however, are arbitrary and adopted for convenience. It might be better if the habit of rejecting or not rejecting a null hypothesis at some arbitrary level (e.g., $p < .05$) were abandoned entirely and associated probabilities simply reported and interpreted. Furthermore, it is important to note that a difference can be statistically significant without being significant from a practical standpoint. Does a time gain of X-minutes per trainee, which may be statistically significant, make any practical difference in the way that instruction is conducted? Such questions are worth asking when one interprets statistical findings.

It is also important to note that an inference drawn from an experiment may be incorrect. Consider the following tabulation:

	True Situation	
Conclusion	No Difference	Real Difference
Real difference	Type I error	Correct
No difference	Correct	Type II error

Thus, if one rejects the null hypothesis by concluding a "real difference" when in fact there is not a real difference, a so-called Type I error has been committed. The probability of a Type I error is equal

to the significance level that has been defined for rejecting or not rejecting the null hypothesis; if the .05 level is defined, then the chances of a Type I error (false rejection of the null hypothesis) are 5 in 100. The significance sword has two edges, however. If one sets a more stringent level (say, .01) for rejecting the null hypothesis, the chances increase that a Type II error (false acceptance) will be committed. The chances of a Type II error can be reduced by relaxing the level for rejecting the null hypothesis (e.g., $p < .05$ instead of .01 or $p < .10$ instead of .05), using a more powerful statistical test, and increasing the sample size.

This discussion of hypothesis testing is relevant to experimental design considerations and suggests the following guidelines:

1. Samples should be randomly drawn if one expects to make appropriate use of the methods of statistical inference. If one is comparing the effectiveness of rival methods for instruction, the people who experience the methods should be assigned to one method or another by a randon procedure. As will be noted later, one must pay attention to randomization if the intent is to avoid erroneous conclusions.

2. Samples should be as large as reasonably possible. Larger samples mean smaller errors and fewer errors of inference.

3. Hypotheses or propositions about expected effects of rival treatments should be as specific as possible. A hypothesis asserting that Treatment A will be superior to Treatment B permits a directional statistical hypothesis to be tested. Directional hypotheses are more powerful than nondirectional ones. If the theory or arguments that led to the design of alternate treatments is sufficiently persuasive for one to predict the direction of difference between rival treatments, then the experimental hypothesis should be so stated.

4. View hypothesis testing in probability terms rather than in categorical "reject-not reject" terms. A sizeable difference obtained in an experiment with small samples may fail to satisfy some predetermined level for rejecting the "no difference" hypothesis (say, $p = .20$ instead of $p < .05$). One certainly would not claim discovery of immutable truth on the strength of such weak statistical support. On the other hand, if the difference is in the hypothesized direction and the principal reason for the high probability value ($p = .20$) associated with the finding appears to be the small sample size, one would want to test the hypothesis again with a new and larger sample, rather than abandoning it. Conversely, a "so what?" question is appropriate even when "statistical significance" is achieved -- a difference can be reliable without being of practical worth.

Certain conventions have become common for representing an experimental or quasi-experimental design (see Campbell & Stanley, 1963). To illustrate, Campbell and Stanley portray the Pretest-Posttest Control Group Design (a "true" experimental design) as follows:

R O X O

R O O

In this representation, time runs from left to right. If symbols are vertical relative to one another, they occur at the same time. If symbols are on the same line, the events are experienced by the same persons. Symbols have the following meaning:

R = Random assignment
O = Observation (measurement) of some kind
X = Exposure of a group to an experimental variable or event; i.e., the experimental "treatment"

Thus, the representation above shows that (a) subjects are randomly assigned to experimental (X) or control (blank) treatments, (b) observations (measurements) are made prior to exposure (the pretest), (c) experimental treatment occurs, and (d) observations (measurements) are made following exposure (the posttest).

This notation scheme uses a blank space to represent the treatment experienced by the control group. This does not necessarily mean that nothing occurs. The blank space usually will denote a rival treatment -- for example, the "old way" to which the "new (experimental) way" is being contrasted.

Campbell and Stanley also use a horizontal dashed line separating two groups to indicate that groups on either side of the line are not equivalent to one another (for example, intact classes or groups not randomly assigned). To illustrate, a non-equivalent control group design (a quasi-experimental design) is symbolized as follows:

O X O

O O

The discussion that follows builds on both Section I and Appendix C. The emphasis is on clarifying who will be measured, when they will be measured, the kinds of measures to be obtained, and how the data from measurements will be processed and analyzed.

Treatment Groups

In instructional evaluation research, a treatment group is a batch of people defined by the treatment or instructional program they experience. Evaluation of alternative treatments implies comparisons between at least two groups -- one that experiences the existing

instructional program and one that experiences an alternative to the existing program. If more than one alternative is being considered simultaneously, then there will be as many experimental groups as there are alternative treatments to be compared.

The term, control group, often is used to denote the reference group against which the experimental group or groups will be contrasted. Perhaps a more descriptive term would be "comparison group," thus reserving control group for describing true experiments in which one can assure insulation of one group from another.

Evaluations of alternative instructional treatments are made to help guide such decisions as the following:

1. Which approach is better, on the average?

2. Which approach is better for trainees of such-and-such characteristics?

3. Which approach is most difficult to implement?

4. Which approach is preferred by trainees?

5. Which approach is preferred by the instructional staff?

To provide credible evidence to support conclusions about which approach is best, preferred, least difficult to manage, or whatever, it is critical not to "stack the cards," intentionally or by accident, in favor of one rival approach over another. By far the best way to guard against card-stacking is to assign people to treatments by a random procedure -- that is, by a procedure that gives everyone an equal chance of being assigned to one of the treatments.

Randomization gives each treatment an equal chance of being applied to any subject (person). However, randomization does not guarantee to balance out natural differences among subjects. If some personal attribute is known (or strongly suspected) to be related to the dependent variable performance, various straightforward methods can be used to approximate balance between or among groups on that attribute. For example, if reading speed is likely to be related to performance on the dependent variable measure and one has prior measures of reading speed, all eligible subjects can be numbered from high to low on reading speed in advance of their random selection for treatment group assignment. Using an unbiased method of selection, subjects then can be assigned systematically from highest to lowest to one or another treatment. For instance, with only two treatments, one might designate subjects for treatment by the flip of a coin (e.g.,"heads = odd numbers = Treatment A"). This procedure would tend to balance reading ability between the two treatments.

With three or more treatments to compare, systematic sampling can be followed. For example, with three treatments, one wants three

groups of approximately equal size and average reading speed from the numbered list generated as just described. One may scan a column from a table of random nunbers in search for last-place digits of 1, 2, or 3. The first of these digits found -- 1 or 2 or 3 -- will designate Treatment A and the second digit found will designate Treatment B; since both Treatments A and B are then defined, Treatment C is defined by the remaining number. If the first number found was "2," then Subject Number 2 and every third number thereafter (i.e., 5, 8, 11, 14, and so on) would be assigned to Treatment A. Such a systematic procedure satisfies two requirenents: (1) individual assignment to treatment group has been determined randomly and (2) each treatment group will include approximately the full range on the reading speed measure, thus essentially equating treatment groups on reading speed.

Sometimes it is not possible to make individual assignments of subjects to groups -- administrative groupings or natural groupings must be left intact. For example, with classroom instruction that is instructor-paced, it is seldom possible (and may not be appropriate in any event) to break up the class of students. Now one unit of analysis becomes the class and randomization must be applied to classes defined by times of day, instructors' names, etc.

Difficult problems arise in the analysis of data obtained from intact groups. Such factors as instructor and class group now become factors in the design -- the comparison is no longer a simple Treatment A vs. Treatment B comparison between randomly assigned students, but Treatment A with Instructor Jones in the morning class vs. Treatment A with Instructor Smith in the afternoon class vs. Treatment B with Instructor Brown in the evening class, and so on. Furthermore, the characteristics of persons in each class may be systematically different (e.g., Instructor Jones has mostly "fast" students and Instructor Smith has mostly "slow" students and Instructor Brown has the students who enrolled last).

Statistical problems can be managed, although not always neatly. For example, each combination of treatment by instructor by intact group can be coded as a unique treatment in a multiple regression analysis. The interpretation of findings is more difficult than in simpler arrangements, however, since so many factors likely to influence the findings are confounded.

If circumstances require that intact groups be used, the principle of randomization of treatment assignment is no less important. Given time and resources, it may be possible to repeat the experiment enough times so that each instructor applies each treatment with several classes. Larger samples and more replications (repetitions) usually turn out to be about the best one can do to balance things out in experiments that must be carried out with intact groups.

The designation of who uses which treatment with which group in which replication can still be made by random procedures. It is really impossible to pay too much attention to randomization. As a principle,

if a random procedure can be used to decide who experiences what treatment when, then a random procedure should be used.

Times and Kinds of Measurement

Measurements may be made at many different times relative to the period of instructional treatment depending on what the measures are expected to reveal.

The most important measurement point is that immediately following exposure to the rival treatments. For example, using the Campbell and Stanley notation, a posttest-only control group design would be symbolized as follows:

R X O
R O

The random assignment of subjects to treatment guards against systematic bias in the characteristics of subjects, including their readiness for the instructional treatment. Thus, there may be no important purpose served by a pretest. The posttest, however, is the direct measure of effects; this measure should be made as soon as possible following treatment so that other influential factors do not intervene between treatment and measurement of effect.

Sometimes it may not be possible to do more than estimate the effects of a variation in treatment on the momentum established by prior instruction. Such a time series design may be symbolized as follows:

O O O OXO O O O

Measures preceding the experimental treatment establish a trend line or norm of progress prior to the experimental intervention. In plotting measures over time, one might hope to find a disjunction in the trend line at the time when the treatment is introduced. This is a quasi-experinental design in which subjects essentially serve as their own controls. It sometimes is referred to as a "regression discontinuity model."

In self-paced instruction, it may be possible to combine the benefits of time-series measurements with true experiments embedded within a progression. The following would symbolize such a design:

$$O\ O\ O\ O \begin{bmatrix} R & X & O \\ R & & O \end{bmatrix} \begin{matrix} O & O \\ O & O \end{matrix} \Big\} O\ O\ O$$

The single string of observations preceding the bracketed posttest-only control group design denotes that all trainees experience common

treatment up to the point at which a random split is made to test an experimental segment. Following the posttest, trainees may merge again into a single group. Alternatively, differentiated measurement could continue for a time as a way of measuring the persistence of effects from the experimental treatment.

One difficulty in symbolizing this design is the fact that trainees in self-paced instruction are likely to be distributed at various stages of instruction. Thus, the representation shown denotes a subset of trainees (possibly as few as two) who have reached a particular stage of instruction. Data would be accumulated over several such small-sample experiments before final analyses were carried out.

Finally, it is important to emphasize that all measures (observations) of interest do not have to be made at the same time. For example, one might be interested in several consequences associated with an experimental treatment, such as trainee attitudes toward the course, trainee achievement, and instructors' assessments of the method. One set of measures might be obtained as a time series, another might be posttest only, and the third variable be measured in a pretest-posttest pattern.

Data Processing and Analysis

Long periods of data collection and many different measures of interest can create problems of data storage, retrieval, and processing, even when the number of cases is small. When sample sizes become fairly large, data handling can become a formidable problem. A complete evaluation plan will include consideration of how data will be recorded and organized to facilitate retrieval for use in analyses.

Table 4 suggests a skeleton layout for recording information about participants in an experimental trial of alternative instructional treatments. Although arranged as though one anticipated paper-and-pencil analyses with a desk-top calculator, it should be easy to see that the categories translate into instructions for encoding data for machine storage.

If one has the benefit of computer support, the skeleton layout illustrated in Table 4 may approximate the structure of an analysis file created by extracting data from several different special purpose files or an omnibus data bank. If viewed as the structure for an analysis file, note that many of the illustrative variables defined by column heads or listed in the footnotes may be irrelevant for any particular planned analysis. For example, if no use is planned for Armed Services Vocational Aptitude Battery (ASVAB) scores in an analysis, then that entry is extraneous.

Table 4 is simply one of many possible layouts for a subject-by-variable raw data matrix. Each row identifies a participant in an instructional treatment experiment. Each column provides some piece of

information necessary for grouping data for analyses or computing measures of association between pairs of entries. The manner in which the data are grouped for analysis will depend upon the analytic model chosen. The array (whether viewed as a spread sheet for paper-and-pencil tabulation or as fields into which data may be encoded for machine tabulation and computation) assumes a multiple regression analysis as the most likely statistical model. Reference back to Section I ("Planned Experiments with Alternative Instructional Treatments") and to Appendixes C and D may be useful at this point.

Table 4

SKELETON LAYOUT FOR CODING OR RECORDING TRAINEE CHARACTERISTICS AND PERFORMANCE DATA

TRAINEE		PRE-ENROLLMENT BACKGROUND 1/	SCHOOL HISTORY 2/	PRE-INSTRUCTION APTITUDE PROFILE 3/	DATA FROM EXPERIMENT											
					TREATMENT GROUP 4/						PERFORMANCE DATA 5/					
					GROUP 1		...	GROUP n			TEST 1 CODE ____		...	TEST n CODE ____		
NAME	ID				CODE	ASSIGN DATE	...	CODE	ASSIGN DATE		ITEM SCORE	DATE	...	ITEM SCORE	DATE	

1/ Examples. Date of birth, Sex, Race, Ethnic group, Marital status, Home of record, Highest year of school completed, Social Security number, Branch of service, ASVAB profile, AFSC/MOS, Most recent duty assignment, Pay grade, etc.

2/ Examples. Date of enrollment for each course in which enrolled at school or facility.

3/ Examples. Reading skills (vocabulary, comprehension, speed), Mathematics skills, Media preferences (projected visual, printed, aural), Attitudes, Interests, Traits (anxiety, curiosity, etc.), etc.

4/ Examples. Identifies treatment group membership (Treatment A group, Treatment B group, etc.) and date of assignment to group. Multiple entries provide for changes in group membership in multi-stage experiments.

5/ Examples. Tests and other measures (1, 2, ... n) coded for identity of measure. Cell entries are raw item scores and dates of test administrations.

HA-423582-16

REFERENCES

1. Air Force. Interservice procedures for instructional systems development. AFM 50-2.

2. American Psychological Association, American Education Research Association, National Council on Measurement in Education (Joint Committtee). Standards for educational and psychological tests. Washington, D.C.: American Psychological Association, Inc., 1974.

3. Anderson, S. B., Ball, S., Murphy, R. T., & Associates. Encyclopedia of educational evaluation: Concepts and techniques for evaluating education and training programs. San Francisco: Jossey-Bass, 1975.

4. Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally & Company, 1963.

5. Cooley, W. W., & Lohnes, P. R. Evaluation research in education. New York: Irvington Publishers, Inc., 1976.

6. Cronbach, L. J. Essentials of psychological testing (3rd ed.). New York: Harper & Row, 1970.

7. Cronbach, L. J., & Snow, R. E. Aptitudes and instructional methods: A handbook for research on interactions. New York: Irvington Publishers, Inc., 1977.

8. Dixon, W. J., & Brown, M. B. (Eds.). BMDP-81: Biomedical Computer Programs, P-Series. Berkeley, CA: University of California Press, 1981.

9. Kerlinger, F. N., & Pedhazur, E. J. Multiple regression in behavioral research. New York: Holt, Rinehart and Winston, Inc., 1973.

10. Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, MA: Addison-Wesley Publishing Company, 1968.

11. McNemar, Q. Psychological statistics (4th ed.). New York: John Wiley and Sons, Inc., 1969.

12. Mood, A. M. Partitioning variance in multiple regression analyses as a tool for developing learning models. American Educational Research Journal, 1971, 8, 191-202..

13. Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., & Bent, D. H. SPSS: Statistical package for the social sciences (2nd ed.). New York: McGraw-Hill Book Company, 1975.

14. Nunnally, J. C. Psychometric theory. New York: McGraw-Hill Book Company, 1967.

15. Scriven, M. The methodology of evaluation. In Perspectives of curriculum evaluation (AERA Monograph Series on Curriculum Evaluation, No. 1). Chicago: Rand McNally, 1967, 39-82.

16. Snedecor, G. W., & Cochran, W. G. Statistical methods (7th ed.). Ames, IA: The Iowa State University Press, 1980.

BIBLIOGRAPHY

The sources listed below cover a range of topics concerning approaches and procedures for research and evaluation in instruction. Each source also includes additional references on general and specific topics relevant to design and evaluation of instructional methods.

1. Bloom, B. S., Hastings, J. T., & Madaus, G. F. Handbook on formative and summative evaluation of student learning. New York: McGraw-Hill Book Company, 1971.

2. Delaney, H. D., & Maxwell, S. E. The use of analysis of covariance in tests of attribute-by-treatment interactions. Journal of Educational Statistics, Summer 1980, 5(2), 191-207.

3. Ghiselli, E. E. The validity of occupational aptitude tests. New York: John Wiley & Sons, Inc., 1966.

4. Hays, W. L. Statistics for psychologists (3rd ed.). New York: Holt, Reinhart and Winston, 1981.

5. Lewis-Beck, M. S. Applied regression: An introduction. In J. L. Sullivan (Ed.), Series: quantitative applications in the social sciences (No. 07-022). Beverly Hills, CA: Sage Publications, 1980.

6. Morris, L. L. (Ed.). Program evaluation kit. Beverly Hills, CA: Sage Publications, Inc., 1978 (Books in kit: Evaluator's handbook, How to deal with goals and objectives, How to design a program evaluation, How to measure program implementation, How to measure achievement, How to measure attitudes, How to calculate statistics, How to present an evaluation report).

7. Snow, R. E. Learning and individual differences. In L. S. Shulman (Ed.), Review of research in education (Vol. 4). Itasca, IL: F. E. Peacock Publishers, Inc., 1976.

Appendix A

A THREE-TRIAL EXAMPLE FOR EVALUATING CANDIDATE TEST ITEMS THROUGH USE DURING ACTUAL SELF-PACED INSTRUCTION

TABLE OF CONTENTS

LIST OF TABLES

A THREE-TRIAL EXAMPLE FOR EVALUATING CANDIDATE TEST ITEMS THROUGH USE DURING ACTUAL SELF-PACED INSTRUCTION

Introduction

This appendix presents step-by-step detail for evaluating candidate test items within the context of actual self-paced instruction. The model for evaluation would fit a situation where a new course was being tried out, or where several lessons within an existing course were being revised, and there was need to develop criterion tests for new or revised segments of instruction. The model also would fit an established course in which student achievement measures were being revised but course content was unchanged.

The data used in the example are hypothetical. For convenience of illustration, the detailed example has 16 students and eight candidate test items. In practice, both the number of students and the number of candidate items should be greater.

Following presentation of the three-trial example, suggestions are presented for extending the evaluation model to more than three trials.

Scheduling Testing Points

The plan for trials of candidate items must assure variability in student performance. In self-paced instruction where progress from one instructional segment to the next is determined by successful performance on a criterion test, the easiest way to assure that trainees are distributed at differing stages of progress is to time the measurements to occur at approximately equally-spaced time intervals.

The first trial administration should precede the beginning of instruction; that is, be a pretest composed of all candidate items before any instruction in the first lesson. The second and third trials should be about equally spaced in time and scheduled so that the third or final trial occurs before the first student who completes the course has been reassigned and departed from the school.

With a new course, estimating these times involves judgment. If about one-third of all students have completed the mid-lesson before the scheduled time for the second trial has been reached, try to conduct the second trial at about that point. If about one-fourth of all students have completed the final lesson before the scheduled time for the third trial, try to conduct the third trial at that time to minimize losses of students from the trial group.

Weighting Item Response Patterns

On a simple two-valued pass-or-fail scoring system, there are eight possible response patterns for each item that could occur over a three-trial sequence. Using "P" to denote pass and "F" to denote fail, these eight possibilities over three successive trials are as follows: F-F-F, F-P-F, P-F-F, F-F-P, F-P-P, P-F-P, P-P-F, and P-P-P. Some of these possible pat- terns conform to an acceptable performance and others do not. The eight patterns may be rearranged and grouped into at least three classes according to the desirability of their response patterns as shown in Table A-1.

Table A-1

ITEM RESPONSE PATTERNS ACCORDING TO DESIRED PERFORMANCE FOR A THREE-TRIAL SEQUENCE

Response Pattern					
Trial 1	Trial 2	Trial 3	Response Scoring	Qualitative Rating	Quantitative Rating
F	F	P	112	OK	3
F	P	P	122	OK	3
P	P	P	222	Acceptable	2
F	F	F	111	Acceptable	2
P	F	P	212	Acceptable	2
F	P	F	121	Not OK	1
P	F	F	211	Not OK	1
P	P	F	221	Not OK	1

The above ratings reflect two considerations: improvement over time and consistency of response. A rating scale with more than three points might be used. For example, the P-F-P pattern rated as "2" is less consistent than either P-P-P or F-F-F, also rated as "2." In practical applications, however -- especially in early or intermediate stages of item development and refinement -- the three-point scale provides acceptable weights for items to guide decisions about items to retain, items to revise, and items to reject.

A procedure for using response pattern weights in evaluating items is shown below in some detail. With small numbers of students available on which to try out training performance test items, the procedure is simple enough to be applied with paper and pencil. Obviously, a procedure such as the one illustrated in the following paragraphs also could be analyzed with even less effort through the use of a computer.

An Illustrative Layout for Item Response Data

Table A-2 shows an illustrative layout with hypothetical data for a three-step trial of eight items with 16 students. (As a reminder, the numbers of items and students have been kept small to simplify the illustration; in practice, probably many more than eight items would be tried out and more than 16 students would serve as trial subjects.)

Table A-2

HYPOTHETICAL DATA FOR A THREE-STEP TRIAL OF EIGHT ITEMS WITH 16 STUDENTS

Response Scoring	Weight	Pattern Frequency by Item: Item a	Item b	Item c	Item d	Item e	Item f	Item g	Item h
112	3	4	6	3	3	2	2	2	5
122	3	2	1	3	1	1	3	4	1
222	2	3	-	1	2	5	2	-	4
111	2	-	2	4	1	4	1	1	1
212	2	1	3	-	1	1	3	3	1
121	1	2	2	1	5	1	3	2	1
211	1	3	1	3	1	2	2	3	-
221	1	1	1	1	2	-	-	1	3
	Total	16	16	16	16	16	16	16	16
Weighted Means for Items by Trial	Trial 1	2.750	2.688	2.438	2.312	2.875	2.750	2.625	2.938
	Trial 2	2.938	2.562	2.875	2.625	2.875	3.000	2.938	3.062
	Trial 3	3.625	3.875	3.312	2.875	3.312	3.562	3.500	3.875

The weighted means by trial for each item are obtained by multiplying the response for the trial by the weight for the response pattern by the frequency of the response pattern and dividing the product by the total number; i.e., [(Weight)(Trial Response)(Frequency)]/N. This computation is shown in detail below for Trial 1 on Item a.

Response Pattern	Pattern Weight	Item a: Trial 1 Response Score	Response Frequency	Computation (Wt.)(Resp.)(Freq.)
FFP	3	1	4	3 x 1 x 4 = 12
FPP	3	1	2	3 x 1 x 2 = 6
PPP	2	2	3	2 x 2 x 3 = 12
FFF	2	1	0	2 x 1 x 0 = 0
PFP	2	2	1	2 x 2 x 1 = 4
FPF	1	1	2	1 x 1 x 2 = 2
PFF	1	2	3	1 x 2 x 3 = 6
PPF	1	2	1	1 x 2 x 1 = 2
			16	44

The weighted mean for Trial 1 on Item a is 44/16 = 2.750.

The weighted means by trial for each item shown in Table A-2 indicate that seven of the eight items reflect the desired quality of improvement over time, as illustrated in Section II, Table 3. Item b displays a slightly lower weighted mean at Trial 2 than at Trial 1, thus departing from the dsired form. Item e has identical weighted means at both Trials 1 and 2, but shows an increased mean at Trial 3 and is therefore acceptable.

Statistical Analyses of Item Response Data

Three statistical analyses of the item data are appropriate at this point to help determine which items to retain in their present form, which ones to revise, and which ones to reject. More extensive discussion and examples of these analyses are shown in Appendix B. The types of analyses and the main principles underlying them are as follows:

1. Determine the homogeneity of the items as a set and search for relatively homogeneous subsets within the larger set. Measurement theory assumes that test item scores are combined by summing (with or without differential weighting of items) into a total score. Unless items are more or less homogeneous -- that is, unless they "go together" statistically as well as according to their manifest content -- it makes little sense to think of the items as comprising a single test. The fact that an item in a set of candidate items does not prove to be homogeneous with other items in the set is not sufficient for rejecting the item for possible use. Nonhomogeneity of an item, however, is sufficient reason for rejecting the item as one of a set of items called a test. An item that stands apart from companion items belongs in another group with new companions.

   Evidence of homogeneity requires data from actual trials of items with persons like those for whom the test is intended. The extent to which items "go together" implies correlational analysis. With a large pool of items, the approach that requires the least computation is to correlate each item with a total score made up from the set of items (or from reduced sets that exclude each item in turn). Items that correlate with the total score also correlate with one another. Examining patterns of intercorrelations among items is somewhat more sensitive. This approach is computationally more tedious, however, for it requires at least $(n^2 - n)/2$ computations instead of n computations (where n = number of items).

   The principle, then, is homogeneity; the means for estimating it is through part-whole or item vs. item correlational analyses.

2. Estimate the reliability of the set of items and of subsets of the most homogeneous items. Once sets of items that are reasonably homogeneous have been identified, estimate the length of test required to achieve a desired level of reliability.

   Reliability is closely related to homogeneity. One approach to estimating reliability from a single test administration makes use of the ratio of the sum of item variances to the variance in total scores. The other ingredient in that index of reliability is the number of items.

   Errorless test measurement -- perfectly reliable measurement -- assumes a test of infinite length. The tolerable upper limit on test length is far short of infinity. In addition, there is a practical limit to the number of different items that can be conceived for most behavioral domains. Even so, the most direct approach for increasing measurement reliability is to increase the number of items. A procedure for estimating the number of items required to increase the reliability of a short test to a higher, desired level is discussed and illustrated in Appendix B.

3. Examine the performance characteristics of candidate items. The likelihood of passing an item should be related positively to time in training or some independent index of proficiency in the subject matter of the test. In criterion-referenced measurement in a self-paced instructional environment, items that do not conform to a monotonic form (i.e., "OK" or "acceptable" response patterns from Table A-1) usually should be rejected.

Estimating Item Homogeneity and Test Reliability

Table A-3 shows the results of an analysis to estimate the homogeneity of the eight candidate items, a-h, as summarized earlier in Table A-2. Item means by trial are repeated in Table A-3 from Table A-2. The standard deviations of item responses by trial reflects the spread of scores around the mean; in statistical jargon, the standard deviation is the square root of the variance. The figures of greatest interest in Table A-3 are the correlation coefficients that indicate how well the items "go together." (See Appendix C for discussion and examples of correlation.)

Two sets of correlation coefficients are shown in Table A-3. The larger coefficients, identified in the table as "item vs. total score," are based on the hypothetical responses of 16 persons to each item in each of the three trials. Each response was weighted according to the procedure described earlier. Each person's item scores were then paired with that person's total score made up of the sum of the responses to all items. These first correlations are inflated, however, since part of the total score variation is controlled by the item itself.

Table A-3

ITEM VS. TOTAL SCORE CORRELATIONS AS A BASIS FOR ESTIMATING HOMOGENEITY AMONG ITEMS

Test Item	Trial	Weighted Mean	Standard Deviation	Correlation Coefficients: Item vs. Total Score	Correlation Coefficients: Item vs. Total of Remaining Items
a	1st	2.750	0.9682	.4611	.1381
	2nd	2.938	1.5194	.5159	.1709
	3rd	3.625	2.1759	.4497	.1041
b	1st	2.688	0.9164	.4248	.1157
	2nd	2.562	1.0588	.0018	-.2464
	3rd	3.875	2.1176	.3956	.0532
c	1st	2.438	0.7043	.2590	.0140
	2nd	2.875	1.6910	.3373	-.0757
	3rd	3.312	2.2000	.3349	-.0056
d	1st	2.312	1.1022	.4174	.0367
	2nd	2.625	1.1659	.4568	.1930
	3rd	3.875	2.1176	.3521	.0058
e	1st	2.875	0.9922	.7848	.5781
	2nd	2.875	1.3170	.4790	.1800
	3rd	3.312	1.7219	.4951	.2383
f	1st	2.750	1.0897	.1299	-.2443
	2nd	3.000	1.6583	.0911	-.2970
	3rd	3.562	2.0300	.0684	-.2558
g	1st	2.625	0.9270	.3583	.0377
	2nd	2.938	1.8530	.4067	-.0451
	3rd	3.500	2.2079	.4379	.0848
h	1st	2.938	0.8992	.1544	-.1588
	2nd	3.062	1.0879	.6545	.4598
	3rd	3.875	1.9961	.4043	.0845
Total All Items	1st	21.375	2.8696		
	2nd	22.875	4.1363		
	3rd	27.938	6.1081		

Note: N = 16 for all computations. Basic response pattern data (16 test-takers, 8 items, 3 trials at equally spaced times) are all hypothetical; computations are for ilustration only.

The right-hand column of Table A-3, labelled "item vs. total of remaining items," corrects for this inflation by correlating each item response with a sum composed of all the remaining items. As can be seen, the adjusted part-whole correlations are substantially smaller and several actually shift from positive to negative in direction.

It is evident from the right-hand column of Table A-3 that only about half of the eight candidate items show signs of belonging to a more or less homogeneous set. Items e and a look encouraging whereas Item f rather obviously belongs somewhere else since it is negatively related to what the other items are measuring. For illustration, assume that Items a, e, g, and h were singled out as the items to retain and refine. How does the homogeneity of that subset compare to the homogeneity of the remaining items? Table A-4 provides an answer.

When the total of eight items is split into subsets based on how the individual items correlate with the total score (see Table A-3), the meaning of homogeneity among items becomes clear, and the effect of item homogeneity upon test reliability is evident.

The total of eight items is not at all promising as a homogeneous test, as shown in the right-hand column of Table A-4. As can be seen, the sum of variances of responses to individual items is virtually as great as total score variance. This leads to very low estimates of test reliability. The eight items would need to be increased to perhaps several hundred before a test with such low homogeneity could meet desired reliability standards.

When Items a, e, g, and h are singled out as a "subtest," the situation becomes more promising. As a group, these four items display total score variance that exceeds item variance by a sufficient margin for that subset to be considered reasonably homogeneous. Homogeneous, in this case, is a relative matter -- certainly they are far more homogeneous than the full set of eight items. Using the Spearman-Brown formula to estimate reliability of a lengthened test made up of similar items, it appears that a test of from 30 to 50 similar items would have a reliability in the range of about .70 to about .80. (See Appendix B for discussion of the Spearman-Brown formula and a detailed example of its application.) Reliabilities of .70 to .80 are not high by commercial test standards but such a level is quite satisfactory for progress measures. In short, these four items provide a foundation on which to construct some additional measures to enhance measurement reliability.

The residual of items -- b, c, d, and f -- would be rejected on statistical grounds as part of a test that also included Items a, e, g, and h. Item variance for this subset is high relative to total score variance for the subset. The negative reliability coefficients shown in Table A-4 are very unlikely to be encountered in actual test development, particularly with achievement measures. In this example, the negative coefficients are a consequence of the quasi-random pattern of item responses by persons that were generated to provide data for an illustration. The resulting example is dramatic but not at all common.

Table A-4

ESTIMATES OF TEST RELIABILITY BASED ON ITEM AND TOTAL SCORE VARIANCE

Statistic	Subtest of Items a, e, g, h	Subtest of Items b, c, d, f	Total of 8 Items
Weighted mean			
Trial 1	11.1875	10.1875	21.3750
Trial 2	11.8125	11.0625	22.8750
Trial 3	14.3125	13.6250	27.9375
Total Score Variance			
Trial 1	4.5273	3.0273	8.2344
Trial 2	11.4023	4.6836	17.1094
Trial 3	19.3398	12.6094	37.3086
Sum of individual item variances			
Trial 1	3.5898	3.7383	7.3281
Trial 2	8.6602	8.0898	16.7500
Trial 3	16.5586	17.9297	34.4883
Coefficient alpha*			
Trial 1	.2761	-.3131**	.1258
Trial 2	.3207	-.9697**	.0240
Trial 3	.1917	-.5626**	.0864

*Coefficient alpha is an index of test reliability based on item homogeneity. See Appendix B for comment. The equation is as follows:

$$r_{kk} = (k/k-1)\,[1 - (\Sigma s_i^2 / s_t^2)]$$

where r_{kk} = reliability of test of $\underline{k}$ items

k = number of items

Σs_i^2 = sum of item variances

s_t^2 = variance of total scores

**See text for discussion of negative coefficients.

Table A-3 (as well as Table A-2) showed that all but one of the eight candidate test items displayed an increasing monotonic form; i.e., lowest mean scores occurred on Trial 1, higher means on Trial 2, and highest mean scores on Trial 3. By that standard, these items might be judged as adequate. Unless an item is homogeneous with other items in a set, however, the item should not be thought of as part of a test for which a total score is obtained by summing across items.

The preceding point has been made before but is important enough to warrant repeating. A single item may be an appropriate measure of one of the effects of instruction even though it does not prove to be homogeneous with other potential (or candidate) test items. When such deviant items are used, it amounts to creating a test with a single item. Sampling theory, as well as common sense, reminds us that the average of several measures of something is a better estimate of a "true" value than is a single measure. The implication, then, is that other items must be developed that are homogeneous with the original one if one wishes to increase the reliability of measurement.

Contrasting Item Response Profiles over Trials

When all or nearly all candidate test items display an increasing monotonic form of responses obtained at spaced intervals coincident with instruction and there is need to select the best of an apparently good lot, the "distance" (D) statistic may prove useful. The distance between profiles for entities a and b (e.g., students a and b, items a and b) for any number of variables ($\bar{k}$) is the square root of the following expression:

$$D^2_{ab} = (X_{a1} - X_{b1})^2 + (X_{a2} - X_{b2})^2 + \ . \ . \ . \ + (X_{ak} - X_{bk})^2$$

For example, assume that an item performance characteristic defined by the following pattern provided a desired idealized model against which to contrast obtained responses. ("Pass" = 2 and "fail" = 1 in the following example, just as in the preceding examples.)

Response Pattern	Pattern Weight	Response by Trial			Relative Response Frequency
		Trial 1	Trial 2	Trial 3	
112	3	1	1	2	3
122	3	1	2	2	2
222	2	2	2	2	1
111	2	1	1	1	2

Computation of Trial Means: [(Wt.)(Resp.)(Freq.)] / N

Trial 1: (3x1x3)+(3x1x2)+(2x2x1)+(2x1x2) / 8 = 2.875

Trial 2: (3x1x3)+(3x2x2)+(2x2x1)+(2x1x2) / 8 = 3.625

Trial 3: (3x2x3)+(3x2x2)+(2x2x1)+(2x1x2) / 8 = 4.750

None of the response patterns shown earlier in Table A-3 corresponds exactly with this particular idealized profile. The "D" statistic permits the similarity between each obtained profile and the idealized one to be expressed quantitatively. In addition, the idealized profile can be compared to a "chance" profile, where "chance" (random) is defined by each of the eight possible patterns having the same frequency. For example, with eight possible response patterns scored and weighted as before, the "chance" profile would be as follows:

Response Pattern	Pattern Weight	Trial Means Trial 1	Trial 2	Trial 3
112	3	2.625	2.750	3.125
122	3			
222	2			
111	2			
212	2			
121	1			
211	1			
221	1			

Illustrative Computation: Trial 1

(1+1)3 + (2+1+2)2 + (1+2+2)1 = 21

21 / 8 = 2.625

Table A-5 summarizes the result of a set of "D" computations for Items a through h and for the above "chance" profile.

The D^2 and D measures shown in Table A-5 were computed from the formula given previously. For example, the D^2 value for Item a was computed as follows before rounding:

$$D^2 = (2.875-2.75)^2 + (3.625-2.9375)^2 + (4.75-3.625)^2 = 1.7539062 \cong 1.754$$

The D-statistic is a convenient and easily calculated way to express similarity between multi-point profiles. Usually one or another of two approaches is followed when using the D-statistic to compare or contrast profiles. One approach, such as summarized in Table A-5, is to compute a set of distances relative to a chosen reference profile so that all distance measures are from a common reference. This allows a ranking or other expression of similarity between each profile and the reference profile.

Table A-5

DISTANCE (D) MEASURES FOR PROFILES OF ITEMS a THROUGH h AND FOR A CHANCE PROFILE CONTRASTED TO AN IDEALIZED PROFILE

Profile	Weighted Means by Trial			Distance Relative to Ideal Profile		Simi- larity Rank	Better (+) or Worse (-) Than Chance
	Trial 1	Trial 2	Trial 3	D^2	D		
Idealized	2.875	3.625	4.750	--	--	--	
Chance	2.625	2.750	3.125	3.469	1.862	--	
Item a	2.750	2.938	3.625	1.754	1.324	2nd	+
Item b	2.688	2.562	3.875	1.930	1.389	4th	+
Item c	2.438	2.875	3.312	2.820	1.679	7th	+
Item d	2.312	2.625	2.875	4.832	2.198	8th	-
Item e	2.875	2.875	3.312	2.629	1.621	6th	+
Item f	2.750	3.000	3.562	1.816	1.348	3rd	+
Item g	2.625	2.938	3.500	2.098	1.448	5th	+
Item h	2.938	3.062	3.875	1.086	1.042	1st	+

Another approach is to compute the D-values between all pairs. This matrix, of course, may include one or more arbitrary reference profiles in addition to other profiles of interest. Had that approach been followed for Table A-5, a 10 x 10 matrix with 100 D-values would be displayed. (Only 45 different computations would have been needed, for the diagonal of the matrix will display "0" and the distance from A to B equals the distance from B to A; thus, (10 x 9)/2 = the number of different values.)

Inter-Item Correlations Compared to Item Response Profiles

It also should be emphasized that the D-statistic is not a substitute for a correlation between individual values that have been averaged to provide profile points. For example, the profile points for item responses are based on the mean of 16 responses at each of three trials for eight different items. Correlations between item responses will reflect variability due to individual differences among persons responding. Thus, a pair of items may not correlate well with one another but be essentially identical in their profile form based on mean values that "wash out" individual differences among persons.

An example -- extreme to emphasize the point -- is provided by Item a vs. Item f. The distance (D) between these two profiles was the smallest of all the pairs of items, thus indicating great similarity between the two profiles based on trial mean scores. (Inter-item profile differences (D-scores) ranged from a low of .088 between Items a and f to a high of 1.258 between Items d and h; the smaller the D-score, the greater the similarity between profiles.) Note, however, the crosstabulations of responses by the 16 imaginary test-takers to these two items, as shown in Table A-6 below.

Table A-6

FREQUENCY CROSS-TABULATIONS OF WEIGHTED SCORE RESPONSES TO ITEM a AND ITEM f FOR EACH TRIAL

Item a Scores	Trial 1 Item f Scores					Trial 2 Item f Scores							Trial 3 Item f Scores						
	1	2	3	4	Sum	1	2	3	4	5	6	Sum	1	2	3	4	5	6	Sum
6					NA*			1	1			2	3	1		1		1	6
5					NA*							0							0
4			2	2	4		2	1				3				2		2	4
3	1	3	1	1	6	2	2					4							0
2	2		1	1	4		1		1		2	4							0
1	_	_	1	1	2	_	2	_	_	_	1	3	2	_	_	2	_	2	6
Sum	3	3	5	5	16	2	7	2	2	0	3	16	5	1	0	5	0	5	16

Trial	Item	Statistic Mean	Std. Dev.	Correlation (a vs. f)	"D" Based on Trial Means
1	a	2.7500	0.9682		
	f	2.7500	1.0987	.1185	
2	a	2.9375	1.5194		
	f	3.0000	1.6583	-.1736	
3	a	3.6250	2.1759		
	f	3.5625	2.0300	-.2069	.0884

*NA = Not applicable; 4 = maximum possible score on Trial 1.

The correlations between items by trial are low; for Trials 2 and 3, they also are negative in direction, thus indicating a weak tendency for a high score on one item to be associated with a low score on the other item. Simple visual examination of the cross-tabulations, even without confirmation from the correlation coefficients as descriptive statistics, affirms that the responses to Item a are not associated strongly with responses to Item f. The distance (D) statistic, however, shows that the the item profiles are very similar.

The correlation coefficients and the D-statistic are complementary rather than contradictory. The correlations indicate that the two items are not homogeneous if thought of as companion items in the same test. Presumably, the items are measuring different things and thus would not be combined intentionally. The D-statistic shows that both items conform approximately to a desired item operating form -- that is, that item performance improves with instruction. On that standard, they both may be "good" items but scores from that pair should not be added together.

Finally, the cross-tabulations in Table A-6 underscore the need for repeating a reminder -- the number of imaginary test-takers was made small (16) to simplify the presentation of examples of useful evaluation analyses. Ideally, the number of persons involved in empirical trials to evaluate test items would be substantially larger than 16. A generally accepted guideline is that the number of cases (persons) should be at least five times the number of items being evaluated to reduce sampling error. The examples, therefore, should not be taken as models of appropriate sample sizes.

Extending the Evaluation Method to More than Three Trials

Several trials provide better, more dependable, indicators of performance than do only a few trials, just as a test with many items is a more reliable indicator of ability than a test with only a few items. The foregoing illustration of ways to evaluate candidate test items with three empirical trials of actual students during a period of instruction can be extended readily to four or more trials. In most practical training situations, more than five trials would be difficult, if not impossible, to arrange. The following paragraphs extend the logic of a three-trial sequence to ones of four trials and five trials.

If four trial administrations of candidate test items can be scheduled at approximately equally spaced time intervals, the number of possible response patterns will be double the number that were possible for three trials. (Since the scoring of performance is binary -- F or P -- the rule is 2^n where n denotes the number of trials or opportunities. Thus, $2^3 = 8$, $2^4 = 16$, $2^5 = 32$, and so on.)

Weights for Response Patterns from Four Trials and Five Trials

Table A-7 shows the 16 possible response patterns in a four-trial sequence, arranged according to the judged quality of the pattern for criterion-referenced measurement. Table A-8 shows the 32 possible atterns and quality ratings for a five-trial sequence.

As was the case with the three-trial response patterns shown earlier in Table A-1, the ratings in both Tables 10 and 11 are judgmental. Finer-grained scales certainly could be used, and some might judge the ordinal position of some response patterns to require adjustment. The weighted means for chance responses by trial provide logical validation of the weights, however. As can be seen, a graphic plot of the chance means for both the four-trial and five-trial sequences would reveal curves of the desired consistent improvement form.

A summary of response data obtained from four empirical trials may be arranged in the same manner as shown earlier in Table A-2. For example:

Response Pattern	Weight	Pattern Frequency by Item					
		Item a	Item b	Item c	Item d	Item d	. . . . .
1112	5						
1122	5						
1222	5						
2222	4						
.	.						
.	.						
.	.						
2211	1	____	____	____	____	____	. . . . .
	Total						
Weighted Means for Items by Trial	Trial 1						
	Trial 2						
	Trial 3						
	Trial 4						

The data summary layout for a five-trial sequence would be similar except for an increase in possible response patterns to 32 and a different set of weights.

Weighted means by trial for each item may be computed as described earlier for hypothetical data in Table A-2. When computed, these values define profile points from which item characteristic curves may be plotted or otherwise compared.

Table A-7

ITEM RESPONSE PATTERNS ACCORDING TO DESIRED CHARACTERISTICS FOR A FOUR-TRIAL SEQUENCE

Response Pattern				Qualitative Rating	Quantitative Rating
Trial 1	Trial 2	Trial 3	Trial 4		
1	1	1	2	Very good	5
1	1	2	2	Very good	5
1	2	2	2	Very good	5
2	2	2	2	Good	4
1	1	1	1	Good	4
2	1	2	2	Good	4
1	2	1	2	Fair	3
2	2	1	2	Fair	3
1	1	2	1	Fair	3
2	1	2	1	Poor	2
1	2	2	1	Poor	2
2	2	2	1	Poor	2
2	1	1	2	Very poor	1
2	1	1	1	Very poor	1
1	2	1	1	Very poor	1
2	2	1	1	Very poor	1

Trial	Weighted Means for Chance Patterns
1	4.0000
2	4.1875
3	4.5625
4	4.7500

Note: 1 = Fail, 2 = Pass.

Table A-8

ITEM RESPONSE PATTERNS ACCORDING TO DESIRED CHARACTERISTICS FOR A FIVE-TRIAL SEQUENCE

Response Pattern					Quantitative Rating
Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	
1	1	1	1	2	8
1	1	1	2	2	8
1	1	2	2	2	8
1	2	2	2	2	8
1	2	1	2	2	7
2	1	2	2	2	7
2	2	2	2	2	7
2	2	1	2	2	6
2	1	1	2	2	5
1	1	2	1	2	5
1	2	2	1	2	5
2	2	2	1	2	5
1	1	1	1	1	4
1	1	1	2	1	4
1	1	2	2	1	4
1	2	2	2	1	4
2	1	2	1	2	4
2	2	2	2	1	4
1	2	1	2	1	3
2	1	2	2	1	3
2	2	1	2	1	3
1	2	1	1	2	2
2	1	1	1	2	2
2	2	1	1	2	2
2	1	1	2	1	1
1	1	2	1	1	1
1	2	2	1	1	1
2	1	2	1	1	1
1	2	1	1	1	1
2	1	1	1	1	1
2	2	1	1	1	1
2	2	2	1	1	1

Trial	Weighted Means for Chance Patterns
1	5.59375
2	5.81250
3	6.06250
4	6.50000
5	6.71875

Recap of Main Steps in the Method

To review briefly, three analyses in the following order are appropriate in evaluating candidate test items over repeated trials:

1. Correlational analyses to estimate homogeneity among items, first considering the items as a whole set and then as selected subsets.

2. Reliability estimates to affirm homogeneity among items and to estimate the number of additional items needed to satisfy desired reliability standards.

3. Item characteristic analysis to assure that items adhere to the desired monotonic or "consistent improvement" form.

Each of these analyses was described and illustrated in the example of the three-trial evaluation. Appendix B contains further detail regarding correlational analyses and procedures for estimating effects of changes in test length.

The total number of cases needed for acceptably dependable estimates of item characteristics depends on both practical limitations and how one defines "acceptably dependable." If only a few students are available for the trials of candidate test items, then one has no recourse but to rationalize "acceptably dependable" as whatever one can obtain with the small number of available students. If the number of available students is not so small, then a randomly drawn sample of 30 or so students should yield fairly stable estimates of item characteristics.

A reminder -- the scheduling of item trials should be defined by clock time rather than by student progress through a self-paced instructional unit. The primary purpose at this stage of development is to evaluate items so that "good" tests can be constructed for evaluating student progress with later groups of students. By scheduling item trials to occur at approximately equally spaced time intervals prior to, during, and shortly after instruction is completed for most students, variability in student performance is guaranteed.

Appendix B

EVALUATING CANDIDATE TEST ITEMS AND DEVELOPING TESTS THROUGH TRIALS WITH CROSS-SECTIONAL SAMPLES

TABLE OF CONTENTS

LIST OF TABLES

EVALUATING CANDIDATE TEST ITEMS AND DEVELOPING TESTS THROUGH TRIALS WITH CROSS-SECTIONAL SAMPLES

Introduction

This appendix describes an approach for evaluating candidate test items with samples of persons representing the range of competence or proficiency in the performance area to which the items apply. The approach may be used when it is not feasible to evaluate test items through multiple trials during actual instruction, as described in Appendix A. Thus, the approach can be viewed as a complement to the multiple-trial approach or as an alternative to that approach when multiple trials during actual instruction are not possible.

This appendix also contains discussion of problems of test reliability, concepts of test validity, and the relation of empirical validity to reliability. Certain statistical procedures appropriate to those issues also are illustrated. These portions of Appendix B may be applicable to the multiple-trial approach to evaluating test items as described in Appendix A.

Format for Recording Response Data

Table B-1 illustrates a layout for organizing data from trials of candidate items with a cross-sectional sample or samples of persons like those for whom the items are intended. The table is arranged to show actual responses from persons in the sample tested. Since no weighting of responses is necessary with the cross-sectional sample approach (and hence no need to multiply a response code by a weight), it is computationally more convenient to use "1" to denote a correct response and "0" to denote an incorrect one. With the "1" and "0" response coding, the mean score for an item is the proportion of correct responses.

All data in this example are hypothetical.

Sample Sizes for Persons and Items

No hard-and-fast rules govern the minimum number of candidate items to be subjected to trial or the minimum number of persons on whom the items should be tried. Two generally accepted rules, however, are that first, the pool of candidate items should contain at least twice the number one wishes to retain for final use, and second, the number of persons should be about five times as large as the number of items to be tried.

Table B-1

ILLUSTRATIVE LAYOUT FOR RECORDING AND SUMMARIZING RESPONSE DATA FROM TRIALS OF CANDIDATE TEST ITEMS WITH A STRATIFIED SAMPLE

Proficiency Stratum	Student ID	Item 1		Item 2		. . .	Item n	
		1st Admin.	2nd Admin.	1st Admin.	2nd Admin.	. . .	1st Admin.	2nd Admin.
High	101	0	1	1	0		1	1
	102	0	1	1	1		1	0
	.							
	.							
	.							
Sub-Total Correct		—	—	—	—	. . .	—	—
Moderate	201	1	1	1	0		0	0
	202	0	0	0	1		1	0
	.							
	.							
	.							
Sub-total Correct		—	—	—	—	. . .	—	—
Low	301	0	0	1	0		0	1
	302	1	0	0	0		1	0
	.							
	.							
	.							
Sub-total Correct		—	—	—	—	. . .	—	—
Total Correct		—	—	—	—	. . .	—	—

Note: 1 = correct response, 0 = incorrect response.

The practical value of the first rule is obvious; some candidate items will not perform well regardless of effort in initial design and it is more efficient to discard poor items than to have to design new ones and arrange more trials. In Table B-1, Item n is the last item in a pool of size n. Thus, if the goal is a 10-item test, n should equal 20 or more.

The basis for the guideline about the total number of persons in the sample may be less obvious. Simply put, the rule of "5 times items" provides some protection against taking advantage of chance (i.e., capitalizing on sampling errors) during item analysis. Some authorities consider five to be too small a ratio and argue for the ratio of cases to items to be 10:1 or more (see reference 14). The possibility of faulty inference is always present with sample data, but in the end, decisions involve balancing what is feasible against the risks one is willing to take. Tests constructed from item analyses based on small samples are subject to much more fluctuation in their behavior from time to time than are tests constructed from item analyses based on large samples.

Number of Data Collection Points

The layout shown in Table B-1 is for a test-retest design. Such a design has some advantages over a single test administration design. Some of the advantages are (a) averaging two responses to get a better estimate of "true" performence, (b) correlating first and second administrations to provide an additional index of measurement reliability, and (c) holding out a random half of both first and second administration data to verify results from the other random half. Such advantages aside, it is not essential that test item trials involve a test-retest design. The layout of Table B-1 is equally appropriate to once-only measurement.

Finally, Table B-1 implies that all data are obtained at the same time. It is not essential that this be so -- data may be accumulated over time and compiled periodically for analyses. It is important that the conditions for testing be as nearly alike as possible from one administration to another which argues against a lengthy period of data collection. It is entirely reasonable, however, to build up sufficient numbers of appropriately selected cases from several independently conducted test administrations.

Summary Item Response Data for Procedural Example

To provide data for an extended example and discussion of procedures for evaluating criterion-referenced test items through the use of cross-sectional samples, a set of hypothetical data was constructed for eight candidate items and 50 test subjects. Table B-2 shows a summary of the resulting frequency distributions by item (1-8), response (0 or 1), and proficiency stratum (low, moderate, high). (The original 8 x 50 table with random entries of 0 or 1 is not shown, but it was organized in a format similar to Table B-1).

Table B-2

FREQUENCY DISTRIBUTION OF RESPONSES TO EIGHT CANDIDATE TEST ITEMS BY 50 PERSONS ACCORDING TO ESTIMATED PROFICIENCY IN UNDERLYING ATTRIBUTE

Candidate Item	Response*	Proficiency Stratum			Total		Response vs. Proficiency (Product moment correlation)
		Low	Moderate	High	No.	Prop.	
1	1	6	6	14	26	.52	
	0	12	10	2	24	.48	
	Total	18	16	16	50	1.00	.4394
2	1	6	8	10	24	.48	
	0	12	8	6	26	.52	
	Total	18	16	16	50	1.00	.2411
3	1	8	12	12	32	.64	
	0	10	4	4	18	.36	
	Total	18	16	16	50	1.00	.2671
4	1	10	8	12	30	.60	
	0	8	8	4	20	.40	
	Total	18	16	16	50	1.00	.1586
5	1	6	6	12	24	.48	
	0	12	10	4	26	.52	
	Total	18	16	16	50	1.00	.3383
6	1	0	12	8	20	.40	
	0	18	4	8	30	.60	
	Total	18	16	16	50	1.00	.4362
7	1	10	8	8	26	.52	
	0	8	8	8	24	.48	
	Total	18	16	16	50	1.00	-.0467
8	1	2	6	8	16	.32	
	0	16	10	8	34	.68	
	Total	18	16	16	50	1.00	.3456

*1 = corect response, 0 = incorrect response.

Visual inspection of the response frequency distributions for each item by proficiency stratum is sufficient to indicate that responses to most items are a function of proficiency level. (This result was guaranteed by the sampling rules followed in generating the data.) The only exception to the generalization is Item 7 which shows an essentially "flat" distribution.

The right-hand column of Table B-2 is an aid to visual inspection. The values shown in that column are the product-moment correlations for each 2 x 3 cross-tabulation of response (scored 0 or 1) by proficiency stratum (scored 0, 1, or 2). As the coefficients indicate, the response pattern for Item 7 is clearly unrelated to proficiency and the pattern for Item 4 reflects only a weak relationship.

Given only the data shown in Table B-2, one could be tempted to celebrate a modest victory in test item design -- at least six of eight candidate items show response patterns that support the assumption of relationship between test item performance and proficiency.

Response Profiles for Candidate Items

Table B-3 converts the frequencies shown in Table B-2 to proportions of people in each stratum who responded correctly. To make the illustration more comparable to Appendix A, a distance (D) statistic also is shown as an index of similarity between each profile and an arbitrarily defined "ideal" response profile. (See Table A-5 and accompanying text in Appendix A for an example and discussion of the D-statistic.)

Table B-3

RESPONSE PROFILES FOR EIGHT CANDIDATE TEST ITEMS AND "D" MEASURES FOR EACH PROFILE COMPARED TO AN IDEALIZED PROFILE

Profile Points by Proficiency Strata	Proportion of Correct Responses								
	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Idealized Profile
High	.875	.625	.750	.750	.750	.500	.500	.500	.875
Moderate	.375	.500	.750	.500	.375	.750	.500	.375	.500
Low	.333	.333	.444	.556	.333	.000	.556	.111	.125
Mean Proportion Correct for Item	.520	.480	.640	.600	.480	.400	.520	.320	.500
Distance (D) Relative to Ideal Profile	.243	.325	.424	.448	.273	.468	.571	.396	

The correlation coefficients shown in Table B-2 and the D-statistics shown in Table B-3 are related but imperfectly so. (Recall that the higher the correlation, the stronger the relationship whereas the smaller the D-statistic, the greater the profile similarity.) When the two sets of indices are compared to one snother, the largest discrepancy in rank-preference is for Item 6. Item 6 discriminates very well at the lower end of the proficiency scale but not so well at the upper end, thus yielding a poor correspondence to the idealized profile. Even with that discrepancy, which could be anticipated from the unusual response pattern for Item 6, the two summary statistics -- the correlations and the D-statistics -- tend to reinforce one another.

Combining Items to Construct a Test

The problems inherent in attempting to measure human performance argue for redundancy in measurement; several indices, considered together, are more likely to provide a dependable estimate of "true" performance than is a single index. For this reason, measurement through testing usually means combining several individual items, each of which contributes to an additive total test score. Items may be weighted equally or differentially, but all tests assume that item scores will sum to a total test score.

Table B-4 shows how well the eight candidate items, considered as an eight-item test, were related to the independently defined scale of "proficiency" that the test items purport to measure. As can be seen in Table B-4, the sums of item scores do correlate quite well with the proficiency scale. This relationship can be seen in the frequency distribution, the differences in mean scores by proficiency stratum, and the correlation coefficient that expresses the relationship between the two scales.

Part-Whole Correlations to Estimate Item Homogeneity

Given the evidence from Tables B-2, B-3, and B-4, one is tempted to conclude that the eight items can be combined into a single test. Such a conclusion, however, would be premature. Before combining subsets of items and calling the combination a test of some attribute, the homogeneity of the items in the combination must be examined. Unless the items that make up a test correlate positively with one another, the items as a group are not homogeneous.

If a test is not homogeneous, then more than one attribute is being measured. When item scores are summed to create a test score, it is assumed that each item adds something to the others. Unless items share a common attribute or factor, it make no sense to sum the item scores. The resulting sum would not have a meaningful interpretation. Criterion-referenced measurement absolutely requires that an obtained score can be interpreted with reference to a standard of mastery.

Table B-4

COMBINED DISTRIBUTION OF TOTAL NUMBER OF CORRECT RESPONSES TO EIGHT CANDIDATE TEST ITEMS BY 50 PERSONS ACCORDING TO THEIR ESTIMATED PROFICIENCY IN THE UNDERLYING ATTRIBUTE

Total Number of Items Answered Correctly	Proficiency Stratum			
	Low	Moderate	High	Total
8	-	-	-	0
7	-	-	-	0
6	-	-	6	6
5	-	6	8	14
4	4	6	2	12
3	8	4	-	12
2	4	-	-	4
1	-	-	-	0
0	2	-	-	2
Total	18	16	16	50
Mean Correct	2.667	4.125	5.250	3.960
Std. Deviation	1.155	0.781	0.651	1.399

Note: Correlation between proficiency stratum (scored 0, 1, 2) and number correct = .7621.

With a large pool of candidate items and many test subjects, an economical approach to estimating homogeneity is to calculate the correlation of each item to total score. Items that correlate most highly with total scores are the "best" items -- they share more of the variance attributable to the common factor among the items and they add more to the reliability of the test. Thus, with a large pool of items, a simple procedure is to rank items from high to low according to the magnitude of their part-whole correlation coefficients and select items in blocks from the top down until reliability objectives have been met. Given reasonable homogeneity in the set, the number of items

needed will depend largely on the reliability one wishes for the test. (How to estimate reliability will be discussed shortly.)

The example used in this report is limited to eight items and 50 cases -- scarcely a "large pool of candidate items and many test subjects." Because the number of items is small, the correlation between an item and total score will be inflated since each item also is part of the total and therefore is being correlated with itself as well as with all other items. To correct for this, each item is correlated with the test score from the other items.

Table B-5 shows the results of such a procedure for the eight candidate items. Adjusted part-whole correlations are shown as the bottom row of the table. All the correlations are low. Most significant, however, is the fact that five of the eight coefficients are negative. Obviously, the set of eight items is not a homogeneous one.*

If the adjusted part-whole correlations shown in Table B-5 were taken as the only index of homogeneity, then Items 5, 4, and 2 would be selected as a relatively homogeneous set, and the remaining items would be discarded. Three items are not likely to make a reliable test, however, so the only recourse would be to develop new items against the model of the few that appear to define a homogeneous set and repeat the trials of test items with a new set of candidates. (Recall the rule that the pool of candidate items should contain at least twice the number desired for the final test.)

Item Intercorrelations to Estimate Item Homogeneity

The illustrative set of eight candidate items is small enough to permit an easy example of a more refined approach to the search for homogeneous sets of items. The part-whole correlation approach is a substitute for examining the patterns of intercorrelations among items. With only eight items in the set, the matrix of intercorrelations is small (i.e., (8x7)/2 = 28) and examining the patterns is informative.

---

*The original item score matrix on which the illustration is based -- 50 cases by eight items -- was constructed by drawing odd numbers (odd = 1) and even numbers (even = 0) from a random number table. Sampling ratios differed slightly by proficiency stratum to assure an overall pattern similar to that shown in Table B-4. No effort was made to assure intercorrelations among items. Considering the manner in which the score matrix was generated, therefore, it is not surprising that the adjusted part-whole correlations center near zero.

Table B-5

ITEM vs. TOTAL RESPONSE DISTRIBUTIONS BY 50 PERSONS ON EIGHT CANDIDATE TEST ITEMS

8-Item Total		Frequency by Response to Item According to Total Score*															
		1		2		3		4		5		6		7		8	
Score	Freq.	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
8	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
7	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
6	6	-	6	2	4	2	4	2	4	-	6	2	4	-	6	4	2
5	14	8	6	4	10	4	10	-	14	6	8	8	6	6	8	6	8
4	12	4	8	6	6	2	10	6	6	6	6	6	6	8	4	10	2
3	12	8	4	8	4	6	6	8	4	8	4	8	4	6	6	8	4
2	4	2	2	4	-	2	2	2	2	4	-	4	-	2	2	4	-
1	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0	2	2	-	2	-	2	-	2	-	2	-	2	-	2	-	2	-
	Total	24	26	26	24	18	32	20	30	26	24	30	20	24	26	34	16
Mean Percent Total Items Correct		44	55	42	57	43	53	40	56	41	58	45	56	44	55	46	56
Percent Passing Item		52		48		64		60		48		40		52		32	
Correlation: Item vs. Sum of Remaining Items		-.043		.078		-.069		.104		.145		-.037		-.043		-.071	

*1 = correct response, 0 = incorrect response.

Table B-6 shows the intercorrelations between all pairs of the eight items. The column and row headings in the matrix are arranged to highlight the presence of three clusters of items:

1. One cluster of fairly homogeneous items is composed of Items 2, 4, 5, and 7. The adjusted part-whole correlations shown earlier in Table B-5 identified only Items 2, 4, and 5 as apparently homogeneous. Item 7 belongs with this set statistically for it correlates positively with each of Items 2, 4, and 5 and correlates negatively with all the remaining items. (Item 7 might be rejected on other grounds, however. Recall from Table B-2 that it did not discriminate by proficiency strata.)

2. A fairly strong two-item cluster consists of Items 1 and 3 (the bottom right corner of the matrix in Table B-6). Item 3 correlates positively only with Item 1, and Item 1 correlates more strongly with Item 3 than with any other item.

3. A third two-item cluster composed of Items 6 and 8 is outlined in the center of Table B-6. This is a weaker cluster than the other two, but a legitimate one nevertheless.

Table B-6

CORRELATIONS AMONG EIGHT CANDIDATE TEST ITEMS

	Candidate Test Items							
Items	4	5	2	7	8	6	3	1
4	--	294	131	196	035	-167	-102	-131
5	294	--	199	122	-144	033	-113	-038
2	131	199	--	122	-144	033	-113	-038
7	196	122	122	--	-028	-196	-053	-282
8	035	-144	-144	-028	--	140	-021	-028
6	-167	033	033	-196	140	--	-068	131
3	-102	-113	-113	-053	-021	-068	--	280
1	-131	-038	-038	-282	-028	131	280	--

Note: (1) Decimals omitted from coefficients.
(2) Product-moment correlation coefficients computed from the formula for the fourfold point or phi coefficient.

Table B-5 demonstrated that the eight candidate items were not homogeneous. Although performance on the items (with the apparent exception of Item 7) correlated positively with the global quality of "proficiency" (see Table B-2), this global quality is not well defined by a single set of measures.

Having discovered that the eight items as a set were not homogeneous, the intercorrelations among items shown in Table B-6 helped identify three fairly homogeneous subsets of items. These three subsets can now be looked at more closely to help guide further test development. One serious problem in criterion-referenced measurement now can be avoided -- that of defining the mastery criterion for a segment of instruction as some fraction of a set of test items without first establishing that the set of test items is homogeneous.

Determining Test Reliability from Item Homogeneity

Frequent reference has been made in the text to the notion of measurement reliability. Reliability means that the measurements are repeatable within a tolerable range of fluctuation. If two appropriately sized samples of people, drawn randomly from the same population, were to take the same test under similar conditions, the results will be very similar if the test is reliable.

Three basic approaches are available for estimating the reliability of tests:

1. A test-retest procedure in which results at one time are correlated with results at another to provide a coefficient of stability.

2. A parallel-form procedure in which two closely comparable versions of a test are administered at a common time and correlated with one another to provide a coefficient of equivalence.

3. An internal-consistency procedure which provides a good approximation of the parallel or equivalent form procedure and also yields a coefficient of equivalence.

It can be shown that the reliability of a sample of test items is determined by the number of items and the average correlation among items. Without attempting to prove this assertion, the following equations can be used to estimate the reliability of a test of any length based on the internal structure of the test. Equation (1) is referred to as "coefficient alpha" and is the more general form. Equation (2) is a special case of Equation (1) for tests composed entirely of dichotomous items; Equation (2) is referred to as "KR-20" (for Kuder-Richardson Formula 20).

$$(1)\quad r_{kk} = \frac{k}{k-1}\left(1 - \frac{\Sigma\, s_i^2}{s_y^2}\right)$$

$$(2)\quad r_{kk} = \frac{k}{k-1}\left(1 - \frac{\Sigma\, p_i q_i}{s_y^2}\right)$$

Definitions: r_{kk} = reliability of test of k items

k = number of items

s_i^2 = variance of item i

s_y^2 = variance of scores on the total test

p_i = proportion passing (scoring "1") on Item i

$q_i = 1 - p_i$ (proportion not passing)

Σ = sigma, standing for the operation "the sum of." Precise notation would be

$$\sum_{i=1}^{k}$$

meaning "the sum of k values beginning with $i=1$ and ending with $i=k$."

Applying either equation to test data is easy, especially for tests that are scored dichotomously (e.g., pass = 1, not pass = 0) so that Equation (2) applies. References 10 or 14 show the computation of total score variance. The procedure also is illustrated below with an example from the eight candidate items used in the illustration preceding this point.

Table B-6 -- the matrix of item correletions -- indicated three clusters of relatively homogeneous items. From the full set of item scores by person (for brevity, not included in this report but summarized as Table B-2), the following frequency distribution can be tabulated to show total scores for 50 persons on a test composed of the largest cluster, Items 2, 4, 5, and 7:

Total Score (X)	Frequency (f)	X x f	X^2 x f
4	8	32	128
3	10	30	90
2	16	32	64
1	10	10	10
0	6	0	0
Total	50	104	292

$$\text{Mean} = 104/50 = 2.08 = 52\%$$

$$\text{Variance} = (292/50) - (104/50)^2 = 1.5136$$

The proportion passing and not passing Items 2, 4, 5, and 7 can be read directly from Table B-2. Thus, $\Sigma pq = (.48 \times .52) + (.60 \times .40) + (.48 \times .52) + (.52 \times .48) = 0.9888$.

Values can now be substituted in Equation (2) to estimate the reliability of a four-item test composed of Items 2, 4, 5, and 7:

$$r_{kk} = (4/3)[1 - (.9888/1.5136)] = (4/3)(1 - .6533) = .4623$$

Table B-7 shows the results of similar computations for three subsets of the eight candidate test items and for the entire set.

Table B-7 illustrates rather dramatically how reliability can be increased by grouping items into more or less homogeneous sets. Considered as a whole, the eight candidate test items are a catchall collection; the overall reliability coefficient of .0187 affirms their heterogeneity.

Lest there be any doubt that the eight candidate items should be separated into separate clusters, consider the correlations between scores on item clusters in the following summary:

Clusters Composed of Items	Clusters Composed of Items 2, 4, 5, 7	1, 3	6, 8
2, 4, 5, 7	--	-.2207	-.1549
1, 3	-.2207	--	.0085
6, 8	-.1549	.0085	--

Table B-7

RELIABILITY ESTIMATES FOR THREE SUBSETS AND THE WHOLE SET OF EIGHT CANDIDATE TEST ITEMS

	Response Frequency Distributions			
Total Score	Items 2, 4, 5, 7	Items 1, 3	Items 6, 8	All Items (1-8)
8	-	-	-	-
7	-	-	-	-
6	-	-	-	6
5	-	-	-	14
4	8	-	-	12
3	10	-	-	12
2	16	20	8	4
1	10	18	20	-
0	6	12	22	2
Total	50	50	50	50
Mean	2.08 (52%)	1.16 (58%)	0.72 (36%)	3.96 (49.5%)
Variance	1.5136	0.6144	0.5216	1.9584
$\sum pq$	0.9888	0.4800	0.4576	1.9264
(KR-20) r_{kk}	.4623	.4375	.2454	.0187

If a similarly heterogeneous batch of items were used as a criterion test for a segment of instruction to which some general rule was applied such as "X% correct defines acceptable mastery," diagnostic interpretation of the total score would be impossible. For a total score of a test to make sense, the items that comprise that test must be reasonably homogeneous. If the items are reasonably homogeneous, and there also are enough items, then the measure defined by number or percent correct also will be acceptably reliable.

Reliability does not guarantee validity, but reliability is a necessary condition for validity.

Test Validity and Its Relation to Reliability

The preceding analyses, applied to provide concrete examples of ways to evaluate and refine test items and to construct tests, have shown that homogeneity among items results in more reliable measures. The analyses also have demonstrated an approach for sorting a collection of heterogenous items into more homogeneous subsets. Because

these analyses were performed on a limited amount of data, the apparent result is three very brief "tests" -- one with four items and two with two items each -- that are made up of relatively homogeneous items but are much too short as they stand to provide acceptably reliable measurement. Before examining ways to decide how much longer the tests should be, it is instructive to consider the relationship between reliability and validity of measurement since, in the end, validity of measurement is the ultimate concern.

In the most general sense, a measuring instrument is valid to the extent that it does what it is intended to do. Validity is absolutely specific to purpose and application. The concept has meaning only with reference to the purpose of the measurement to which the concept applies.

Types of Validity

Convention recognizes three categories of validity: content validity, criterion-related validity, and construct validity:

1. Content validity requires that the behaviors demonstrated in testing be a representative sample of the behaviors that define the objectives of a program or program element, such as a unit or course of instruction. In the context of instruction, content validity often is called "curricular validity."

2. Criterion-related validity expresses the extent to which scores on a measure relate empirically to scores on an external criterion. For example, when scores on a paper-and-pencil test about steps in a trouble shooting routine are correlated with the time required or errors committed in actually performing a specified trouble shooting routine, the correlation coefficient expresses the criterion-related validity of the paper-and-pencil test for that routine. When a test is given and external criterion performance also is measured at about the same time, the relationship between the two measures is referred to as the test's concurrent validity. When the criterion performance is more remote in time, such as "success on the job" in relation to "success in training" (as measured by an examination score or total time to the instructional criterion), the relationship is called predictive validity.

3. Construct validity refers to the degree to which test scores allow inferences about underlying qualities or traits. For example, claims of construct validity for a test of "anxiety" would require evidence that persons scoring in one direction on the test were more likely to display both physiological and psychological indicators of apprehensiveness than were persons scoring in the other direction on the test. Construct validity usually is estimated from patterns of

relationship; that is, scores on the measure in question should be related to other scores on theoretically relevant measures (convergent evidence) and also should not be related to other scores on theoretically unrelated measures (discriminant evidence).

Much more extended discussion of the concept of validity will be found in references 2, 3, 6, 10, and 14.

In training directed toward developing knowledge, skills, and attitudes appropriate to effective performance on a job, the classes of measurement validity of greatest interest and importance are content validity and predictive validity.

1. Is the content of each test a full and fair representation of the substance of instruction that the test purports to measure?

2. Does test performance following a segment of instruction effectively identify persons who are prepared to undertake the next segment of instruction?

3. Does the aggregate of performance on all tests throughout instruction effectively identify people who will perform satisfactorily on the job toward which the training is directed?

Limits on Empirical Validity

Although validity, in one or more forms, is the most critical quality of a test, the limits to empirical validity (such as concurrent and predictive validity) are determined by the reliability of the measures involved. It can be shown that the upper limit of a correlation coefficient between two variables is defined by:

$$r_{xy} = r_{tt} \sqrt{r_{xx}\, r_{yy}}$$

Where: r_{xy} = correlation between variables x and y

r_{tt} = true relationship between x and y

r_{xx} = reliability of predictor variable x

r_{yy} = reliability of criterion variable y

It follows from the preceding equation that the obtained correlation between a predictor variable and a criterion variable -- the predictive validity coefficient -- cannot exceed the square root of the least reliable measure. Expressed in the form:

$$r_{tt} = r_{xy} / \sqrt{r_{xx}\, r_{yy}} \quad ,$$

the equation is called the "correction for attenuation." The practical use of the equation is to help point the way toward improvement of prediction.

Strategies for Increasing Test Reliability

Consider again the example of the eight candidate test items. When we abandoned the candidate items to discuss some basic notions about validity and its relationship to measurement reliability, the following points had been established:

1. The eight items could be broken into three fairly homogeneous clusters.

2. Each cluster was negatively related or unrelated to the other clusters.

3. The reliabilities of the three clusters of items, when considered as tests, were low.

We now can ask and answer two practical questions:

1. What can be done to increase the reliability of the three brief tests (i.e., the three clusters of items from the set of eight candidate items)?

2. How much effort should be invested in attempting to increase measurement reliability?

Table B-8 expands an earlier tabulation and sets the stage for answers to the above questions. Table B-8 shows six correlation coefficients -- each item cluster with the overall "proficiency" classification and each item cluster with the other clusters.

Table B-8

CORRELATIONS OF ITEM CLUSTERS WITH THE PROFICIENCY SCALE AND WITH EACH OTHER

	Proficiency	Cluster A	Cluster B	Cluster C
Proficiency Categories (0, 1, 2)	--	.2795	.4436	.5191
Cluster A (Items 2, 4, 5, 7)	.2795	--	-.2207	-.1549
Cluster B (Items 1, 3)	.4436	-.2207	--	.0085
Cluster C (Items 6, 8)	.5191	-.1549	.0085	--

The correlations shown in Table B-8 are derived directly from the same set of hypothetical data summarized earlier in Table B-2. For example, the two-way frequency tabulation for the correlation between the proficiency classification and responses to Cluster B (items 1 and 3) is as follows:

Proficiency Category	Cluster B (Items 1, 3): Number of Items Correct 0	1	2	Total	
High (2)	2	2	12	16	
Moderate (1)	2	10	4	16	
Low (0)	8	6	4	18	
Total	12	18	20	50	$r_{xy} = .4436$

The intercorrelations in Table B-8 allow the computation of first-order and second-order partial correlations of interest, i.e., the correlation between two-variables with the influence of one or both of the remaining two "partialled out" or controlled. First-order and second-order partial correlations are shown below in Table B-9.

The first-order partial correlations in the upper portion of Table B-9 are the relationships between pairs of variables freed from the influence of a third variable. For example, the partial correlation between the proficiency scale and item Cluster A scores, freed from the influence of Cluster B, is .4317. This coefficient, symbolized $r_{pA.B}$, may be contrasted to the simple correlation of .2795 between proficiency and Cluster A shown earlier in Table B-8. The smaller simple correlation coefficient of .2795 reflects the negative relationship between Clusters A and B and between Clusters A and C which influence the relationship of Cluster A to proficiency. These inter-cluster relationships were shown earlier in Table B-8.

The first-order partial correlations supply the bases for computing the second-order partial correlations shown in the lower portion of Table B-9. The second-order partial correlations are relationships between each of the Clusters A, B, and C with the proficiency scale when the influence of the other two clusters has been removed. We will take second-order partial correlations as the best available bases for estimating "true" relationships of each cluster score (as a predictor variable) to proficiency (as the criterion variable). A "true" relationship, in measurement theory, implies a hypothetical measuring instrument of infinite length and therefore of perfect reliability. We will use the coefficient of .6459, symbolized by $r_{pA.BC}$, as an estimate of the "true" relationship between the proficiency scale and cluster A scores.

Table B-9

FIRST-ORDER AND SECOND-ORDER PARTIAL CORRELATIONS FOR PERFORMANCE CLASSIFICATION SCORES AND ITEM CLUSTER SCORES

Partial Correlations	Variables Correlated	Variables Controlled	Notation	Coefficient
First-Order	Proficiency v. Cluster A	Cluster B	$r_{PA.B}$	.4317
	Proficiency v. Cluster A	Cluster C	$r_{PA.C}$	.4262
	Proficiency v. Cluster B	Cluster A	$r_{PB.A}$	.5396
	Proficiency v. Cluster B	Cluster C	$r_{PB.C}$	.5139
	Proficiency v. Cluster C	Cluster A	$r_{PC.A}$	.5929
	Proficiency v. Cluster C	Cluster B	$r_{PC.B}$	.5751
	Cluster A v. Cluster B	Cluster C	$r_{AB.C}$	-.2220
	Cluster A v. Cluster C	Cluster B	$r_{AC.B}$	-.1569
	Cluster B v. Cluster C	Cluster A	$r_{BC.A}$	-.0267
Second-Order	Proficiency v. Cluster A	Cluster B, Cluster C	$r_{PA.BC}$	.6459
	Proficiency v. Cluster B	Cluster A, Cluster C	$r_{PB.AC}$	.6899
	Proficiency v. Cluster C	Cluster A, Cluster B	$r_{PC.AB}$	.7215

Note: Proficiency categories scored as follows: High = 2, Moderate = 1, Low = 0. Cluster A composed of Items 2, 4, 5, 7; Cluster B composed of Items 1, 3; Cluster C composed of Items 6, 8.

Setting Test Reliability Targets

The development of the example of steps in evaluating candidate test items and constructing homogeneous tests has now reached the point where a key test development strategy question can be asked and answered:

Q: Give r_{xy}, how reliable would the criterion (y) and the predictor (x) have to be in order to achieve the estimated true relationship (r_{tt})?

A: Define various realistic target values for the reliability of the criterion measure (r_{yy}). Substitute these values, along with estimates for an obtained correlation between predictor and criterion (r_{xy}) and for the true relationship between predictor and criterion (r_{tt}), in the equation for correction for attenuation. Solve the equation for reliability of the predictor measure (r_{xx}).

Table B-10 shows the results of such an exercise. The values shown in the body of Table B-10 are reliabilities (r_{xx}) of the predictor tests made up of items such as those in Clusters A, B, and C that would satisfy the following form of the "correction for attenuation" equation:

$$r_{xx} = r_{xy}^2 \; / \; r_{tt}^2 \; (r_{yy})$$

Table B-10

ESTIMATED RELIABILITY STANDARDS (r_{xx}) FOR TESTS COMPOSED OF ITEMS LIKE THOSE THAT DEFINE CLUSTERS A, B, AND C

Cluster	r_{tt}	r_{xy}	r_{yy} = .60	.70	.80
A	.65	.28	.309	.265	.232
(Items 2, 4, 5, 7)		.43	.729	.625	.547
B	.69	.44	.678	.581	.508
(Items 1, 3)		.54	N.A.*	.875	.766
C	.72	.52	.869	.745	.652
(Items 6, 8)		.58	N.A.*	.927	.811

*N.A. - Not applicable; > 1.0.

The values for r_{tt} in Table B-10 can be recognized as rounded versions of the second-order partial correlation coefficients from Table B-9. The smaller values for r_{xy} are rounded versions of simple correlations between predictor and criterion from Table B-8. The larger values for r_{xy} are rounded versions of average first-order partial correlations from Table B-9. These values for r_{xy} are merely rough guesses about what r_{xy} might be as a function of varying reliabilities. Values for r_{yy} are assumed achievable target values for the reliability of a criterion measure of proficiency.

Estimating Test Length Needed for Desired Reliability

Examination of Table B-10 suggests that predictor measure reliability (r_{xx}) of .80 would exceed most of the values in the table while roughly approximating the rest. With this decision, a second test development strategy question can be asked and answered:

Q: How many items must there be in a homogeneous test of the attribute measured by each cluster of candidate items for such a lengthened test to have a reliability of .80?

A: Use the Spearman-Brown formula for the reliability of a composite test having parallel components.

The general Spearman-Brown formula is usually expressed in the following form:

$$R_{KK} = \frac{Kr_{11}}{1 + (K-1)r_{11}}$$

where R_{KK} = Reliability of the lengthened (or shortened) composite test

K = Multiple of the number of test items in the original test to be lengthened (or shortened)

r_{11} = Reliability of the original test to be lengthened (or shortened)

In the above form, the formula is a handy one for the question, "What if we change the length of the test of n items with reliability r_{11} by a factor of K (i.e., add or subtract i items so $K = (n+i)/n$)?" The formula may be rearranged to solve directly for K, however, if one has an estimate of a target value for R_{KK}. In that form, the Spearman Brown formula becomes:

$$K = R_{KK}(1-r_{11}) \; / \; r_{11}(1-R_{KK})$$

Solving the Spearman-Brown formula for K, using values from previous computations or analyses for r_{11} and R_{KK}, produces the

findings shown in Table B-11. In the computations for Table B-11, the input values for r come from the bottom row of Table B-7 where they were identified as r_{KK} computed from the Kuder-Richardson Formula 20 (KR-20). The input values for R_{KK} come from decisions following examination of Table B-10 -- R_{KK} for all computations in Table B-11 equals .80, the target value decided upon for reliability of the predictor tests. (In the notation of the equation for correction of attenuation, the equivalent term was denoted as r_{xx}).

Table B-11

MINIMUM NUMBER OF HOMOGENEOUS ITEMS NEEDED IN LENGTHENED TESTS OF ATTRIBUTES MEASURED BY CLUSTERS A, B, AND C FOR EACH TEST TO HAVE A RELIABILITY OF .80 OR MORE

	Test Item Clusters		
	A	B	C
KR-20 reliability of original item set (r_{11})	.4623	.4375	.2454
Desired reliability of lengthened test (r_{KK})	.8000	.8000	.8000
Number of items in set to which r_{11} applies	4	2	2
Multiplier for number of items (K)	4.6524	5.1429	12.2999
Total items needed in lengthened test (K x no. original items, rounded upward)	19	11	25

Recap of Main Steps and Decisions in Test Development for Criterion-Referenced Measurement

The test development agenda is now clear. The implications of analyses to this point may be summarized as follows:

1. The original set of eight candidate test items appeared to differentiate reasonably well among levels of overall proficiency to which the test was directed. The individual items were not equally strong in their ability to differentiate among proficiency levels, as was shown in Table B-2. However, when item scores were summed to a total score, the eight-item test looked reasonably good, as shown by Table B-4. Such an evaluation is not warranted, however, until homogeneity of the test items is examined. Without demonstrating homogeneity among the items, the test must be considered a catchall collection that cannot be useful for differential diagnosis of proficiency.

2. The first analysis of test item homogeneity was to correlate each item score with the total score -- literally, to correlate each item with the sum of scores of all the remaining

items. This analysis was shown in Table B-5 and demonstrated that the items were not homogeneous. Based only on the part vs. whole analysis, only three of the original eight items (items 2, 4, and 5) would be retained. If the pool of candidate test items had been substantially larger than eight, the part vs. whole analysis would be an efficient way to screen items since it involves only as many indices as there are items.

3. A more sensitive analysis of homogeneity among items involves correlating each item with all others and examining the pattern of intercorrelations for subsets of items that appear to go together. This requires more computations -- $(n^2-n)/2$, where n = the number of items -- but the computations are simple when items are scored as "pass" or "fail" (1 or 0) and pose no real burden if data are encoded for computerized computation. Intercorrelations among items were shown in Table B-6 and revealed three clusters of relatively homogeneous items: Cluster A (Items 2, 4, 5, 7), Cluster B (Items 1, 3), and Cluster C (Items 6, 8).

4. The discovery of three fairly homogeneous clusters of items challenged the assumption that "proficiency" is a unitary quality. Correlations of cluster scores with one another and with the proficiency score, as shown in Table B-8, suggest that "proficiency" may be made up of three components or factors. This impression was strengthened by the partial correlation analysis summarized in Table B-9. These analyses suggested that three tests are needed rather than one, or a test for each apparent component of proficiency.

5. Two important principles of measurement were asserted without complete proof but can be proved: (a) measurement reliability is a joint function of the homogeneity of items that comprise the measure and the number of items in the measure, and (b) the upper limit of empirical validity is bound by the reliabilities of the measures involved. The question of how long a measure should be -- that is, how many items it should include -- depends largely on the standards of reliability and the limits of predictive validity that one wishes to achieve.

There are both practical and statistical limits to the previous assertion, of course. First, as the correction for attenuation demonstrates, the upper bound of predictive validity is a function of the reliabilities of measures of both predictor and criterion and also of the "true" relationship between them. Any "true" bivariate relationship, in measurement of human performance, almost certainly will be considerably less than 1.0 simply due to the number of different factors that influence performance. For example, for want of a better basis for inference, a coefficient of

.65 was posited as an estimate of the upper bound for a correlation between a test composed of items like Cluster A and an independent measure of the dimension of proficiency to which such a test applied.

As shown in Table B-11, the Cluster A-type test must be increased to at least 19 items if it is to reflect a reliability of .80. With a reliability of .80 for both predictor and criterion, one could anticipate a correlation of about .52. If the predictor test were increased to some 89 equally homogeneous items, a reliability of .95 would result. That increase in reliability, purchased at the price of an additional 70 test items beyond 19, might increase the obtained correlation between predictor and criterion from about .52 to about .57. The practical gain would not be worth the effort.

The test development agenda, then, calls for two complementary efforts: first, to construct additional items that are homogeneous with those represented by item Clusters A, B, and C from the original candidate items and second, to refine, and expand as necessary, independent measures of the criterion performance.

For internal validation of a training program, the overall criterion may be some weighted combination of several indicators, such as "measured time to subordinate criteria," "number of attempts to subordinate criteria," ratings by instructors of practical work, self-ratings by trainees of confidence in their mastery of the training content, and any other behavioral indicators that are accepted as differentiating among students.

For external validation of a training program, the criterion must be some weighted combination of indicators of how well a graduate of the training program performs on the job, because training, in the end, is effective only to the degree that it contributes to on-the-job competence.

Development of predictor measures is more straightforward than development of criterion measures, and is defined by the number of items needed in the predictor tests (as shown in Table B-11). Developing additional items that are homogeneous with those on which the analyses were based may prove difficult; in some areas, it may not be possible to create enough new items to meet the quantitative goals and also satisfy the requirement of homogeneity.

If either the quantitative requirement or the homogeneity requirement must be compromised, the quantitative goal is less important than the requirement for item homogeneity. Without homogeneity among the items that comprise a measure of performance, simply summing the scores to a total score will not make sense. Instead, homogeneous subsets should be treated as subscores in a criterion profile, and criterion performance should be defined in terms of the profile rather than

according to a summation across items. For example, if a criterion test were composed of three subtests similar to Clusters A, B, and C, the criterion performance might be defined as passing X% of items from Subtest A and Y% of items from Subtest B and Z% of items from Subtest C.

Without a differentiated definition of the criterion performance in training for a lesson, segment, unit, or block of instruction, attempts to devise differentiated treatments to best fit learner aptitudes is almost certainly doomed to failure or, at best, very inconsistent success.

Appendix C

REGRESSION ANALYSIS IN THE EVALUATION OF INSTRUCTIONAL TREATMENTS

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

REGRESSION ANALYSIS IN THE EVALUATION OF INSTRUCTIONAL TREATMENTS

Introduction

This report can only introduce some of the details that must be considered in performing a multiple regression analysis of data from an instructional treatment experiment designed to search for aptitude-by-treatment interactions. In this appendix, three topics basic to such an analysis are presented:

1. An overview of the idea of regression analysis, beginning with simple regression (two variables) and extending by analogy to multiple regression (three or more variables).

2. How to create variables by coding so that inter-group contrasts can be made from regression analyses.

3. How to represent aptitude-treatment interactions in a multiple regression model.

The appendix closes with references to packaged statistical programs for computers and to other instructional sources for details beyond the scope of this report.

Brief Overview of Regression Analysis

Understanding some basic ideas of regression analysis, if not the how-to-do-it details, can begin with the equation for a straight line. A straight line can be defined as the connection between two points. If these two points are places on a map drawn with perpendicular coordinates -- north-south and east-west -- the line from P to Q can be defined by the coordinates of one of the points and the compass direction from that point to the other point. If this familiar idea is sketched on paper, with a vertical axis (Y) and a horizontal axis (X), a picture like the solid-line portions of Figure C-1 might be drawn.

Now, instead of using Point P as part of the definition of the straight line connecting P and Q, that line could be extended to cross the Y axis, thus defining an intercept point on the Y axis. The slope of the line (rate of change in the Y-direction relative to the rate of change in the X-direction) is analogous to compass direction. And there we have it -- the equation for a straight line:

$$Y = a + bX$$

where: a = intercept constant
b = slope

If Y is related to X without error, then once we know X, we also know Y. For example, Fahrenheit temperature is an exact straight-line function of temperature in Celsius, and this relationship can be expressed in the form of a straight-line equation:

$$F = 32 + 1.8C$$

If this line were graphed, 32 would define the intercept on the F-axis and 1.8 would define the slope of the line (i.e., for every unit change in C there is a 1.8 change in F).

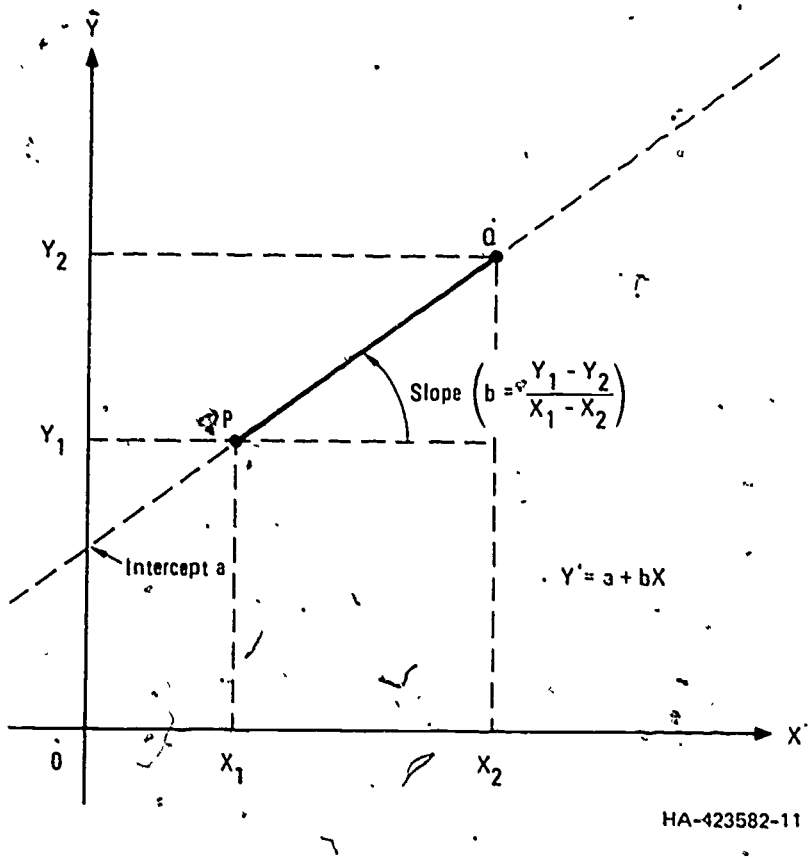

FIGURE C-1 REGRESSION EQUATION AS A STRAIGHT-LINE DEFINED BY Y-AXIS INTERCEPT AND SLOPE

A regression equation involving two variables (such as time to criterion and criterion score) takes the same form as the equation for a straight line. Thus:

$$Y' = a + bX$$

where: Y' = predicted Y score for dependent variable
a = intercept constant
b = regression coefficient or weight
X = score of an independent (predictor) variable

Y (or Y') is called the dependent variable because it is assumed that it varies "depending" on the value of X (the independent or predictor variable).

The line defined by $Y' = a + bX$ is that one which best fits a set of paired X and Y values. The "best fit" is defined by the line that minimizes the sum of the squares of the differences between the values of Y and the values predicted by the regression equation. Thus, the intercept constant (a) and the slope (b) are called "least squares" estimates.

The idea of the least squares best fit line is illustrated in Figure C-2. Figure C-2 is a graph or "scattergram" of 12 paired scores of X and Y values. (For example, X might be test scores at Time 1 and Y might be test scores at Time 2 for a class of 12 trainees.) It is evident to the eye that X and Y tend to vary together; in general, a low score on X means a low score on Y and a high score on X means a high score on Y. The statistic that expresses the relationship between X and Y is called a correlation coefficient, conventionally symbolized as r. In this case, r = .86.

Correlation coefficients can range in absolute value from .00 to 1.00, or from no relationship to perfect relationship. The sign of r (+ or -) denotes the direction of relationship. A positive sign means that high tends to go with high and low with low, whereas a negative sign means that high tends to go with low and low with high.

The regression equation shown in Figure C-2 ($Y' = 1.447 + .909X$) is a precise expression (within rounding error) of the regression line that best fits the scatter of points. Notice that the line in this example does not literally pass through any of the 12 points that comprise the scatter even though the line is the analytical "best fit." The difference between an estimated value of Y (denoted Y') and an actual value of Y for values of X is an error of estimate. (These differences, or errors, are often called "residuals" in regression analysis.)

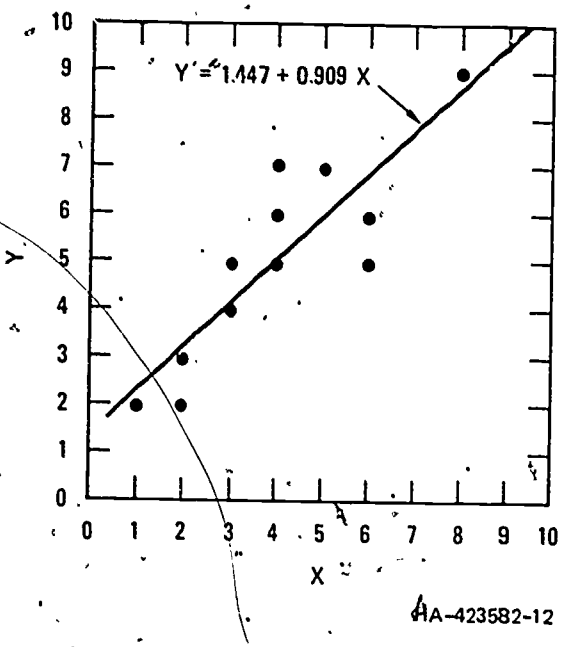

FIGURE C-2 REGRESSION LINE FOR SCATTER OF 12 POINTS DEFINED BY X AND Y VALUES

The average of the squared errors is the variance of the error distribution; in regression analysis, this term usually is called the "mean square error." The square root of the mean square error is the standard deviation of the distribution of differences between actual and predicted Y-values. This standard deviation is called standard error of estimate. It also is a key term involved in defining some other terms: (1) standard deviation of b, the regression coefficient, (2) sample standard deviation of estimated Y as an estimate of population mean, and (3) sample standard deviation of estimated Y as an estimate of a new point Y.

Table C-1 converts the information from Figure C-2 to numbers. The third column of Table C-1 shows the values of Y estimated for various values of X from the regression equation. These points fall on the line shown in Figure C-2. The fourth column shows the errors of estimate or the differences between actual and estimated values of Y. Summary statistics at the bottom of Table C-1 include the terms necessary to separate variability in the dependent variable, Y, into two parts: that variability accounted for by the regression line, and that variability which is unexplained (the sum of the squared errors or residuals). Note that these two parts add to the total variability.

In the bottom portion of Table C-1, the entries are referred to as sums of "squared deviations." Deviation refers to the difference between a given score in a distribution and the mean of that whole distribution of scores. It is conventional to denote a raw score in a

Table C-1

NUMERICAL VALUES TO SUPPLEMENT FIGURE C-2

Case	Actual Scores X	Actual Scores Y	Predicted Y Scores (Y')	Error Scores (Y - Y')
01	1	2	2.356	-.356
02	2	2	3.265	-1.265
03	2	3	3.265	-.265
04	3	4	4.174	-.174
05	3	5	4.174	.826
06	4	5	5.083	-.083
07	4	6	5.083	.917
08	4	7	5.083	1.917
09	5	7	5.992	1.008
10	6	5	6.902	-1.902
11	6	6	6.902	-.902
12	8	9	8.220	.280
Sum of Scores	48	61	61.000	0.000
Sum of Squared Scores	236	359	346.447	12.553
Sum of Cross products (XY)	(284)			
Mean of Scores	4.000	5.083	5.083	0.000
Standard Deviation of Scores	1.915	2.019	1.741	1.023

Regression (explained) sum of squared deviations = $346.447 - 61.000^2/12 = 36.364$

Error (unexplained) sum of squared deviations = $12.553 - 0.000^2/12 = 12.553$

Total sum of squared deviations = $359 - 61^2/12 = 48.917$

Correlation between X and Y = $r_{xy} = \dfrac{(12 \times 284) - (48 \times 61)}{\sqrt{12 \times 236 - 48^2}\ \sqrt{12 \times 359 - 61^2}} = .8622$

Explained variation / Total variation = $R^2 = 36.364 / 48.917 = .7434$

$r^2_{xy} = R^2 = .8622^2 = .7434$

Note: Values rounded following computations

distribution by a capital letter, such as X. (Both X and Y are used in Table C-1 to distinguish two variables.) The symbol, $\bar{X}$ (read as "X-bar"), is commonly used to denote the mean of the distribution of X-scores. A deviation score, then, is defined as $(X - \bar{X})$ and symbolized by the lower case, x. The phrase, "sum of squared deviations," could be symbolized as Σx^2 or as $\Sigma(X-\bar{X})^2$, where "Σ" means "add all terms."

The sum of squares of deviation scores (or "sum of squared deviations") in Table C-1 can be shown with simple algebra to have the following identity with raw scores:

$$\Sigma x^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$$

Raw score values are shown in the computations at the bottom of Table C-1.

The term, Σx^2, denotes the sum of squared deviations of scores about the mean or, broadly, the variability in the distribution. As noted earlier, the mean of that term -- $\Sigma x^2/N$ -- is commonly called the "mean square." It also may be called "variance."* The square root of the variance is the standard deviation of the distribution. Note in Table C-1 that the standard deviation of raw Y-scores = 2.019. This is the square root of the whole term, 48.917 (the "total sum of squared deviations") divided by 12 (the N or number of cases); i.e., $2.019 = \sqrt{48.917/12}$.

Based on the raw scores that comprise score distributions, the terms needed in most commonly used statistical calculations are (a) the number of cases, (b) the sums of scores in each single distribution, (c) the sums of squared scores in each single distribution, and (d) the sums of cross-products in each distribution of paired scores. Appendix E of this report contains a selected collection of commonly encountered statistical formulas.

A procedure for partitioning variance into additive parts is shown at the bottom of Table C-1. Also shown is the meaning of correlation in terms of the ratio of explained variation to total variation. Computations of the slope and intercept in the regression equation are described below.

Recall that the basic form of the regression equation is $Y' = a + bX$, where Y' = estimated Y-variable score, a = intercept, b = slope, and X = any X-variable score. The slope may be computed directly from

---

*This report bypasses the problem of specifying appropriate divisors for various mean squares. This is not to belittle its importance but to note that the general problem has too many specific answers to be treated here.

a combination of the basic sums. It is handier to compute the slope term (the regression coefficient) before computing the intercept, so that one can use the obtained slope term in the intercept computation.

(1) Slope = $b = \frac{\sum XY - \sum X \sum Y/N}{\sum X^2 - (\sum X)^2/N}$

(2) Intercept = $a = \sum Y/N - b\sum X/N$

Equations for bivariate case only

Using equation (1), the slope term for the data in Table C-1 is computed as follows:

$$b = \frac{284 - [(48)(61)/12]}{236 - (48^2/12)} = .90909\ldots = .909$$

Using equation (2), the intercept term for the data in Table C-1 becomes:

$$a = 61/12 - [(b)48/12] = 1.446969\ldots = 1.447$$

Thus, the completed equation for estimating the regression of Y on X is as shown in Figure C-2:

$$Y' = 1.447 + .909X$$

Before the above computational detour, the meaning of error of estimate was introduced and illustrated in Table C-1. In Table C-1, we also showed how total variability (i.e., the "sum of squares" or "sum of squared deviations") can be partitioned into a fraction explained by the regression and a fraction that is unexplained, and that this ratio = R^2. The unexplained fraction is variously referred to as "error" or "residual." The additive property of the sums of squares made up of explained and unexplained variability leads to need for a term to denote "error" in the generalized regression equation. In the simple two-variable form, this equation is usually written as:

$$Y = a + bX + e$$

where: Y = value of Y
X = value of X
a = intercept constant
b = slope or regression coefficient
e = error

The objective is to minimize error (i.e., to maximize explanation). It is precisely this objective that leads to such efforts as sharpening measurement to reduce measurement error, transforming scales to increase the straight-line nature of relationships between variables, and sifting predictor variables so that only relevant ones are included.

As soon as two or more predictor variables are employed in an effort to explain variability in a third variable (the dependent variable), the regression analysis becomes a multiple regression analysis.

The general multiple regression equation is a logical extension of the simple form. Multiple regression uses many variables to predict Y. This can be expressed as follows:

$$Y = a + b_1X_1 + b_2X_2 + \ldots + b_kX_k + e$$

In this form, the subscripts (1, 2, ... k) denote different X variables, each of which has an associated b or regression coefficient. As before, error is denoted by e. The regression coefficients -- the b values -- are weights. Thus, the objective is to find the best-weighted combination of X values to predict Y.

Instructors, subject matter experts, and other staff responsible for the development, conduct, and evaluation of technical training may encounter two kinds of problems with respect to multiple regression analyses -- interpreting the work of others and planning, conducting, and interpreting their own analyses. In both cases, knowing something about the meaning of terms in a regression equation is essential; these topics are treated briefly below. The more creative enterprise is to plan, conduct, and interpret one's own analyses. The closing portion of Section I of the report offers some prescriptive counsel regarding "do-it-yourself" regression analyses.

The Scale of Measurement

The discussion to this point has largely assumed that the variables in the regression equation are in terms of original measurements (i.e., test scores, time to complete, etc.). We call these "raw" scores. Since the results of a regression analysis are not changed by multiplying any variable by a constant, or by adding a constant, it is frequently more convenient to "standardize" the variables so that their average is zero (by adding a constant) and their standard deviation is one (by multiplying by a constant). Standard scores are commonly called z-scores.

When variables are standardized, computer printouts, for example, often designate the slope as a "beta weight," "b-weight," or "partial regression coefficient." Using raw scores, the slope is often designated "B-weight," or simply "B," or "raw score weight."

Comparisons across samples for a single variable probably are best made using unstandardized, rather than standardized, regression coefficients since the beta weight is so sensitive to variability in the distributions. However, the question of relative importance among several predictors in a regression equation can be approached only when the coefficients are standardized. It is only when all variables are

in standard (z-score) terms and the coefficients are expressed as beta weights that the relative importance of the predictors (in accounting for total variance in the dependent variable) can be estimated. Variables with larger beta weights, regardless of their sign, are more important.

How to Represent Categorical Variables in Regression Analysis

When multiple regression analysis is used to analyze experimental data -- for example, when one wishes to estimate the effects of alternative treatments on some dependent measure of performance -- then independent variables can be created to provide a way of quantitatively coding subjects according to the treatment they experienced. Furthermore, when one is interested in possible interactions between characteristics of subjects and the treatment they experienced, additional variables can be created to represent such interactions.

The following two sections discuss ways to deal with these two issues: (1) using categorical or nominal variables (e.g., sex or treatment group) along with continuous variables (e.g., test scores or performance ratings) in regression analysis, and (2) creating variables to represent interactions between treatments and personal characteristics or aptitudes.

Coding Categorical Variables

When a multiple regression approach is used to assess the effects of alternative instructional treatments on performance, the analyses must accommodate both categorical or nominal variables (e.g., instructional treatment, sex, race) and continuous variables (e.g., years of service, measures of prior performance, aptitude test scores, interest inventory scores).

Categorical variables, such as type of instructional treatment, can be used in a multiple regression analysis by representing them through what are called "dummy variables." A dummy variable is created by treating each category of a variable as separate. For example, one may wish to use "prior course experience" as one of the predictor variables in the evaluation of an entirely new course. Imagine that assignment to the new course is made from among persons who have previously completed any one of existing courses A, B, C, or D. The evaluation question concerns which prior course experience is the best predictor of performance in the new course. Dummy variables D1, D2, and D3 would be created to represent prior course experience as shown in Table C-2.

Table C-2

DUMMY VARIABLE SCORES FOR THE CATEGORICAL VARIABLE, "PRIOR COURSE EXPERIENCE"

Prior Course	Dummy Variables D1	D2	D3
A	1	0	0
B	0	1	0
C	0	0	1
D	0	0	0

Note in Table C-2 that the number of dummy variables needed is one less than the number of categories to be represented. As shown in Table C-2, "prior course D" is fully determined by the other three categories (i.e., zero on all three variables). In this example, "prior course D" becomes the "reference category." It is not excluded from the analysis; rather, it is the reference value against which the other variables are contrasted.

Dummy variables also can be created to represent the experimental variable of "instructional treatment." An approach exactly analogous to Table C-2 could be used if there were three or more instructional approaches to be contrasted. Again, the guiding rule is that when a categorical variable has C categories, use one less than C dummy variables to represent it. Thus, with three alternative treatments to contrast, two dummy variables would be needed and the third treatment would be the reference category.

Creating Variables to Represent Interactions

In analysis of variance (ANOVA) with two or more independent variables (as illustrated in the body of the report in Figure 5 and accompanying text), the analysis produces terms that represent interactions between the combinations of the independent variables. For example, in a two-way design such as shown in Figure 5, the independent variables were Treatment (X_1, with levels A and B) and AFSC (X_2 with levels 1 and 2). The ANOVA produces estimates of X_1 variance, of X_2 variance, and of variance due to interaction between X_1 and X_2, or X_1X_2. Simply stated, an interaction between two independent variables implies that lines connecting cell means, when plotted as in Figure 5, are not parallel.

With three independent variables (e.g., Q, R, S), ANOVA would yield (1) three main effect estimates, Q, R, and S, (2) three two-way interactions, QR, QS, and RS, and (3) one three-way interaction, QRS.

By analogy, the number of interaction terms can be seen to expand with the inclusion of each additional independent variable.

In multiple regression analysis, the analyst is seeking a regression equation that minimizes the error of estimate. If interaction is either suspected or expected, variables to represent the interaction must be created and included in the analysis. Symbolically, an interaction term for two independent variables would be shown as follows:

$$Y' = a + b_1X_1 + b_2X_2 + b_3X_1X_2$$

In the above expression, X_1X_2 is the product of variables X_1 and X_2 and b_3 is the regression coefficient associated with that created interaction variable.

In instructional treatment experimentation, it is customary to create interaction variables involving treatment (the experimental independent variable) and one or more of the individual differences variables or "aptitudes" descriptive of persons in the experiment.

Table C-3 illustrates procedures for creating aptitude-treatment interaction variables with dummy coding variables.

Table C-3

CREATING APTITUDE-TREATMENT INTERACTION VARIABLES
(Hypothetical Data)

			Dummy Coding	
Treatment Group	Dependent Var. Y	Aptitude Var. X_2	Treatment Code X_1	Aptitude Treatment Interaction X_1X_2
A	7	6	1	6
	8	7	1	7
	9	8	1	8
	10	9	1	9
B	8	6	0	0
	9	7	0	0
	7	8	0	0
	6	9	0	0

Sources for More Detailed Guidance

Computational procedures for multiple regression analyses involving several variables are sufficiently complicated to make it impossible to address them responsibly in this report. Packaged programs

for computerized solutions are widely available (see, for example, Reference 13, Statistical Package for the Social Sciences (SPSS) or Reference 8, BMDP Biomedical Computer Programs).

The following references contain computational examples: Kerlinger & Pedhazur (1973), McNemar (1969), Nunnally (1967), Snedecor & Cochran (1980). The Kerlinger and Pedhazur volume contains several step-by-step illustrations.

Appendix D

ANALYSIS OF TRAINEE CHARACTERISTICS AND PERFORMANCE DURING INSTRUCTION AS A BASIS FOR DEVELOPING ALTERNATIVE INSTRUCTIONAL TREATMENTS FOR SUBSEQUENT EXPERIMENTATION

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

ANALYSIS OF TRAINEE CHARACTERISTICS AND PERFORMANCE DURING INSTRUCTION AS A BASIS FOR DEVELOPING ALTERNATIVE INSTRUCTIONAL TREATMENTS FOR SUBSEQUENT EXPERIMENTATION

Introduction

This appendix contains an example of kinds of analyses that might be performed to provide bases for developing alternative instructional treatments intended to improve the performance of particular trainees and therefore the average performance of all trainees. The analyses are relevant to each of the first three functions of internal evaluation as pictured in Section III, Figure 9: (1) analyze needs, (2) specify objectives and design approach, and (3) develop approach.

The first part of this appendix discusses the conceptual and statistical relationships among four measures of trainee performance during self-paced instruction in a computer-managed instructional environment. Actual data are used in the first part of the appendix; they were obtained during a study to generate instructional strategies that held promise of reducing learning time. The first part of the appendix closes with some speculations about stylistic differences among trainees that appeared to influence the way in which they performed.

In the second part of the appendix, a fictitious data base is introduced. The data are similar to those in the real sample; however, to simplify presentation of some examples, scales have been compressed and simplified to show how student characteristics might be combined with training performance data to help in conceiving instructional treatments. At that point, the appendix leads back to Section III of the report.

Relationships Among Four Measures of Student Performance During Training

Four measures of student performance are obtained routinely in an Air Force computer-managed instructional setting. The meaning of these measures and their relationship to points in an instructional sequence are illustrated in Figure D-1.

1. MTM (first attempt measured time) defines instructional clock-time from the beginning of instruction to the first attempt of the criterion test.

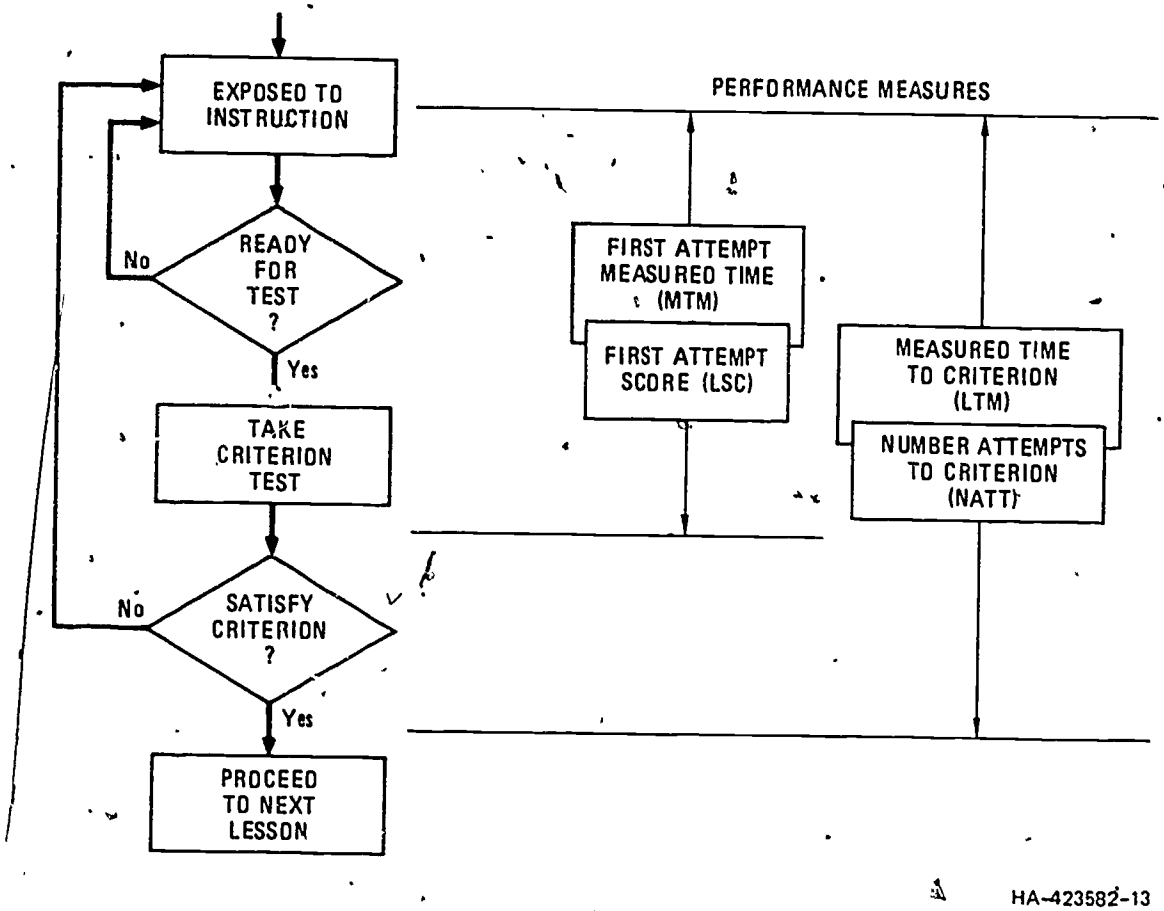

FIGURE D-1 PERFORMANCE MEASURES IN RELATION TO SELF-PACED INSTRUCTIONAL SEQUENCE

2. LSC (criterion test score on first attempt) is recorded following the first attempt.

3. LTM (measured time to criterion) identifies the total instructional and test-taking clock-time from the beginning of instruction until the test criterion is satisfied.

4. NATT (number of attempts to criterion) is a count of the number of times a student takes the criterion test.

In an analysis of student performance in a course at the computer-managed instructional site, composite indices for each of the above measures were constructed to describe the average performance of 136 trainees through a sequence of three consecutive lessons. Scores on each measure for each of the three lessons were transformed to standard scores or z-scores (i.e., each student's score was expressed as the deviation from the mean score of the group). This transformation put all scores from the three lessons on a common scale, thus adjusting for differences between lessons in the study time required and the number of items in the criterion tests. The resulting standard scores for each measure from each lesson were then summed and composite scores across all three lessons were computed from the combined distributions of standard scores for each measure. These composite scores were labelled K in subsequent analyses -- KMTM, KLSC, KLTM, and KNATT.

Table D-1 shows intercorrelations among these composite scores as well as means and standard deviations for each. (Recall that the composite of individual lesson z-scores created an abstract scale for each measure. For example, KMTM scores ranged from a "slow" score of -5.798 to a "fast" score of 3.937. As shown in Table D-1, the mean KMTM score was 0.000 and the standard deviation was 2.223.)

Figure D-2 shows histograms for the four frequency distributions. These histograms illustrate the negative skewness (lack of symmetry) of each, the most extreme of which is the KNATT measure.

The data in Table D-1 and Figure D-2 provide one basis for identifying stylistic differences among trainees. Differences among trainees, in turn, suggest alternatives in instructional approaches that invite experimentation and evaluation.

The time relationships among the four performance measures, as illustrated earlier in Figure D-1, should be kept in mind when interpreting the correlation coefficients shown in Table D-1.

1. The LTM and NATT measures constitute completion of a lesson or segment of instruction. Either or both are appropriate dependent variables or outcome measures.

Table D-1

CORRELATIONS AMONG COMPOSITE PERFORMANCE MEASURES AND DESCRIPTIVE STATISTICS FOR EACH

	Correlations Between Composite Measures			
Composite Measure	KMTM	KLSC	KLTM	KNATT
KMTM	--	.3600	.9267	.2592
KLSC	.3600	--	.5574	.8135
KLTM	.9267	.5574	--	.4803
KNATT	.2592	.8135	.4803	--

Descriptive Statistics	KMTM	KLSC	KLTM	KNATT
Median	0.080	0.451	0.071	0.513
Mean	0.000	0.000	0.000	0.000
Standard Deviation	2.223	2.132	2.317	2.072
"Best" Score	3.937	2.941	4.072	1.778
"Worst" Score	-5.798	-6.927	-6.198	-8.406

Note: N = 136 for all computations.

All variables scaled in the "desirable" direction. Positive signs denote (1) short time to first attempt (KMTM), (2) high first attempt scores (KLSC), (3) short time to criterion (KLTM), and (4) few number of attempts to criterion (KNATT).

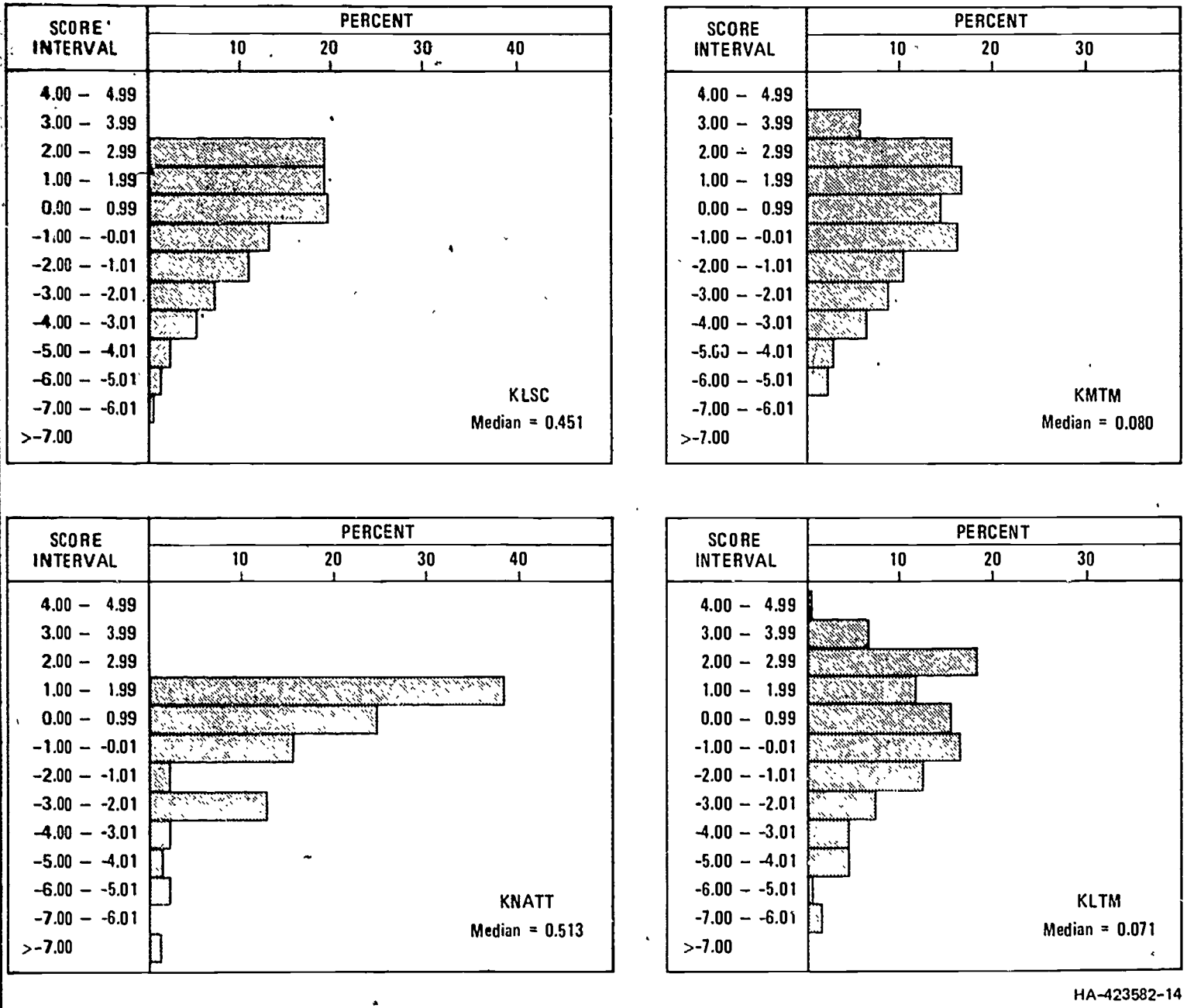

FIGURE D-2 HISTOGRAMS OF FOUR COMPOSITE PERFORMANCE MEASURES

2. The NATT measure also is an independent, or predictor, variable with reference to LTM. Time to criterion (LTM) depends in part on number of attempts (NATT). It does not make conceptual sense, however, to view NATT as depending on total time.

3. The LSC measure is an appropriate predictor of either LTM or NATT or both. The first attempt score is pivotal in its influence on the other performance measures. If the first attempt score satisfies the criterion, then the first attempt is the only attempt; that is, NATT = 1. Also, if LSC satisfies the criterion, then MTM = LTM. On the other hand, if LSC does not satisfy the criterion, then (a) NATT > 1 and (b) LTM = MTM + (Time for Attempts Beyond the First Attempt).

4. The MTM measure is an appropriate predictor of NATT. The MTM measure also may be used as a predictor of LTM. The correlation between MTM and LTM is inflated, however, since MTM is a component of LTM; that is, LTM = MTM + (Time for Attempts Beyond the First Attempt). For the majority of the 136 trainees whose performance was summarized in Table D-1, LTM = MTM on each of the three individual lessons from which the composite scores were constructed. When performance scores were combined over three lessons, however, KLTM > KMTM for a substantial fraction of trainees.

Three complementary correlational analyses help clarify the pattern of relationships among the composite performance measures as summarized in Table D-1. Consider first the relationship between the KMTM and KLTM scores. As noted above, KLTM = KMTM + X, where X = time for attempts beyond the first one. The correlation between KMTM and KLTM is high (.9267) due to the part-whole relationship between the two measures. Also, the extreme skewness in KNATT scores indicates that most trainees require only one or a very few attempts to satisfy the criterion. The correlation between KMTM and the additional time, X, can be computed as -.0868. This indicates that KMTM is essentially unrelated to the additional time beyond KMTM that represents the difference between KLTM and KMTM -- high KMTM is as likely to be accompanied by low additional time as by high additional time.

Consider next a series of partial correlations involving the four performance measures. A partial correlation estimates the relationship between two variables when the influence of one or more other variables has been eliminated or "partialled out." Table D-2 shows five sets of partial correlations. In the first three sets, KLTM is the dependent variable with KNATT, KLSC, and KMTM, in turn, used as the single predictor variable. In the remaining two sets, KNATT is the dependent variable with KLSC and KMTM, in turn, treated as the single predictor variable.

Table D-2

SIMPLE AND PARTIAL CORRELATIONS AMONG PERFORMANCE MEASURES

Variables Dependent	Correlated Predictor	Zero-order Correlation	First-order Partial Correlation: Variable Partialled Out	First-order Partial Correlation: Coefficient	Second-order Partial Correlation: Variables Partialled Out	Second-order Partial Correlation: Coefficient
LTM	MTM	.92670	LSC	.93732		
			NATT	.94697	LSC, NATT	.94405
LTM	NATT	.48025	MTM	.66146		
			LSC	.05548	MTM, LSC	.32715
LTM	LSC	.55740	MTM	.63820		
			NATT	.32682	MTM, NATT	.24923
NATT	LSC	.81353	MTM	.79930		
NATT	MTM	.25916	LSC	-.06220		

The middle column of Table D-2 shows first-order partial correlations, or the relationship between the dependent and predictor variables when one or another third variable is partialled out. The right-hand column shows second-order partial correlations, obtained when the joint influence of two variables is partialled out.

Several inferences may be drawn from the partial correlation analysis shown in Table D-2:

1. The spuriously high correlation between KLTM and KMTM is well illustrated by the sequence of computations in the first row of Table D-2. The simple correlation between KLTM and KMTM is .92670. Partialling out KLSC and KNATT, either singly or as a pair, has very little effect on the relationship. The dominance of KLTM by KMTM, of course, follows from the definitions of the two variables, as noted earlier.

2. When the effects of KMTM on KLTM are eliminated, the apparent relationship between KLTM and KNATT and between KLTM and KLSC

are somewhat greater than indicated by the simple correlations. For KLTM vs. KNATT (second row of Table D-2), the correlation increases from .48025 to .66146 when KMTM is partialled out. In the KLTM vs. KLSC relationship (third row of Table D-2), the coefficient increases from .55740 to .63820.

3. KNATT is predicted rather well by KLSC, as the definitions of the variables would lead one to expect. The simple correlation between KNATT and KLSC is .81353, as shown in the fourth row of Table D-2. When the effect of KMTM is partialled out, the relationship is decreased only very slightly to .79930. The strength of the relationship between KNATT and KLSC also is demonstrated by (a) the very low correlation between KLTM and KNATT when KLSC is partialled out (.05548 as shown in the second row of Table D-2) and (b) the very low correlation between KNATT and KMTM when KLSC is partialled out (-.06220 as shown in the fifth row of Table D-2).

4. The second-order partial correlations (a) between KLTM and KNATT with the joint influence of KMTM and KLSC partialled out (.32715 as shown in the second row of Table D-2) and (b) between KLTM and KLSC with the joint influence of KMTM and KNATT partialled out (.24293 as shown in the third row of Table D-2) both reinforce the apparent low-moderate relationship between KMTM and KLSC that was shown earlier in the simple correlation of .36004 (see Table D-1). There is a slight tendency for a short time to first attempt to be associated with a successful first attempt. However, KMTM does not have much practical use as a predictor of KNATT. As will be shown below, a cross-tabulation of KMTM and KLSC serves to identify two subgroups that differ stylistically from one another -- some (who might be characterized as "gamblers") who achieve criterion after many attempts and many errors and others (who might be characterized as "sure bettors") who achieve criterion with fewer attempts but longer study time than the "gamblers."

A third way to look at the relationships among the performance measures is in a multiple correlation sense; that is, the relationship between a dependent variable (KLTM or KNATT) and a best-weighted combination of predictors.

As the partial correlation analysis showed, KMTM alone is virtually interchangeable with KLTM. When KMTM is paired with either KNATT or KLSC as a predictor of KLTM, essentially identical multiple correlation coefficients of about .96 are obtained, thus indicating that KNATT or KLSC contribute little to explained variance in KLTM beyond that attributable to KMTM. When KNATT and KLSC are paired as predictors of KLTM, the obtained coefficient of .5593 is virtually identical to the coefficient of .5574 that describes the direct relationship between KLTM and KLSC.

A similar situation applies to the prediction of KNATT as a dependent variable. KNATT is fairly well predicted by KLSC alone (.81353), and not well predicted by KMTM alone (.25916). Combining KLSC and KMTM as predictors of KNATT yields a coefficient of .81433 -- in short, KMTM adds nothing to KLSC as a predictor of NATT.

We noted above that a cross-tabulation of First Attempt Measured Time (KMTM) and First Attempt Score (KLSC) suggests substantial stylistic differences among trainees in the manner in which they approach trials on criterion tests in the self-paced instruction.

The cross-tabulation between KLSC and KMTM is shown in Figure D-3. In this form, the plot is called a "scattergram." Plotting the scatter of paired X and Y values for each case is a useful way to get a visual idea of the shape of a distribution that underlies a correlation coefficient.

Before discussing Figure D-3, it is important to emphasize that the KMTM scale in the scattergram (the X-axis) runs in a direction opposite that used in the descriptive statistics reported earlier in Table D-1 and shown as a histogram in Figure D-2. In both Table D-1 and Figure D-2, scores had been scaled to reflect "desirable" directions.* Since "short time to first attempt" was considered more desirable than "long time to first attempt," the scales in Table D-1 and Figure D-2 show "short time" as a positive score and "long time" as a negative score. In the scattergram shown in Figure D-3, the conventional practice has been followed of showing values as increasing upward on the vertical (Y-axis) and to the right on the horizontal (X-axis). This reversal means that the correlation coefficient computed from Figure D-3 is -.3600 in contrast to the coefficient of +.3600 reported in Table D-1. Also, the median for ZMTM from Figure D-3 is -0.080 rather than +0.080 as reported in Figure D-2.

Returning to the scattergram in Figure D-3, the moderately weak relationship denoted by the correlation coefficient is evident from the relatively formless scatter of the points (X, Y pairs). The horizontal and vertical dashed lines mark the means of both scales; their intersection is the arithmetic mean of the combined distributions and since both means = 0.0, this also is the origin of the two-way plot.

---

*It is common practice, and frequently helpful in interpreting findings when several variables are being analyzed, to reflect some scales (multiply by -1) so that the "good" or "desirable" ends of all scales have the same sign. Reflecting scales does not affect the strength of a relationship but will reverse the signs of correlation coefficients so that all "good-to-good and bad-to-bad" associations carry a positive sign.

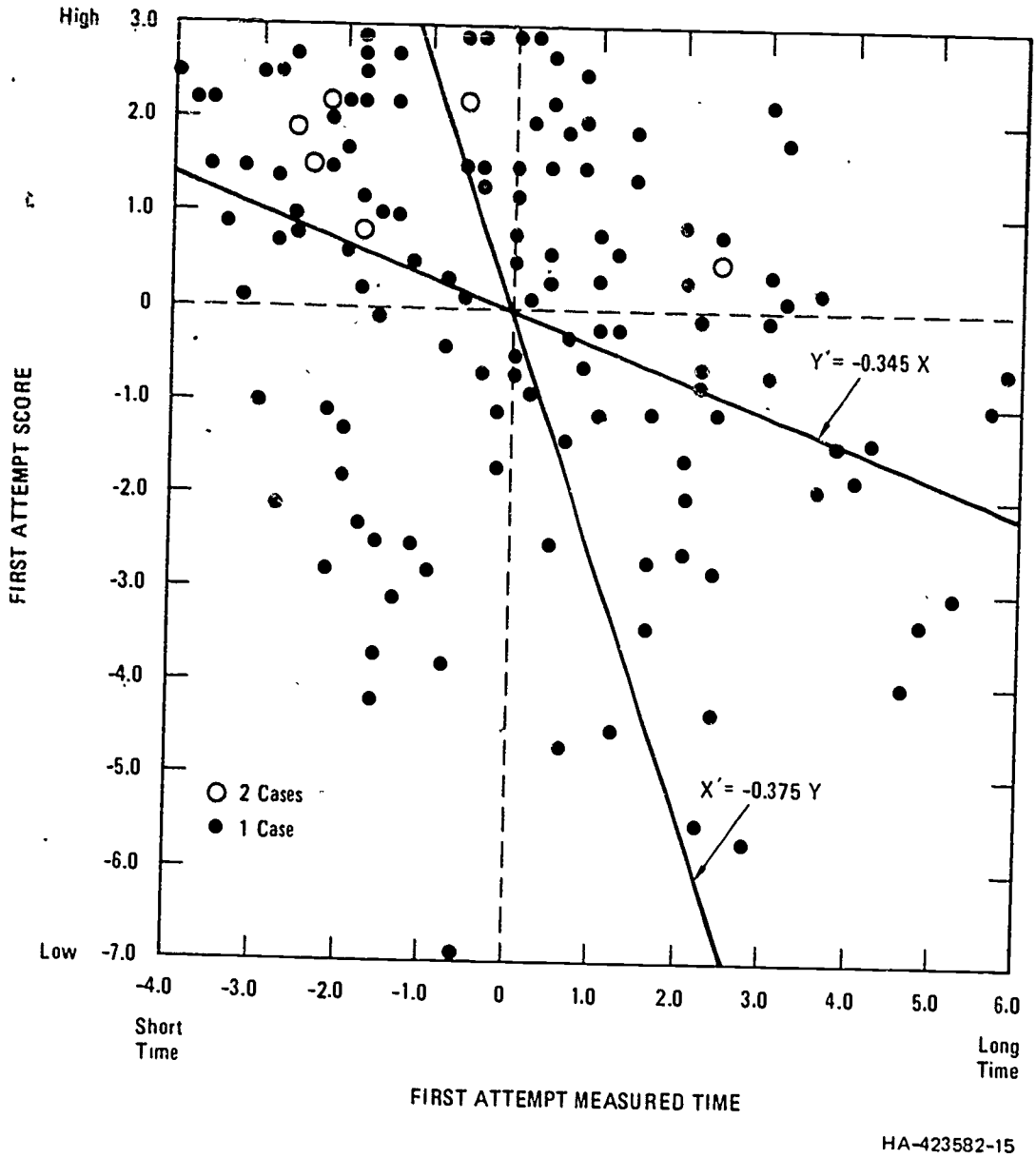

FIGURE D-3 RELATIONSHIP BETWEEN FIRST ATTEMPT SCORE AND FIRST ATTEMPT MEASURED TIME FOR 136 TRAINEES

The two solid lines that intersect at the mean are the two regression lines. The less-steep line, labelled Y' = -.345 X, is the regression for Y predicted from X or First Attempt Score (KLSC) predicted from First Attempt Measured Time (KMTM). The equation says that, on the average, for every unit increase in X, there is a .345 decrease in Y. The second line, labelled X' = -.375 Y, is the regression of X on Y or KMTM on KLSC. This equation says that, on the average, for every unit increase in Y, there is a .375 decrease in X. (Note that the average of these two regression coefficients or slopes is equal to the correlation coefficient of -.360. This holds because the two means are zero and both scales are on the same metric.)

Both regression lines were computed by procedures described in Appendix C. Both are least-squares best-fit lines. If there had been no relationship between X and Y (if r = 0.0), the regression lines would have corresponded to the two dashed lines -- one would have predicted any Y-score using the mean of X, and any X-score using the mean of Y. On the other hand, if the relationship between X and Y had been a perfect negative one (if r = -1.0), the two regression lines would have coincided and extended downward from the upper left through the mean to the lower right.

The fact that the relationship between scores on the first attempt (KLSC) and time to the first attempt (KMTM) is quite low suggests stylistic differences within this sample of 136 trainees. Certainly there is a cluster of trainees denoted by points in the lower-left quadrant of Figure D-3 who were relatively quick to attempt the test but also relatively unsuccessful in their first attempt. By contrast, trainees identified by points in the upper-right quadrant took relatively long times before attempting the test but generally performed well when they did attempt it.

These contrasting patterns suggest at least two sets of questions as candidates for guiding instructional treatment experiments:

1. What can be done to stimulate quicker responses from the slow but accurate types represented in the upper-right quadrant of Figure D-3? Is there something in their experience backgrounds that discourages risk-taking? If so, what changes in the instructional approach would be likely to speed them up without jeopardizing seriously their chances of doing well on the test?

2. What can be done to stimulate more deliberation from the fast but inaccurate types represented in the lower-left quadrant of Figure D-3? Why do they seem so willing to guess, as some must have done? Are they unable to evaluate their own readiness for the test, or are they using feedback from the test results as constructive guidance? What changes in the instructional approach would help them make better evaluations of their chances of passing the test without excessively reducing their willingness to test themselves?

Trainees represented by points in the lower-right quadrant of Figure D-3 present instructional problems, too. Here are trainees who are both relatively slow and inaccurate. Are they beyond their depth in the instruction? Are there shortcomings in their backgrounds, such as slow reading speed or problems in reading comprehension, with which they should be helped before they begin new instruction? Are the criterion tests adequate that brought them to this stage of training? Is it possible that faulty measurement at an earlier point suggested readiness before it was justified?

Identifying the Relative Importance of Various Predictors in Accounting for Variability on a Criterion Measure

From this point on, the data used in the illustration are fictitious. To sustain continuity with the preceding discussion, assume that the purpose is to look more closely at variables or factors that affect a criterion measure, "time to criterion." The purposes follow from the speculative discussion in the preceding paragraphs. In the following example, the purpose is to investigate how instructional efforts might be modified to reduce "time to criterion."

Table D-3 shows the full array of data used in the following exercise. Three of the variables are familiar from the preceding discussion: (a) time to criterion (LTM) which will be the criterion or dependent variable, (b) first attempt score (LSC), and (c) number attempts to criterion (NATT). Two hypothetical variables have been added to the array: (a) a "basic skills" factor which might represent measures such as reading comprehension, reading vocabulary, abstract reasoning, or prior learning experiences, and (b) an "anxiety" factor derived from a self-report paper-and-pencil measure. The two factors are not independent of one another; the correlation between them is negative and low-moderate in strength.

To simplify the presentation and make it easier to reproduce the computations, all five of the variables have been represented on three-point scales -- 0, 1, 2. In keeping with common sense equivalences, "0" means "low," "few," or "short" and "2" means "high," "many," or "long." Thus, we will discover that some "desirable" relationships are negative in sign. For example, few attempts to criterion (NATT) is "good" and a high score on the first test attempt (LSC) is "good;" as is apparent to the eye, the relationship between NATT and LSC shown at the far right in the bottom row of Table D-3 will be negative.

Table D-4 provides the descriptive statistics that go with Table D-3. In the top half of Table D-4, correlations between all pairs of variables are shown. The right-hand column shows the correlations between each predictor and the criterion. The next objective is to estimate the degree of improvement possible in the prediction of LTM provided by the Basic Skills factor alone ($r = -.6367$) if use also is made of the other three predictor variables. The square of -.6367 is

Table D-3

FREQUENCY CROSS-TABULATIONS OF HYPOTHETICAL TRAINEES ACCORDING TO SELECTED APTITUDE AND TRAINING PERFORMANCE MEASURES

Variable	Levels Within Variable	Time to Criterion: Short 0	1	Long 2	Basic Skills Factor: Low 0	1	High 2	Anxiety Factor: Low 0	1	High 2	1st attempt Score: Low 0	1	High 2	Total
Basic Skills Factor	High 2	28	12	2										
	1	10	30	7										
	Low 0	5	7	35										
Anxiety Factor	High 2	10	19	21	26	17	7							
	1	15	18	13	14	20	12							
	Low 0	18	12	10	7	10	23							
1st Attempt Score	High 2	25	22	3	7	16	27	20	15	15				
	1	12	11	23	8	26	12	14	16	16				
	Low 0	6	16	18	32	5	3	6	15	19				
Number Attempts to Criterion	Many 2	4	8	18	20	9	1	3	9	18	23	7	0	30
	1	12	19	23	20	23	11	8	13	28	15	32	7	54
	Few 0	27	22	3	7	15	30	29	19	4	2	7	43	52
	Total	43	49	44	47	47	42	40	46	50	40	46	50	136

Note: Variable means, variances, and intercorrelations computed from frequencies as shown in 3 x 3 tabulations. See Table D-4 for summary statistics. See text for explanation of example.

.4054 which indicates that the Basic Skills factor alone accounts for some 40% of the variability in the criterion measure, LTM.

Table D-4

DESCRIPTIVE STATISTICS FOR FREQUENCY CROSS-TABULATIONS SHOWN IN TABLE D-3

Correlation Between Selected Variables (N = 136)

Variable	Basic Skills	Anxiety	1st Attempt Score	No. Attempts to Criterion	Time t Criteri
Basic Skills	--	-.3890	.5544	-.5129	-.6367
Anxiety		--	-.2099	.4974	.2148
1st Attempt			--	-.7455	-.3867
Number Attempts				--	.4620
Time to Criterion					--

Means and Standard Deviations for Variables (N = 136)

Statistic	Basic Skills	Anxiety	1st Attempt Score	No. Attempts to Criterion	Time to Criterio
Mean	.9632	1.0735	1.0735	.8382	1.0074
Standard Deviation (N-1 wtd.)	.8111	.8132	.8132	.7623	.8027

Note: All variables re-scored to a three-valued scale (0, 1, 2). See Table D-3 for corespondence between adjective scores (e.g., high, short, many) and numerical values.

The next step is to determine the multiple regression equation for estimating LTM using all the predictors (Basic Skills Factor, Anxiety Factor, LSC, and NATT). Some of the discussion below nay be helped by referring to Appendix C.

Table D-5 summarizes the results of the multiple regression analysis. The full regression equation may be written as:

$$Y' = 1.2244 - .6169X_1 - .1734X_2 + .2073X_3 + .4067X_4$$

where Y' = Predicted Value of Time to Criterion (LTM)

X_1 = Basic Skills Factor

X_2 = Anxiety Factor

X_3 = 1st Attempt Score (LSC)

X_4 = Number Attempts (NATT)

The obtained R^2 of .4564 is somewhat greater than .4054 or the square of the simple correlation between Basic Skills and LTM; the other predictors have added about 5% to the variability accounted for in LTM.

Table D-5

SUMMARY OF MULTIPLE REGRESSION ANALYSIS OF FICTITIOUS DATA SHOWN IN TABLES D-3 and D-4 FOR DEPENDENT VARIABLE OF "MEASURED TIME TO CRITERION"

Variable	Parameter Estimates	Standardized Coefficients
Intercept	1.2244	.0000
Basic skills factor (X_1)	-.6169	-.6234
Anxiety factor (X_2)	-.1734	-.1757
1st attempt score (X_3)	.2073	.2100
Number of attempts to criterion (X_4)	.4067	.3862
Square of multiple coefficient		.4564
Standard error of estimate		.5919

The step of significance testing of the separate regression coefficients will be skipped on the grounds that we are only "data snooping." It is apparent from the standardized coefficients shown in Table D-5 that the Basic Skills factor is the most potent of the predictors in this hypothetical example.

Some ideas for instructional treatment design may be drawn from an analysis such as this (remembering, of course, that all the data are fictitious):

1. We discovered that the Anxiety Factor operates in a somewhat unanticipated fashion. From the correlations shown in Table D-4, we might have expected the Anxiety Factor to work against LTM. The regression coefficient, however, suggests that some anxiety may be a good thing. If various combinations of Basic Skills values and Anxiety values are substituted in the regression equation and LSC and NATT are held constant at their mean values, an increase in the Anxiety factor leads to lower LTM values at every level of the Basic Skills factor. The Anxiety effects are not nearly so great as the Basic Skills effects, but they do run in a direction that suggests that more, rather than less, anxiety is predictive of lower LTM scores. Could this be interpreted to mean

that generating a little anxiety among the slow-but-accurate trainees would be helpful in reducing time to criterion for them? How might this be done? Should some moderate anxiety be generated by imposing some time constraints on self-pacing, especially for trainees with high Basic Skills?

2. The LSC variable also does not operate as anticipated. Its regression weight is positive in sign which is contrary to the sign of its simple correlation coefficient. This indicates that a high first attempt score, other things being equal, appears to be associated with longer, rather than shorter, time to criterion. This is a reflection of the "sure bettor" phenomenon, described earlier. It suggests that training time might be reduced by encouraging this group to test somewhat earlier.

3. The potency of the Basic Skills Factor is affirmed by the regression analysis. What implications can we derive from this? For example, it may suggest that more practice -- even remedial instruction outside the mainstream of the instructional course -- should be provided for incoming trainees who show deficiencies in aptitudes underlying the Basic Skills Factor.

Appendix E

COMMONLY ENCOUNTERED STATISTICAL CONCEPTS

Appendix E

COMMONLY ENCOUNTERED STATISTICAL CONCEPTS

The symbols and notation in this selective compendium are not universal but reflect notation used in this report. See the reference list in the report for several widely used textbooks and reference works in statistical methods and test theory for further detail and explanation.

Statistic	Symbol	Formula	Comment
Raw score of individual i on variable X	X_i	--	Paired scores usually symbolized by X (independent variable) and Y (dependent variable). Therefore, Y_i = raw score of individual i on variable Y.
Number of individuals in sample	N	--	
Sum of N values X_i, beginning with i = 1 and ending with i = N.	$\sum_{i=1}^{N} X_i$	$X_1+X_2+\ldots+X_N$	Raw scores. Subscripts and superscripts often omitted when meaning is clear; thus, $\Sigma X = X_1+X_2+\ldots+X_N$

Statistic	Symbol	Formula	Comment
Mean	$\bar{X}$	$\frac{\sum_{i=1}^{N} X_i}{N}$	Often written as $\Sigma X/N$ when rules of summation are obvious.
Deviation score	x_i	$X_i - \bar{X}$	Sum of deviation scores = 0; $\Sigma x = 0$.
Sum of squares of deviations from the mean	Σx^2	$\Sigma X^2 - [(\Sigma X)^2/N]$	Equivalent formula: $\Sigma x^2 = \Sigma X^2 - N\bar{X}^2$ Reminder! $\Sigma X^2 \neq (\Sigma X)^2$
Variance of specified sample	S^2	$\Sigma x^2/N$	Deviation score form
	S^2	$(\Sigma X^2/N) - \bar{X}^2$	Raw score form
Unbiased estimate of variance in population (universe) from which sample is drawn	s^2	$\Sigma x^2/N-1$	Deviation score form
	s^2	$\frac{\Sigma X^2 - [(\Sigma X)^2/N]}{N-1}$	Raw score form. Note also that $s^2 = S^2(N/N-1)$
Standard deviation of specified sample	S	$\sqrt{S^2}$	Exponential notation sometimes used instead of radical sign. In general, $\sqrt[n]{A} = (A)^{1/n}$

Statistic	Symbol	Formula	Comment
Unbiased estimate of standard deviation in population from which sample is drawn	s	$\sqrt{s^2}$	Also may be written as: $s = (s^2)^{1/2}$
Standardized score (z-score, deviation score)	z	$(X_i - \overline{X})/S_x$	Distribution of z-scores will have $\overline{X}_z = 0$, $S_z = 1$.
Score of individual i on item g	A_{ig}	--	Conventional item scoring: 1 = correct, 0 = incorrect.
Proportion answering item g correctly	p_g	$\frac{\sum_{i=1}^{N} A_{ig}}{N}$	Number of correct answers divided by total of correct and incorrect answers. Often called "item difficulty."
Proportion answering item g incorrectly	q_g	$1 - p_g$	$p + q = 1.0$
Variance of item g	s_g^2	$p_g q_g$	Equivalent formula: $s_g^2 = p_g - p_g^2$

Statistic	Symbol	Formula	Comment
Sum of N products of paired scores, X_i and Y_i	$\sum_{i=1}^{N} X_i Y_i$	$X_1Y_1+X_2Y_2+\ldots+X_NY_N$	Raw scores. Often expressed simply as ΣXY. Commonly called "sum of cross-products."
Pearson product-moment correlation coefficient	r_{xy}	$\frac{\Sigma xy}{N s_x s_y}$	Definition formula in deviation score terms. Pearson r sometimes called "PM correlation." Also may be referred to as "simple correlation" or "zero-order correlation."
	r_{XY}	$\frac{N\Sigma XY - \Sigma X\Sigma Y}{\sqrt{N\Sigma X^2-(\Sigma X)^2}\sqrt{N\Sigma Y^2-(\Sigma Y)^2}}$	Computation formula in raw score terms. Alternate formula if S_X and S_Y known: $\frac{N\Sigma XY - \Sigma X\Sigma Y}{N^2 S_X S_Y}$
Slope of regression of Y on X	b (in $Y' = a + bX$)	$r_{XY}(s_Y/s_X)$	Easiest formula if r_{XY}, s_Y, and s_X known.
	b (in $Y' = a + bX$)	$\frac{\Sigma XY-[\Sigma X(\Sigma Y/N)]}{\Sigma X^2-[(\Sigma X)^2/N]}$	Raw score form.

Statistic	Symbol	Formula	Comment
Intercept of regression of Y on X	a (in $Y' = a + bX$)	$\overline{X}_Y - (b\overline{X}_X)$	Easiest formula if b, $\overline{X}_X$, and $\overline{X}_Y$ known.
	a (in $Y' = a + bX$)	$\frac{\Sigma Y}{N} - [b(\Sigma X/N)]$	Raw score form; uses slope term, b.