DOCUMENT RESUME

ED 118 579 TM 005 040

AUTHOR	Pasquariella, Bernard G.; Wishik, Samuel M.
TITLE	Evaluating Training Effectiveness and Trainee Achievement: Methodology for Measurement of Changes in Levels of Cognitive Competence. Manuals for Evaluation of Family Planning and Population Programs, Number 8.
INSTITUTION	Columbia Univ., New York, N.Y. International Inst. for the Study of Human Reproduction.
SPONS AGENCY	Agency for International Development (Dept. of State), Washington, D.C.; Ford Foundation, New York, N.Y.
REPORT NO	Man-8
PUB DATE	75
NOTE	221p.; Some of the Figures in the text and some pages in the appendices may reproduce poorly due to small print
AVAILABLE FROM	International Insitute for the Study of Human Reproduction, 78 Haven Avenue, New York, New York 10032 ($3.00)
EDRS PRICE	MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS	Academic Achievement; Achievement Gains; Achievement Tests; *Cognitive Measurement; Comparative Analysis; Computer Programs; Data Analysis; *Educational Programs; Evaluation Methods; Family Planning; *Guidelines; Manuals; Program Effectiveness; *Program Evaluation; Statistical Analysis; Test Construction; *Trainees; Training; Training Objectives

ABSTRACT

This Manual has been designed to provide step-by-step guidelines for conducting an evaluation of a structured training sequence. The assessment design to be presented involves essentially: the testing of a group of trainees before and after a sequence of instruction by administration of the same set of objective-form items under structured testing conditions; and the application of a series of statistical procedures to the resultant scores and individual item responses to determine the magnitude, direction and level of Test to Retest changes in cognitive (subject matter) competence. As will be stressed repeatedly throughout the Manual, the quantitative analysis of the testing data can provide both a measure of trainee achievement and an assessment of training effectiveness, by estimating how much of the increase in levels of subject competence displayed by the trainees at the end of the course can be attributed to the training experience. (Author)

Manuals for Evaluation of
Family Planning and Population Programs

Manual Number
of a Series

8

EVALUATING TRAINING EFFECTIVENESS AND TRAINEE ACHIEVEMENT

Methodology for Measurement of Changes in Levels of Cognitive Competence

Bernard G. Pasquariella
Samuel M. Wishik

DIVISION OF SOCIAL AND ADMINISTRATIVE SCIENCES
INTERNATIONAL INSTITUTE FOR THE STUDY OF HUMAN REPRODUCTION
COLUMBIA UNIVERSITY / 78 Haven Avenue / New York, New York 10032

EVALUATING TRAINING EFFECTIVENESS AND TRAINEE ACHIEVEMENT

Methodology for Measurement of Changes in Levels of Cognitive Competence

by

Bernard G. Pasquariella, M.A.*
Samuel M. Wishik, M.D., MPH**

With the assistance of:

William Wilkinson

*Research Associate (Psychology) and **Director, Division of Social and Administrative Sciences, International Institute for the Study of Human Reproduction; Columbia University, 78 Haven Avenue, New York, N.Y. USA 10032

1975

Published by:
Division of Social and Administrative Sciences
INTERNATIONAL INSTITUTE FOR THE STUDY OF HUMAN REPRODUCTION
Columbia University/78 Haven Avenue/New York, New York 10032

Printed in the United States of America

This work was made possible, in part, by the financial support of the Ford Foundation and the Agency for International Development (Contract AID/csd-2479). This support is gratefully acknowledged.

Library of Congress Catalog Card Number: 75-7901

PREFACE

The assessment procedures that form the basis of this Manual were originally developed in response to a request from the Demographic Association of El Salvador for help in evaluating a series of training programs in Population and Family Planning, and Maternal/Child Health, each directed at a different professional or paraprofessional level.

The one objective common to all programs in the series was that the trainees acquire a body of knowledge (and fundamental skills) in a number of Public Health-related subject areas. Thus, the evaluation of instruction was to focus on an assessment of the amount of substantive learning that occurred among the trainees in the various subject areas. The assessment was to be effected by means of a single objective test instrument administered twice under a Pre-/Post-Instruction design. Rather than create an achievement test instrument on the subject matter and send it to them, it was decided that a better procedure would be to prepare a set of guidelines for the preparation of the test instrument and allow the Association to create its own test to meet the specific program needs. As a result, the idea for a complete Manual was born.

Field testing of the methodology outlined in the guidelines was later conducted, at the invitation of the US Agency for International Development, for a training program in Washington, D.C. involving a government-sponsored Population/Family Planning Program Seminar-Workshop. Additional experience with the methods, leading to some modification of the Manual, was provided by a request for an evaluation consultation by the Department of Health and Family Protection at the National School of Public Health in Rennes, France in November 1973. (A more complete discussion of the background of the Manual is provided in Appendix A.)

While parts of the evaluation procedures are newly introduced here, the steps involved in the instrument design are for the most part not innovative, but are based on what may be taken as standard thinking on the subject of achievement testing and measurement. It was not felt that educational and psychometric theory need be brought into the body of the Manual. It is assumed that the reader will accept the authority of the sources listed in the bibliography. However, an appendix discussing some of the theoretical aspects of achievement test design and the Test/Retest model has been provided.

The Manual has been designed to be a self-contained, complete guide to objective achievement testing for purposes of training program evaluation. The text has therefore been arranged to take the reader step-by-step through the procedures -- from designing the test instrument (i.e., constructing test items and creating the test format), through the actual administration of the test, coding and scoring, and the various statistical analyses that provide the final data for the evaluation. (There are also appendices covering other aspects of the methodology for the reader who requires additional information in designing the achievement instrument or conducting the evaluation.)

NOTE: Information about individual trainees that can be derived from analysis of the testing data is not recommended for use in discriminating between trainees on matters such as job placement, salaries, future promotions, etc. The primary focus on the analysis of test data relating to individual trainees is for purposes of evaluating the effectiveness of a sequence of training, as it is reflected in the performance of the trainees as a group, and by variations in performances among subgroups and individuals.

ACKNOWLEDGMENTS

Appreciation is extended to individual members of the Institute staff for particular assistance. The authors are greatly indebted to William Wilkinson for his invaluable contributions as editor, critic and reviewer throughout the preparation of this Manual. His creativity and imagination, especially in the designing of many of the sample achievement test items provided in the text, is also appreciated. Dr. David Wolfers, Epidemiologist, is named in the text for several valuable original formulations. Dr. Prem Talwar, Senior Statistician, contributed to statistical presentations and Joanne Revson assisted in the development of the test instruments employed in the overseas field trials.

Our thanks also to the staffs of the Governmental Affairs Institute, Washington, D. C. and the Department of Health and Family Protection, National School of Public Health, Rennes, France for assistance in testing our methodology in their training programs.

Those who judge of a work by rule are in regard to others as those who have a watch are in regard to others. One says, "It is two hours ago"; the other says, "It is only three quarters of an hour." I look at my watch and say to the one, "You are weary"; and to the other, "Time gallops with you," for it is only an hour and a half ago, and I laugh at those who tell me that time goes slowly with me and that I judge by imagination. They don't know that I judge by my watch. (Pascal, Pensées)

TABLE OF CONTENTS

FIGURES IN THE TEXT

CHAPTER I

GENERAL DESCRIPTION OF THE METHODOLOGY

This Manual has been designed to provide step-by-step guidelines for conducting an evaluation of a structured training sequence.

The assessment design to be presented involves essentially: the testing of a group of trainees before and after a sequence of instruction by administration of the same set of objective-form items under structured testing conditions; and the application of a series of statistical procedures to the resultant scores and individual item responses to determine the magnitude, direction and level of Test to Retest changes in cognitive (subject matter) competence*. As will be stressed repeatedly throughout the Manual, the quantitative analysis of the testing data can provide both a measure of trainee achievement and an assessment of training effectiveness, by estimating how much of the increase in levels of subject competence displayed by the trainees at the end of the course can be attributed to the training experience.

The Manual is in two parts. The first deals with the construction of the test instrument, its administration, and methods for ensuring that the material covered in the test is compatible with what is being planned for the course and checking that the material which was included in the test was actually covered in the course.

The second part of the Manual deals with comparative analysis of the two applications of the test, with each step explained so that it can be done by either manual/mechanical methods or by computer.

The first part of the Manual will be the larger and more detailed of the two. Although the statistical analysis of the test results is important to the assessment outcome, it is the initial planning, construction and application stages that must be given special attention to ensure that the test instrument will serve its intended purpose.

---

* See Note on Terminology, p. 6. This concept refers basically to the acquisition and mastery (in terms of knowledge, understanding and application) of the subject material of instruction. (Further discussion of this concept is provided on pp. 7-8, and in Appendix B.)

Before presenting the practical, how-to aspects of the methodology, however, it is necessary to provide a small amount of background and to discuss the uses to which an instrument such as this can be appropriately put.

In order to decide on an evaluation procedure for any sequence of training, the objectives of the training must first be determined. Training program administrators have traditionally evaluated the reception and impact of their programs by a number of informal and subjective approaches such as questionnaires, rating scales, and checklists. All of these self-report procedures have been used to try to elicit the following information:

1. In relation to their needs and interests, what the individual trainees got out of the training.

2. Rating of training sessions, typically on a scale from poor to excellent.

3. Rating of individual training sessions in terms of selected aspects.

4. Extent to which trainees felt that the training had prepared them for future work in the field.

5. Rating of instructors in terms of selected considerations.

While these approaches to evaluation may provide qualitative assessments of a program's impact by identifying strengths and weaknesses as reported by trainees and by indicating trainees' feelings about the training, they do not usually supply an administrator with substantive objective feedback of the type required to assess the degree of effectiveness of current training and to implement improvements for the future. The evaluation methodology presented here was developed to provide objective, quantitative feedback to training administrators whose aim is to increase the levels of cognitive competence of trainees in specific subject areas -- that is, it provides an administrator with the means to assess the effectiveness of the training in increasing the trainees' competence with the subject matter of instruction through their ability to adapt and apply what was learned to decision-making and problem-solving situations.

The following procedural format is employed in assessing the effectiveness of instruction in terms of increasing levels of subject matter competence:

1. A pre-instruction baseline level of competence (assessing the degree to which the trainee has already acquired what is to be learned) established by the administration of a series of objective test items covering the subject material to be presented during the course of instruction. (THE PRE-TEST)

2. Review of the planned instruction to determine whether it will adequately meet the needs and demands of the current trainee group, based on the results of the Pre-Test.

3. A Curriculum Audit undertaken during the course of the training, to ascertain how much of the subject material assessed by the test items is in fact covered during instruction.

4. A second administration of the same set of objective-form test items at the end of the training sequence. (THE POST-TEST)

5. A comparative statistical analysis of Pre-Test and Post-Test results to assess the effects of training on levels of subject competence.

The administration of the test instrument provides data which can be broken down into:

1. Data on the training
   a. Total test results
   b. Results on subsets of items
   c. Results on individual items

2. Data on trainee test performance
   a. Total trainee group
   b. Trainee subgroups
   c. Individual trainees

Training Evaluation Hierarchy

It should be emphasized here that the assessment of the impact of training on subject matter competence is only the first level of training evaluation. The ultimate objective is to determine how effective the training has been in increasing the on-the-job capabilities of trainees. Between this ultimate level and the more immediate level (assessing competence)

are several intermediate levels of evaluation. These levels, when listed in terms of time sequence (realization of objectives at increasingly longer range) and measurement complexity, form a training evaluation hierarchy, beginning with the level dealt with in this Manual:

1. Cognitive Competence: How much learning (in terms of ability to use and apply relevant subject matter) can be said to have occurred as a result of training instruction? Measured by objective-type tests administered to trainees before and after training.

2. Relevant Attitude Change: How has training modified the attitudes of the trainees about the subject matter or about their jobs? Measured through structured attitude scales, projective tests, or other special test methods (e.g., the Semantic Differential Technique).

3. Short or Long-Term Retention of New Learning: How much knowledge and understanding of the subject materials do trainees retain after selected periods of time? Measured by delayed re-administration of the original test or administration of a comparable instrument.

4. Subsequent Job Placement: To what extent is the job situation of trainees relevant to the nature of the training program? Measured by structured follow-up interview or questionnaire.

5. Assigned Job Duties and Functions: Is the training content relevant to duties subsumed under the trainee's work role? Measured by structured follow-up interview, questionnaire, or observation.

6. Work Performance: To what extent does the worker employ or fail to employ knowledge and skills acquired during training? Measured by structured follow-up observation.

7. Staff/Client Relationships: What effect has training had on workers' subsequent interaction with those around them (i.e., workers, clients, patients) in the working environment? Measured by structured questionnaire or interview with trainees and others, or by on-site observation.

8. Overall Job Effectiveness: What significant contributions can be attributed to training in terms of increased capacity to meet job demands or to attain goals established by the work role? Measured by impact on achievement of work objectives.

This Manual is one of several proposed to describe the development, application and interpretation of instruments, techniques and designs for assessing the effectiveness of structured training programs at each of these levels. It has been designed to be a self-contained reference source, including the basic information needed to plan, administer and analyze an objective test instrument under a Test/Retest design, and gives additional information for specific situations.

Although primarily written for administrators of training programs concerning population, family planning, and Maternal/Child Health, the Manual may be used in a variety of training situations by individuals whose knowledge of and experience with educational assessment methods may vary. Its design is thus intended to be specific enough to permit unambiguous application of the methodology in specific training programs and flexible enough to be applicable to a variety of settings. No attempt has been made to create an actual test instrument that can be lifted directly from the Manual. Items written into a test instrument will depend on the subject material specific to a particular sequence of instruction and must be designed by those who actually conduct the training.

A Note on Terminology

Since a review of the literature on educational measurement uncovered no standard set of terms, the following terms and their definitions and equivalents should be noted:

1. Cognitive Competence = Subject (matter) Competence: a learning outcome of a structured educational experience involving the acquisition and mastery (in terms of substantive knowledge, understanding and application) of the subject material imparted during a sequence of instruction. For purposes of assessment, the operational definition of competence emphasizes usage, adaptation and application of the material learned rather than simple recognition or demand recall of the material at a later time.

2. Instructional sequence = educational input = sequence of training: the systematic imparting through structured lectures, seminars and/or recitations, of subject material of a highly specific nature. Implicit in the definition is the fact that such learning experiences are directed toward predetermined educational objectives.

3. Test = Pre-Test: the administration of an achievement-measuring instrument at the beginning of an instructional sequence.

4. Retest = Post-Test: the administration of an achievement-measuring instrument at the termination of an instructional sequence.

5. Items = Test Tasks: individual test questions or problems.

6. Item Set = Subset = Subtest: the subdivisions or grouping of items, each subdivision corresponding to a separate subject matter area.

7. Composite Test = Total Test: the total number of items comprising the complete instrument; the sum of the Item Sets.

CHAPTER II

THE TEST INSTRUMENT

Designing the Test Instrument: Overview

A valid assessment of educational achievement is the result of a systematically controlled succession of steps beginning with the identification of relevant objectives, continuing through construction and administration of the assessment instrument, and ending with scoring, analysis, and interpretation of results.

A major purpose of the training process is imparting substantive subject matter to trainees. It is safe to assume that the training instructors desire that the trainees acquire full comprehension of the scope, applications and limitations of the more significant subject matter. In order to assess the extent to which this general learning outcome is achieved, it is first necessary to translate it into components, which in turn will be translated into performance variables that can be observed and subjected to objective, quantitative measurement.

The content of most subject areas covered in training courses consists of methodology, facts, theories, problems, and points of view. In most training programs the emphasis for the trainee is on developing competence in subject content usage and application rather than on content recognition and recall. This is because what the trainee is able to do with the subject material will contribute more toward his subsequent "on-the-job" effectiveness than will simply being able to remember it on demand. Thus for purposes of assessing trainee achievement and training impact, subject matter competence is defined as the trainee's expected ability to perform specific operations on, and make specific application of, the subject material that was encountered during a sequence of training instruction.

There are a number of ways to interpret operations and applications in terms of expected trainee behavior. One common approach, for example, is to classify them in terms of the cognitive functions that contribute to those behaviors--mental processes such as concept formation, inference, analysis, synthesis, abstract reasoning, critical thinking, etc. However, the types of behavioral processes involved in this classification are too numerous and functionally interdependent, and do not readily translate into well-defined test tasks.

As an alternative approach, some type of classification of behavioral learning outcomes should comprise the domain of subject

matter competence. The desired learning outcomes would then be defined in terms of overt performance on specified test tasks. This will provide a most effective approach to the measurement of subject competence since most of the behavioral correlates of achievement can be classified into one or another of several categories. On this basis, a test item would be classified, for example, as one that contributes to the conclusion that the examinee "knows terminology and vocabulary," "knows concepts and principles," "can apply generalizations and principles to new situations," "can make valid evaluative judgments," etc. (1) Test construction guided by a classification such as this will direct the focus away from simple knowledge of definitions and facts to encompass a greater range of more complex cognitive behaviors.

Achievement Areas

The test instrument should be designed to assess a broad range of behavioral learning outcomes, from simple acquisition and comprehension of terminology, facts, and principles to higher-level abilities involving the application of what was learned to new problem-solving and decision-making situations. The seven achievement areas specified by Ebel (2) can serve as the basis for designing test items to effect this type of assessment.

The seven achievement areas and the types of items that can be designed to assess them follow:*

1. Understanding of Terminology: Items designate terms to be defined or otherwise identified. The examinee is provided with a word or words and asked to select the correct or best definition from among several alternatives (e.g., "What is an ectopic pregnancy?" or "The demographic transition is a term that describes . . ."). These are probably the simplest types of objective items to design.

2. Comprehension of Fact or Principle: Items are based on descriptive statements of the way things are. The examinee is asked to select from several alternatives the response that best completes a statement, that best answers a question, or that otherwise shows a grasp of the basic facts and principles of the subject matter at hand. (e.g., "The interrelationship between ovary and pituitary during the menstrual cycle can accurately be described as one in which . . ."; "What is the basic principle underlying the rhythm method of conception control?")

---

* See pp. 28 - 31 and Appendix C for information on item construction.

3. Ability to Calculate: Items require use of mathematical processes to get from the given to the required quantities. The examinee is provided with a well-defined computational problem together with a set of alternative answers. One example of the type of quantitative item employed in the area of family planning and population is, "Out of 200 clients initially enrolled in a family planning program, only 158 remained active one year later. What is the annual dropout rate?"

4. Ability to Explain or Illustrate: Items generally contain the words "why" or "because." This type of item has two forms. The examinee is either asked to select, from the alternatives given, the one that best explains or provides the best reason for the existence or occurrence of the specific situation cited in the item stem (e.g., "If estrogen alone and progesterone alone can successfully prevent ovulation, why is it necessary to administer both under the combined method of oral contraception?") or, the examinee is asked to select the alternative that provides the correct or best answer to the question posed in the item stem and, at the same time, to justify the answer selected, as in the following example:

> When a spermicidal preparation is the contraceptive method employed, should douching be postponed for at least six hours following coitus?
>
> a. Yes, because douching within a few hours following coitus may either remove the spermicide or dilute it to the point that sperm will survive in the vagina.
>
> b. Yes, because irrigation of the vagina within a few hours following coitus will force large numbers of live sperm into the fallopian tubes, increasing the risk of conception.
>
> c. No, because the douching agent will increase the effectiveness of the spermicidal barrier within a few hours after coitus, ensuring greater contraceptive protection.
>
> d. No, because spermicides lose their effectiveness within three hours following coitus, allowing live sperm to remain in the vagina unless removed immediately by douching.

5. Ability to Predict: Items are based on descriptions of specific situations. All conditions are given and the examinee is asked for the future result -- i.e., to select from among several alternatives the most likely outcome, as in the following example:

If the Crude Birth Rate of Hong Kong were reduced immediately to the level of the Crude Death Rate (i.e., 4/1000) and held at that rate indefinitely, assuming no net migration, what would be expected to happen to the population?

a. The population would cease to change, remaining steady at the current level with zero growth.

b. The growth rate would decline, but the population would continue to grow more and more slowly for several decades.

c. Population numbers would decline at an accelerating pace until the population virtually disappeared.

d. The growth rate would commence to oscillate between positive and negative.

6. Ability to Recommend Appropriate Action: Items are based on description of specific situations. Some conditions are given and the trainee is asked to provide by selecting from among several alternatives other conditions or actions that will lead to a specified result. For example:

Since the initial insertion of an intrauterine device can seriously damage a developing embryo (from an undetected pregnancy) the safest time to insert an IUD is

a. just before the expected menstrual cycle.
b. during and immediately after menstruation.
c. only during menstruation.
d. during the time at which ovulation is expected to occur.

7. Ability to Make an Evaluative Judgment: The types of items assessing this level of subject competence involve response options which are statements whose appropriateness or quality is judged on the basis of specific criteria presented in the item stem. For example;

Which of the following ratios provide the best indication of the overall mortality conditions in a developing country?

a. The number of infant deaths in a year per 1000 live births in that year.

b. The number of deaths per 1000 in one year over the total population at mid-year.

c. Deaths to persons over 50 years of age in a year over the total number of deaths in that year.

d. Number of deaths in a year to persons 70-74 years of age per total number of persons aged 70-74 years at mid-year.

It should be noted that the acquisition of higher level abilities (areas 4-7) depends on achievement in the first three areas. That is, it is necessary for the trainee to acquire a certain fund of information (facts, principles, computational skills, etc.) with a higher degree of comprehension before he can adapt and apply this new learning to practical situations. Therefore, items designed to assess higher-order abilities will presuppose the trainee's achievement at the lower levels. The test instrument should contain items assessing the first three areas as well, however. In the later analysis of test data it may be discovered that some of the items assessing high-level abilities were missed because the trainees did not achieve at the lower levels. For example, trainees may have done poorly on an item designed to assess their ability to predict because they didn't understand the basic principles required, or were unable to make an essential calculation.

It is strongly recommended that when designing a test blueprint all the behaviors that will apply to the specific subject material be included. Not all of them, however, will be applicable to every course of training. The relative importance of each of these behavioral outcomes as objectives of instruction will vary from program to program. For example, ability to calculate is an appropriate learning outcome to be expected from a statistical training course, but would not be a relevant outcome in a training program where the focus was on subject areas such as contraceptive technology or the anatomy and physiology of human reproduction. The final decision as to which of the seven behavioral outcomes above constitute relevant course objectives will have to be made by the training staff. The decision will be a subjective judgment based on an analysis of the specific subject matter to be covered during the sequence of training, as outlined in the curriculum plan. However, since the use and application of learned material constitutes the primary domain of subject competence, it is recommended that test items assessing abilities should have greater representation on a test relative to items sampling simple understanding of terminology, fact, principle, etc.

One final comment about constructing items assessing the seven achievement areas: It may not always be possible to write pure items -- i.e., items assessing only one of the areas to the exclusion of the other six. It may sometimes be the case that an item will call into play a number of separate abilities,

each of equal importance. This is quite valid in terms of the instrument being proposed here. These seven achievement areas should not be considered mutually exclusive categories. It is quite acceptable to write an item assessing one or more areas, as long as all areas are given representative coverage. The major purpose of the above discussion was to illustrate that an objective achievement test need not be confined to simple recognition/recall tasks, but can be so designed to assess more sophisticated, higher-order cognitive processes, the types of higher-level processes considered to underlie subject matter competence.

Subject Content Areas

Once the types of test behaviors that the examinees are required to demonstrate in an assessment of cognitive competence have been specified, the subject content areas to be covered by the test items should be determined. The content dimension is very important to the proposed assessment since it is through the course content that the behavioral outcomes are taught and through which they are demonstrated. As will be discussed later, the subject content to be assessed by the test instrument can be derived from the curriculum plan. Like the classification of behavioral learning outcomes, the course content should be arranged (for testing purposes) as a detailed outline of a limited, finite number of discreet subject matter categories. This can be done by taking each proposed training session in the order defined by the curriculum plan and listing the major topics and subtopics to be covered. When completed for all sessions, the test designer will have a complete listing of all the subject matter being proposed for presentation. How to employ this list, together with the classified behavioral dimension, in the construction of specific test items is the subject of the next few pages.

Two-Way Item Specification Table

The behavioral learning outcomes and the content of instruction represent the two dimensions which underlie the test plan. Once each has been specified, as shown above, they should be combined into a framework which will serve as a guide to the development of the test instrument. This

framework, or Item Specification Table, will serve as the test blueprint.

This blueprint, while a practical guide to test construction, is also a theoretical outline of what constitutes competence with the material to be covered during instruction. That is, it specifies which behaviors an examinee must demonstrate in which specific subject areas, in order for him to be considered as having attained a high level of subject matter competence. Properly constructed, the blueprint will illustrate not only which subject areas are to be covered by the test items, but also which of the various learned behaviors are to be expected from each area, and will indicate the relative weights assigned to each subject area and learning outcome, in terms of the number of items to be constructed.

The Item Specification Table is a two-way table that relates specific subject content to expected learning outcomes. A table of this type is easily constructed for any sequence of instruction by designing a two-way grid with the subject content areas listed along the vertical axis (left side) and the behavioral learning outcomes listed along the horizontal (top). Table cell entries will consist of check marks or some other code designating the number of items to be constructed. An example of the general format for an Item Specification Table is illustrated in Figure 1. (The numbers, in parentheses, in each of the table cells indicate the <u>percentage</u> of total items to be devoted to each behavioral outcome within each content area.) Although it was designed for a statistics training sequence, the basic format is applicable to any type of training.

The importance of employing such a table as a blueprint for constructing the test instrument becomes evident when the concept of achievement testing is considered. Any achievement test is a work sample. That is, the aggregate of items that comprises a test covering specific subject material is only a sample drawn from some hypothetical universe or population of all possible items that might be used to make up such a test. In the assessment of some curriculum areas, the population of potential items is limited -- for example, an elementary school class whose spelling competence with five hundred words is being assessed. However, for some test situations there is almost no limit to the number of potential test items that could

FIGURE 1

SAMPLE ITEM SPECIFICATION TABLE FOR THE STATISTICS COMPONENT OF A TRAINING PROGRAM IN QUANTITATIVE RESEARCH METHODS

LEARNING OUTCOME / SUBJECT MATTER	1 Understanding of Terminology, Vocabulary	2 Understanding of facts, Principles, etc.	3 Ability to Explain or Illustrate	4 Ability to Calculate	5 Ability to Predict Under Specified Conditions	6 Ability to Recommend Appropriate Action	7 Ability to make an Evaluative Judgment	total # items	total % items
ITEM SET 1									
Measurement & Scales	(1) #1	(1) #2	(1) #4			(1) #5	(1) #3	5	5
Frequency Distributions (ungrouped & grouped data)			(2) #6, 7	(1) #9		(1) #8	(1) #10	5	5
Measures of Central Tendency: Mean, Median & Mode		(1) #14	(1) #15	(1) #11		(1) #13	(1) #12	5	5
Measures of Dispersion: Range, Semi-Interquart. Range, Std. Dev. & Var.		(1) #18	(1) #16	(1) #20		(1) #19	(1) #17	5	5
Probability Sampling & Sampling Distributions			(1) #23	(5) #24, 29, 30, 31, 34	(1) #21	(3) #22, 25, 33	(5) #26, 27, 28, 32, 35	15	15
Hypothesis Testing: Point & Interval Estimations			(5)	(2) #79, 80	(3) #76, 82, 88	(5) #77, 78, 84, 89, 90	(5) #81, 83, 85, 86, 87	15	15
Analysis of Variance: One-Way Design			(2) #94, 100	(1) #99	(1) #93	(3) #92, 96, 98	(3) #91, 95, 97	10	10
total number of items	5	5	15	15	15	20	25	100	100

be constructed -- for example, the number of qualitative and quantitative items that could be constructed to cover a sequence of training in statistics. In most training courses of the type for which this Manual has been designed, the latter case will probably be more common.

Where a test instructor has no finite, discrete list from which to select the item sample to be used, he is faced with the problem of constructing an aggregate of items that will be an adequate representation of the total universe of items that would be appropriate for both the subject matter and the behavioral learning outcomes.

The purpose of the test instrument is to provide objective data for making inferences about the extent to which a sequence of training increases the levels of competence of trainees -- demonstrated by certain cognitive behaviors -- in specific subject areas. Such inferences will be valid only to the extent to which the test instrument provides a representative sampling of potential items reflecting the entire domain of subject material covered during instruction and all of the expected learning outcomes. Without a carefully developed test plan, ease of construction all to frequently becomes the dominant criterion in selecting and constructing test items (3). That is, items measuring simple knowledge (essentially recall and recognition tasks), because they are easier to construct than those assessing the more complex learning outcomes, might be over represented on the test. As a result, the test might end up assessing a limited and biased sample of behaviors and subject content areas, neglecting those that might be more relevant to the objective under consideration, namely subject matter competence.

The Item Specification Table will facilitate the process of planning both the types and numbers of items to ensure an appropriate, fair, and representative sample written into the test instrument.

In constructing the Table, the seven behavioral learning outcomes are listed regardless of the specific instruction content. However, because not all of the subject matter covered during instruction will be important enough to be assessed, some procedures must be employed by the training

staff for selecting the subject matter that will be entered into the table and thus sampled by the test items. In many training situations, some subject material serves as a background or prerequisite for other more important material. This initial material is important only to the extent to which it facilitates acquisition and understanding of the more relevant subject material and would not be the focus of direct assessment. For example, in a course on statistics, it would be necessary to introduce the students to the concept of probability before proceeding to such topics as hypothesis testing, inference and sampling techniques. While the student must have an understanding of the laws of probability, it is their application to the area of statistical inference that is important. The focus of assessment would therefore be on statistical inference directly, with little attention given to probability in itself.

In situations where the subject matter universe to be sampled for testing purposes is not composed of all the material covered during instruction, the training staff must select from the list of content areas (previously drawn from the curriculum plan) those areas which should be included in the assessment of competence. Then, each staff member would be responsible for selecting the most important topics and subtopics from his area of expertise. Each staff member should base his judgment on the criterion of relevance to competence -- which material, out of the entire range of material comprising the particular subject area covered, should the trainees be capable of dealing with (in terms of use and application) in order to be considered competent in that subject area?

Once the relevant material has been selected from all of the subject areas under assessment, the resulting list will be entered into the Item Specification Table as the content dimension. The completed Table will then be employed in specifying both the types, and numbers of items to be included in a specific test instrument to ensure a balanced assessment coverage in terms of content areas and behavioral outcomes.

Defining Relative Test Emphasis

Prior to the construction of test items it is necessary to determine what proportion of the total items on the proposed test should be constructed for each subject content area and, within each content area, for each of the behavioral learning outcomes.

There are no hard and fast rules for allocating a test item (or items) to a specific subject content area or to a specific behavioral outcome within a subject area.

One approach that is commonly used, and is suggested by a number of testing specialists; is to allocate certain numbers of items to subject areas on the basis of the amount of time devoted to each during the course of training. Unde this system, for example, if the curriculum plan for a ten week training program specifies that subject areas A and B are to each receive 4 weeks coverage, with only 2 weeks devoted to topic C, then the proportion of test items to be allocated to topics A, B and C would be 40%, 40% and 20%, respectively. Or, if the plan delegates three training sessions to coverage of subject X and only one session to subject Y, the number of test items dealing with subject X will be three times greater than those dealing with subject Y. Similarly, the number of items designed to assess the different behavioral learning outcomes within the various content areas would be based on the emphasis placed on each during the training. With the training focus on use and application of the subject material presented, the assessment instrument should emphasize the same behavioral objectives (i.e., behavioral outcomes three through seven, pp. 8-10). The training staff, employing the Item Specification Table would see to it that the relevant behavioral objectives and subject areas are adequately sampled by the test items in proportion to the emphasis given each during the course of instruction.

This approach is appropriate where only the most important subject matter within each area is included in the population from which the sample of test items is to be drawn (i.e., the areas denoted on the Item Specification Table). That is, when prerequisite or other preparatory material is not being considered for assessment, despite the fact that training time had been devoted to it, only the important content areas will appear on the vertical axis of the Table. The staff may then use the time-based approach to determine the number of items to be devoted to each of those areas.

However, a major difficulty arises with a time-based approach where some subject material is simply more difficult to get across to a training group than other material. In such a situation, the subject areas that happen to be more difficult to present may not be more important or more necessary to the training objective under assessment (i.e., subject matter competence), but more time will have to be devoted to their coverage -- thus qualifying them for a larger proportion of test items than others that may be equally important. As this is a fairly common case in training courses, a time-based approach to the allocation of items to content areas

does not guarantee that the test item content will be a representative sample of the subject content actually covered during instruction. This imbalance may well be further reflected in the Curriculum Audit (see Chap. V). At the end of the training sequence, it may be noted that much more time was devoted to some subject topics than others, although this did not reflect the importance of the topics, as determined in the curriculum plan. Where difficulty of certain content areas has been controlled for in the curriculum plan by allocating more time to coverage of these areas, this should not automatically be reflected in the construction of the testing instrument, as would happen if the amount of time alone is the criterion for determining the number of items. It is not the difficulty of the subject material that determines the number of items necessary to properly assess how competent trainees have become with it, but its importance in terms of the training objectives.

An alternative approach would be to have the training staff determine the relative importance of (and thus, the relative test emphasis to be given) each behavioral outcome and each topic within the major subject matter areas to be covered during instruction. Since the training staff is responsible for both defining the curriculum plan and instructing within selected subject areas, the staff members should then be the group most qualified for setting the standards of competence in their respective subject areas. (When the training staff includes outside instructors, as discussed on pp. 32-34. those instructors should also share the responsibility in determining the relative importance, for testing purposes, of the behavioral outcomes and subject topics under consideration.)

The allocation of test items by the training staff will be based on subjective judgments, since, in most cases, there will not be objective criteria for determining what constitutes competence in a specific subject area. Such criteria will usually be set by the staff members, drawing upon their own expertise in a particular subject area. The guiding principle underlying the allocation of test items will be that the test items, in number and content, should maintain the same relative coverage of important subject areas and behaviors that the training staff will try to achieve through instruction. Estimates of relative importance will be expressed as percentage weights to be recorded in the Item Specification Table.

In the example (see p. 17) where three subject areas are covered during a ten week training program, a time-based procedure for weighting these three areas would result in a 40%, 40%, 20% allocation of test items (given that the in-

struction time devoted to the areas was 4, 4, and 2 weeks, respectively). Assume that a total of 4 weeks (out of the 8 weeks of instruction assigned) will be devoted to coverage of prerequisite material in both areas A and B. Now, even though different amounts of time will be devoted to instruction in these three areas, the staff might judge them to be of equal importance in terms of the trainee's post-instruction competence. Thus, each subject area would be assigned percentage weight of 33 1/3% and the areas would be allocated equal numbers of test items (rather than the 40%, 40%, 20% item allocation set by the time-based approach).

In the above example, it was the professional judgment of the staff that the three subject areas are of equal importance in terms of the material that the trainees should be competent with. In this case, the time-based approach would not have been adequate for ensuring a balanced test coverage of the relevant content of instruction.

After the initial weighting, the major subject areas will be broken down into a number of discrete topics in order to provide a broader subject base from which to construct test items. A hypothetical example will outline the procedure.

A sixteen week training program, "Quantitative Methods in Health-Related Research" is composed of three major subject areas:

1. Descriptive and inferential statistics (8 weeks)
2. Techniques of demographic analysis (4 weeks)
3. Survey design (4 weeks)

Although the amount of teaching time to be devoted to area A is twice that of either B or C, the staff judges areas A and B to be twice as important as area C; therefore, each will receive an assigned weight of 40% while C receives 20%. Thus, 80% of all items to be constructed will be divided equally between areas A and B with 20% of the items allocated to assessing competence in area C.

The staff members with particular expertise in these major areas will decide which of the topics within these areas should be included in the subject universe to be assessed (selection to be based upon the "relevance to competence" criterion stated on p. 16). The material selected will then be recorded, together with the list of behavioral learning outcomes, in the Item Specification Table.

A partial listing of the relevant subject matter comprising the program's statistics training component is illustrated in the Item Specification Table in Figure 1. (The content listings for the other two areas would follow the same format

and could be recorded below the statistics section on the same table. However, given the length of combined listings from several subject areas, it might be best to break the Table up into sub-tables, one for each subject area.)

Assigning weights within subject areas Once the most important topics have been delineated for each major subject area (and recorded in the table), another series of judgments must be made by the staff. That is, the decision must be made as to how the total percentage of items allocated to a major area is to be distributed among the topics comprising that area. The allocation of percentage weights to these sub-areas will have to be based primarily upon subjective staff judgments as to the relative importance of competence in each sub-area to overall competence in the major subject area. (The weights selected will be recorded in the Table's right-most column, labelled "total % items.") For example, the staff decided that competence with "Probability Sampling and Sampling Distributions" will contribute more to overall statistical competence than will competence with such material as "Measures of Central Tendency" or "Frequency Distributions." The relative percentage weights assigned to each of these topics (i. e., 15%, 5%, 5%, respectively) are the result of these decisions.

Selecting appropriate behavioral learning outcomes The next set of staff decisions involves selecting the cognitive behavioral outcomes most appropriate to the subject content to be assessed. As pointed out earlier (e.g., see p. 7) the one objective common to most training programs is to develop, in the trainee, the cognitive competence to use and to apply the subject material learned, not simply the ability to remember the material on demand. The test instrument designed to assess competence at this level should therefore focus on the more complex learning outcomes with less emphasis on simple acquisition of knowledge.

The test design should require the trainees to demonstrate behaviors in the upper achievement areas (i.e., areas 3-7), listed on pp. 8-10 with less stress on assessing those cognitive behaviors within areas 1 and 2. Of course, not all higher level learning outcomes will be appropriate for each and every subject topic (e.g., the ability to calculate would be an inappropriate expected behavioral outcome in a seminar on human reproductive anatomy and physiology). However, the general procedure should be to consider the applicability of the higher order achievement areas before considering those lower on the list. The relative emphasis provided each of these behavioral areas when covering a certain subject during instruction will help guide the allocation of items to specific behavioral outcomes. The percentage weights assigned to a particular behavior for a specific subject topic will

be recorded in the appropriate cells of the Item Specification Table. (The numbers, in parentheses within each cell, represent these percentage weights.)

It should be noted here that the table indicates the relative importance of each cell in terms of an assigned percentage weight, not in terms of the actual number of items to be assigned to each cell. Relative numbers of assigned items will depend upon the decision of how many items will comprise the total test. (This decision point will be discussed in the next section.) Before taking up the question of determining optimum test size, one further step must be considered.

In order to obtain a more reliable estimate of the numbers of test items required, it might be necessary to subdivide the topics (within major subject areas) into subtopics. This procedure will help ensure a more balanced test coverage of the total subject area. Furthermore, subdividing will result in a number of discrete, homogeneous subject subtopics for which specific test items can be written. For example, the statistics topic "Measurement and Scales," listed in the table, can be further subdivided into:

1. Variables and constants
2. Discrete variables
3. Continuous variables
4. Nominal measurement
5. Ordinal measurement
6. Interval measurement
7. Ratio scales

With this breakdown of topics within subject area and subtopics within topics, the test designer will have a comprehensive blueprint from which the test instrument can be constructed.

Determining Total Number of Test Items

The number of items to be included in the completed test instrument is the last decision that must be made prior to beginning the task of item construction.

In many testing situations, the number of items administered is determined primarily by the amount of time available for testing. Since it is strongly recommended that the administration of the test instrument be untimed (see p. 42), the factor of time will not put constraints on test size. Since there are no hard and fast rules about numbers of items to be included in a test, the ultimate decision will be a subjective one and should involve the entire training staff. It should be kept in mind, however, that the larger the number of test

items administered, the more adequate will be the test sample (in terms of course content coverage) and the more reliable will be the scores derived from testing.

A general guide can be applied when determining the number of items to be allocated to the subject matter within major areas. When each of the topics comprising a subject area has been further partitioned into a number of discrete subtopics or sub-areas -- e.g., the subdivision of "Measurement and Scales" into 7 constituent subtopics (see p. 21) -- an attempt should be made to construct at least one item for as many of the sub-areas as is feasible. For example, when the number of topics within an area is relatively small (i.e., around 5), it might be feasible to allocate an item (or items) to each of the major subtopics.

However, when the number of topics in an area is large (e.g., those comprising the statistical training component, for which only a partial listing of subtopics was provided in Figure 1) the size of the Item Set that would result if one item were constructed for each subtopic would probably be too large to be administered effectively (especially when combined with the Item Sets sampling the remaining subject areas). In such a situation, the recommended procedure is to sample within subject topics. That is, instead of one item (or more) per subtopic, a balanced coverage of a topic can be obtained by constructing an item (or items) for every other subtopic listed -- e.g., for the "Measurement and Scales" topic, items would be constructed for subtopics 1, 3, 5 and 7. Provided that the selection of subtopics for item writing is random -- i.e., subtopics on the list are selected by ordinal position (every nth one) and not simply because good items can be written for them -- then the resulting item groups will provide a balanced and representative sample of both the subject content and cognitive behaviors to be covered during the course of instruction. Administration of test items, selected according to the above procedures, will allow a valid assessment of both trainee achievement and training effectiveness in raising the level of subject matter competence.

One factor that must be considered when determining the total number of items to include in the test is the amount of prior experience the trainee group has had in taking objective-type tests. Even when the number of test items is relatively small, a lack of experience in dealing with objective items can lower the validity of the test for its intended purpose. (The students participating in the Rennes Francophone Africa FP/MCH Training Program were not familiar with the mechanics of objective-type test taking and required an average time of four hours to complete the 116 item test.) Introducing inexperienced trainees to the mechanics of taking

the objective-type test (prior to the time of administration) will help cut down the time required for test-taking as well as allow a greater number of items to be included in the instrument.

Note: Given the above discussion, it is nonetheless possible to consider a recommendation for a reasonable ceiling on the number of items to be included in the test instrument. In order to do this, another factor, the time allocated to testing, must be taken into account.

In most testing situations, the number of items administered is determined for the most part, by the time limits imposed on the testing. For reasons to be discussed (see p. 42), it is recommended that the test instrument being proposed here be administered without rigid time limits. The only time constraint suggested is that each administration of the test be completed at one sitting so as to minimize the use of training time for testing purposes and to keep the pre- and post-testing situations as uniform as possible. Breaking up the test into two or more sub-units to be administered on successive days is possible, but not recommended since this would involve the trainee in an excessive number of time-consuming test-taking situations (i.e., possibly four or more separate sessions for combined pre- and post-testing). Placing too much emphasis on testing and assessment might have a negative effect on trainee morale and therefore decrease the effectiveness of the training experience. (This is especially true if the trainees are middle to high level professionals -- e.g., FP/MCH physicians and program administrators -- who might resent being subjected to too much testing and personal assessment.)

It is best to administer the test during an afternoon session so that the test period can extend beyond the session to allow most of the examinees to complete the test. (Although the test will be untimed, it should be untimed only to the extent that 80-90% of the examinees finish the total instrument; a small percentage of examinees will always take as much time to complete the test as they are given and therefore have to be limited in the amount of time they can take.)

Based upon our experience in constructing test instruments for the assessment of training in the area of family planning and maternal/child health we feel that we can recommend that (for most testing situations) an instrument made up of 150 items (maximum) should be adequate. If several item formats (e.g., simple and complex multiple-choice, interpretive exercises) are included with more or less equal frequency, a test of this length would require approximately three hours

to complete. (This is based upon an average trainee's read and response time of 30-45 seconds for a simple multiple-choice item; 60-90 seconds for a complex multiple-choice item, and 60-120 seconds for an item requiring computations and problem solving, or based upon an interpretive exercise.) Even allowing for a break, three hours is likely to be the limit of most trainees' endurance for test taking, after which the effects of fatigue could seriously impede test performance. If, out of all the subject matter to be covered during training, only the most important material is selected for assessment (see pp. 15-16), then a test composed of approximately 150 items should be adequate in providing a representative sample of the learning outcomes and subject matter for even a lengthy sequence of instruction covering complex material.

Naturally, the more test items that have been subjected to trial testing and item analysis, the more confident the test designer can feel that his test items are performing effectively for their intended purpose. The creation of an item file based on the analysis data (see pp. 121-127) will aid in constructing a test made up of the least number of items offering the most valid and reliable sample.

Use of Objective Items

The two types of items most commonly employed in tests designed to measure achievement are the open-ended essay item and the objective* or fixed-response item. Which type is most appropriate for use under a Test-Retest design to assess the learning outcomes being evaluated? The final selection of objective items was based on two essential factors:

1. The serious limitation of the unreliability of the scoring of essay-type items (4). Because scoring is usually based on subjective criteria, subject to the impressionistic biases of the examiner, it has been proven time and again that the same essay can receive different grades from different examiners and even from the same examiner at different points in time. To the degree that a test score reflects the private, subjective, unverifiable impressions and values of one particular scorer, it is deficient in meaning and hence in usefulness to the student who received it or to anyone else (i.e., the training staff) who is interested in the ability or achievement (being assessed) (5). If a testing procedure is designed to measure change in subject competence over time, then an unequivocal, objective scoring procedure is required so that changes that do occur can be interpreted as an increase in competence and not due to variations in the criteria for scoring between the first and second administrations of the test.

2. Employing objective items affords the opportunity for a larger and more representative sampling of relevant course subject content than is possible with essay items. Only a limited number of essay items can be given during any one testing session; thus, it is rarely possible to cover all subject areas adequately, and overemphasis on some areas of learning and total neglect of others may result.

Essay items do, of course, have some decided advantages over objective items, especially in assessing learning outcomes where originality or writing ability are important factors. In terms of the objective cited in the Manual, however, essay items pose several problems that overshadow factors favoring their use. The objective item test, presenting far more items than essay tests and reducing all responses to a form that can be easily and unambiguously scored, avoids these confounding problems (6).

---

* The term "objective" item actually refers to the fact that the correct responses are determined at the time of item construction; this helps to ensure uniformity in assessing the correctness or appropriateness of the responses given.

It should be pointed out here that the methodology also calls for the administration of the identical test both before and after the period of instruction. It is possible to employ the Pre-/Post-Test design without using the same test both times by constructing a parallel form* of the test instrument with comparable items. Due to the excessive amount of time required and the enormous technical difficulties encountered when attempting to construct two sets of items, the methodology here recommended calls for the administration of the same set of items on both occasions.

Forms of Objective Items

There are several forms of objective items, each with particular strengths and weaknesses and each requiring special skill in construction. For most applications of this instrument, limiting the number of different forms of the items used to one or two is strongly recommended. This limits the number of item-writing skills that the training staff will have to master and also eliminates difficulties which may be encountered in explaining to trainees the procedures for dealing with a number of different types of tasks. The two forms of items recommended, for reasons described below, are multiple choice items and interpretive exercises. Both forms of items can be employed quite effectively in the assessment of the impact of instruction in the areas of knowledge, understanding and application.

Multiple Choice A multiple choice item is one in which a question is posed and the trainee asked to choose the correct or best answer from a number of listed alternatives. There are many variations on this basic format (such as choosing the word(s) that best completes a statement) but in all cases the trainee is being asked to select the correct or best answer from among other incorrect or less appropriate ones.

It is, of course, possible for a person to guess the correct answer of a multiple choice item, but with only one out of four or five alternatives correct, the probability of his guessing all items is very small. Attention to the details of constructing these items will further reduce the probability of a person's guessing correctly every time. (Appendix C contains a discussion of multiple choice item construction, with examples.)

---

* Essentially two test instruments differing only with respect to the sample of items selected, the two item sets having been equated (through previous trial testing) in terms of content validity, difficulty, etc.

Interpretive Exercises Interpretive exercises are especially useful for measuring achievement of the type discussed on pp. 8-12. That is, these exercises can best be employed to assess understanding and the ability to adapt and apply what was learned rather than for the assessment of simple recall and/or recognition. Here the examinee is given material (such as a table, chart, illustration or a paragraph of text) and is asked a series of questions about it. The examinee's task is to interpret the material in one of several ways. In some cases, he may be asked to identify which statements are true; in others he may be asked to indicate an opinion about a statement's accuracy; or he may be given a series of multiple choice items regarding the material presented. In general, such exercises give the student an opportunity to show whether he has learned to apply new or old skills to the interpretation of unfamiliar data. (Rules for constructing interpretive exercises and examples will be found in Appendix C.)

Other Forms of Objective Items True/false, matching, and short answer or fill-ins are other forms of objective items often used. The major argument against fill-ins and matching items is the difficulty in arranging an answer sheet to accommodate them, especially if the scoring is to be done by computer. Short answer items must always be scored by hand, since the student must write in a word or phrase. Both forms are difficult to construct. Matching items may take a long time to answer and are, in any case, really only another more complicated form of multiple choice item. Short answers or fill-ins may often lead to ambiguous situations where a number of responses could be considered correct. While it is less possible to guess short answer items correctly (since the student must supply missing information), it is nonetheless recommended that short answer items be transformed into multiple choice items.

The chief argument against true/false items--where a student is simply asked to state whether an item is true (correct) or false (incorrect)--is that there is a 50% chance of guessing correctly on every item. However, there are other reasons why they should be avoided. It is difficult to make a statement that may be categorically classified as true or false and any distinction such as "more true" or "more false" clouds the issue and makes the item less valid as a measure of a student's competence. Items must be constructed very carefully to prevent any kind of ambiguity. Again, it is easiest to transform true/false items into multiple choice items, reducing the possibility of correct guessing and, at the same time, making all items consistent in form.

General Guidelines for Construction of Objective Items

Many of the following suggestions may seem obvious, but professionals experienced in reviewing and editing test items indicate that the most obvious faults are the most frequently committed in the preparation of objective tests.

It should be stressed that Pre-/Post-Course testing, if it is to be an accurate means of evaluating both individual achievement and the effectiveness of a sequence of training, must provide the maximum amount of objective data possible. The composition of unambiguous and untricky test items is for this reason absolutely essential. Unless skillfully written, objective items may suffer some of the disadvantages of the essay item in that different answers may be of varying degrees of correctness. Subjectivity will be introduced into scoring even though the items themselves are "objective."

It should also be stressed that the test instrument being constructed is "self-defining" (7), in the sense that the test itself defines what constitutes desired competence. The test, in turn, is constructed by the training staff based upon what they consider constitutes a high level of competence in the subject areas under assessment. How carefully the test is constructed is consequently of utmost importance. Its validity as an assessment of subject matter competence rests on the skillful construction of test items, and on the perception of the training staff in determining relevant subject areas and behaviors. It is for this reason that these aspects of the instrument have been stressed in the Manual.

The following suggestions and guidelines are provided to help avoid the difficulties inherent in the construction of objective test items and to help ensure that the items assess only what they were designed to assess.

General Guidelines (8)

1. Keep the level of reading difficulty low. Complexly written items or the use of unnecessarily technical vocabulary can put an unfair burden on the test taker and interfere with the ability to demonstrate competence in the subject matter.

2. Do not take items verbatim from books or lecture notes. Correct responses to such items may be the result of recognition or rote memorization and may not reflect understanding. In addition, lifting a sentence from its context may change its meaning. It is best to paraphrase material used in items.

3. The intended correct answer must in fact be correct. Whenever a correct answer relies not on undisputed fact, but on knowledge of opinion or point of view, the source of that opinion must be identified (e.g., "According to Malthus ...").

4. Be certain that all items deal with significant subject content. Do not use items that rely on knowledge or understanding of trivialities. Before including any item, ask yourself whether it is relevant to the desired competence of the trainees in the subject area being assessed.

5. Each item should be independent of every other item. A correct answer on one item should never be a prerequisite for answering subsequent items correctly. In addition, items should not provide clues for the solution of other items.

6. Avoid recognizable patterns in positioning of correct responses. Set the position of the correct response in a random manner, to avoid the test takers' trying to outguess you by figuring out the pattern of responses.

7. Only one alternative should be the correct or most correct response. Allowing more than one answer to be correct is confusing to test takers, and in addition turns each item into a string of "true/false" statements. It is best to allow only one answer to be correct. But make sure one is correct.

8. Make all alternatives equally plausible to the examinee who lacks the understanding or ability required to answer the item. If one or more of the alternatives is obviously ridiculous, even to someone who doesn't have any idea of what the correct answer might be, the chances of his guessing correctly are greatly improved.

9. When constructing the items according to the requirements in the Item Specification Table, it is best to construct more than the number planned for the final test instrument. Extra items will replace items which were initially acceptable, but proved to be unsatisfactory (and could not be adequately revised) when subsequently reviewed. A 20-25% reserve per Item Set should be sufficient.

10. Each item constructed should be recorded on a separate 5 X 8 card. This procedure will facilitate the review and editing of items since it is easier to revise defective items and to delete from and add to the item pool when each item is on a separate card. Also, simply by sorting the item cards, the serial placement of items for the final instrument can be arranged and, if necessary, rearranged until an acceptable order has been achieved.

In addition to the item and correct answer, the content area and cognitive behavior assessed by the item should be recorded on the card. This information can be checked against the cells of the Item Specification Table to determine if the final test item pool does, in fact, provide the balanced coverage outlined in the Table.

Finally, if an item analysis is conducted (see pp. 121-127) the results for each item should be recorded on the card. This data will help in the compilation of an effective item card file that can be used in a future evaluation study.

11. Test items should be constructed from two to four weeks prior to the time of the first Pre-Test, put aside for a period of time and then critically reviewed for defects. This procedure will, among other things, help to uncover ambiguities and inconsistencies (in subject content, grammar, vocabulary, spelling, etc.) which were initially overlooked. Whenever possible, independent staff members familiar with the subject material should be called in to review and criticize the items, help revise defective items and select replacements from the reserve sets.

Specific attention should be given to such factors as the appropriateness of the test content as well as of the cognitive behaviors called for within content areas and the accuracy of the scoring key.

In addition to assessing whether each item adheres to the guidelines provided above (i.e., nos. 1 - 8), the item review should involve some additional checks.

a) A check for balanced, representative test coverage. The question to be answered here is, "do the items, as constructed, still relate to the content/behavior cells in the Item Specification Table?" If a discrepancy exists, the item can either be revised to conform to its original purpose or re-classified on the Table according to its new content and/or behavioral objective. When large numbers of items are reclassified, a check of the Table should be made to make sure that all content areas are given representative item coverage (in terms of the percentage weights specified in the table). When necessary, reserve or newly constructed items should be added to under-represented areas.

b) If independent staff members (familiar with subject material under study) take part in the reviews, they should read each item and answer it, in addition to

checking for item defects. Any discrepancy between their answers and the keyed (correct) answer would be evaluated by the staff with appropriate action taken to reduce any ambiguity surrounding the correct answer to that item.

c) After the final item pool has been compiled and ordered serially within Item Sets, a final procedure should be carried out to determine the degree to which the content and behavioral specifications of the items actually constructed agree with the item requirements as originally defined in the Specification Table. This is done by recording the number of each item in the appropriate behavioral outcome/content cell on the Table (see Figure 1). When all item numbers have been recorded, simply check whether the number of items in each cell corresponds to the percentage of total items (the number between the parentheses) originally allocated to that cell. If the relative numbers within the cells agree, then there is evidence that the test instrument will be assessing a representative sample of the behaviors and content areas covered during the course of instruction (provided, of course, that the Table was properly constructed according to the guidelines presented on pp. 12-16). If a large number of cells show major discrepancies between numbers of items proposed and items constructed, then items should be deleted or added until there is concordance between these values in the majority of cells.

NOTE: For more specific guidelines on construction of the recommended forms of objective items, with examples and discussion of poorly constructed items, see Appendix C.

Non-Staff Lecturers and the Problem of Adequate Item Coverage

The guidelines for conducting the assessment study were originally designed for use with training programs having a "resident" training staff. This procedure was based on the assumption that the personnel who designed the curriculum plan will also conduct the instruction sessions. As field applications of the methodology continued, however, it became apparent that the average health-related training program involves input from a core (i.e., resident) staff as well as from a number of outside lecturers who are experts in their respective fields. This is especially true for programs in Family Planning and Maternal/Child Health, which require training in a number of diverse subject areas.*

The training administrators (including staff instructors) are responsible for the structuring of the training program (as discussed in Chapter II, pp. 18-21). They develop the comprehensive session-by-session curriculum plan for the sequence of instruction as well as construct the Item Specification Table for the assessment. When the core staff conducts all of the sessions, it has the added responsibility of constructing all the items comprising the competence assessment test instrument. This situation is somewhat altered when outside experts are called in as lecturers.

While the subject material to be covered by outside lecturers is broadly defined by the training administrators (in accordance with the overall subject theme of the program), the specifics of what is covered are determined by the lecturer (as subject expert). Thus, the lecturer has primary responsibility for the structure and content of his presentations. Since the subject areas covered by outside experts are part of the training curricula, it will be necessary to assess competence in these areas. The visiting lecturer will be the person most qualified to set the standards for competence in his particular area of expertise. Therefore, the non-resident lecturer should become an integral part of not only the training, but also the assessment of its impact.

---

* For example, a four month Francophone African Training Program conducted (in the Spring of 1974) at the National School of Public Health in Rennes, France, involved the sponsoring staff from the Department of Health and Family Protection as well as 30 outside instructor/lecturers from such areas as Demography/Statistics, Human Reproduction & Family Planning, Maternal/Child Health, Clinic Procedures, Health Administration, Health Education, etc.

In order to integrate the outside lecturer into the training program's evaluation framework, with a clear definition of what is required, a number of steps should be taken:

1. When outside experts are initially asked to conduct certain sessions, they should be made aware that submitting test items for the assessment of training would also be necessary. (The immediate and long range importance of training evaluation as well as the need to incorporate evaluation into the training structure during the planning stage should be impressed upon them at that time.)

2. The lecturers should be provided with the general topic areas they will be responsible for covering. They should then be requested to submit to the resident staff a detailed outline of the subject material they plan to present at each session. This outline should also include judgments concerning the relative importance (based on the relevance to competence criterion, p. 16) of the topics and subtopics comprising the subject area to be covered. Relative importance is indicated by assigning percentage weights which will ultimately determine the number of items to be allocated to each content component of the subject area (see pp. 16-21) for a discussion of assigning relative weights).

3. Each submitted outline will be incorporated into a composite Item Specification Table. For each subject area entry both the types of ability to be tested and the number of items to be constructed will be determined and recorded (see Figure 1, p. 14).

4. a. The Item Specification Table will be forwarded to each non-staff lecturer with the types and numbers of items he is to design for each subject entry clearly delineated.

   b. An achievement item "information paper" should also be provided to each non-resident lecturer. It would be a composite paper composed of those sections of the Manual which are relevant to the construction of test items. The paper should include:

      - The 7 Achievement Areas and sample achievement items (pp. 8-12).
      - Use and forms of objective-items (pp. 25-27).
      - Guidelines and rules for constructing objective-form items with examples (pp. 28-31 and Appendix C)

   This material will provide outside lecturers with all the information necessary to construct objective items that will adequately assess substantive knowledge and abilities (i.e., competence) in specific subject areas.

Requiring lecturers to construct such items without proper guidelines will probably result in groups of items which are overly represented by the more common and easily-designed categories (understanding terminology, facts and principles), with underrepresentation of those measuring more complex abilities (i.e., ability to predict, recommend appropriate action and make evaluative judgments).*

5. Each of the items submitted should be checked for adequacy in terms of the subject content and ability assessed. Appropriate changes in item structure and content should then be made. The types of items submitted should also be checked off on the Item Specification Table so that an assessment of item representation can be done.

   If the final selection of items by test trials with follow-up item analysis is not feasible (see pp 121-127 for a discussion of item tryouts and analysis), the items should be put away for awhile and then rechecked for structure and content by the resident staff before putting together the completed instrument.

   Requests for items from visiting lecturers should be made (when timing permits) several months before the start of instruction. This will allow each lecturer several weeks to comply as well as provide adequate time to check submitted items, set them aside for a short period and then recheck, revise and select the final items for the test instrument.

---

* This happened in one field situation and necessitated major revisions in many of the items submitted by visiting lecturers in order to make the overall item content more representative of the wide range of abilities subsumed under the assessment of subject competence. This extra work on the part of the resident staff might have been avoided if each visiting lecturer received a copy of the written guidelines and item samples.

THE TEST FORMAT

Once the Item Specification Table has been constructed and the individual items designed, the next step is to determine the most effective format for presenting these items to the examinees in a structured testing situation.

General Considerations

The procedures for test construction and administration to be described were developed to meet two basic criteria:

1) For trainees--neither the test directions or format, nor the testing environment should produce variation in test performance that is not correlated with differences in levels of competence among trainees.

2) For training administrators--the test format should facilitate error-free scoring and should provide economy of cost, time and effort.

The testing procedure recommended is the one most commonly used in administering group achievement tests. Each examinee is given his own set of test items and instructions (in booklet form) together with a separate sheet for recording item responses. The examiner could administer the test orally, allowing a uniform response interval following each item. However, while more economical in terms of cost of test materials, this method is not acceptable for several reasons:

1) Many of the forms of objective items to be included are too complex to be easily followed by ear;

2) Examinees cannot return to previously given items, thus encouraging guessing rather than thinking through each item carefully;

3) A structured time limit per item is imposed on the testing situation. (As will be described later in this section, every effort should be made to keep the testing untimed.)

Selecting the best method for reproducing test materials (i.e., item booklets and answer forms) should take into account such factors as the types of facilities readily available to the test constructor, the number of copies required, the types of items to be reproduced (e.g., simple worded items vs. drawings,

pictures, complex diagrams), and funds available for such use. (A detailed discussion of available document reproduction techniques is well beyond the scope of this manual; the reader is advised to discuss his particular requirements with those who specialize in document reproduction.) In most cases, the reader can employ the almost universally available and simple-to-use techniques of mimeographing and photocopying when duplicating test materials.

The Test Item Booklet

A variety of methods has been recommended for the layout of test items, usually suggesting the grouping of all items by item format (e.g., multiple choice, true/false) and arranging all items by increasing level of difficulty. However, several factors, specific to this proposed instrument, require a different item layout than is usually the case. For this instrument, a satisfactory arrangement of items must include the following considerations:

1) This instrument will contain separate subsets, each composed of items of varying format.

2) The testing will be administered without time limits.

3) The test items will be administered twice to the same examinees (and possibly again to future trainees).

Given the above factors specific to this test instrument, together with a number of factors related to achievement testing in general, the following rules and guidelines should be employed in the construction of the test item booklet:

1) A separate response sheet for recording answers should be employed; no marking should be done in the test booklet. It is then only necessary to reproduce new answer sheets, using the same item booklet when re-administering the instrument. In addition, hand scoring and key punching will be facilitated when all item responses are displayed serially on one answer sheet.

2) A complete set of directions, with sample items correctly answered, covering procedures for answering items with different formats should be displayed on the first pages of the test booklet. As stated earlier, each item subset may contain several item formats. Covering all types of items at the outset makes it unnecessary to repeat directions each time a different item format is confronted. Note: The only exception to this rule is the Interpretive Exercise Item (see p. 27) which may require

its own unique set of directions which should accompany the item. (Structure and content of directions are discussed on pp. 42-43.)

3) Within each subset, items can be grouped by format or by levels of difficulty, or both. Either way is equally acceptable since the conditions requiring one or the other of the methods do not apply here; that is, grouping by format is required when separate directions for each type are given. Also, since the test is untimed and will be given to adult examinees, it will not be necessary to arrange items entirely according to increasing difficulty. (The major assumption underlying arrangement by difficulty being that, in timed tests, if the examinee does not have enough time to finish the test, he will not have attempted those items he probably would not have answered correctly had he reached them.)

4) The entire item (i.e., the problem statement plus the response alternatives) should be placed on the same page, or on facing pages. Furthermore, if tables, graphs diagrams, etc., are presented, they should be placed on the same page as all items referring to them, or on a facing page. This is probably the most important rule related to the arrangement of items. Having to turn pages back and forth to obtain a complete idea as presented in an item can be confusing to an examinee, expecially when reference is made to graphs, tables, etc. Since it is not absolutely necessary to arrange items in any particular order within item sets, items can be rearranged on the various pages so that this rule is not violated.

5) When arranging items, correct responses to successive items should follow no pattern. When a tentative ordering of items is completed, the corresponding correct answers should be checked to determine if a repetitive pattern of response positions has resulted (e.g., a,b, a,b,c,a,b,a,b,c, etc.) When this occurs, the pattern should be altered, either by rearranging the order of alternatives of several items so that the correct answer appears in a different position among the possible alternatives, or by changing the order of several items.

6) The test title, introductory section (containing a statement of test purpose) and instructions to examinees should take up the first two pages of the test booklet with the first set of items following immediately. Although usually required in timed tests, this instru-

ment does not require the inclusion of a separate cover or Instruction and Item Section separator pages. They serve no useful purpose and only add unnecessary materials and reproduction cost. (See p.45 for a discussion of the rationale for employing a detachable Introduction/Instruction Section when designing the test booklet for Pre- and Post-Test use.)

7. Labeling item sets and individual items. Each item set is considered a separate subtest (to be subjected to independent analysis) and should be labeled as such. Each subset should be preceded by a leading (e.g., Item Set 1, Item Set 2,etc.) and the items within each subset numbered in succession beginning with Number 1. Arrows can be placed between item sets to serve as separators and to show the progression of test items.

   Labeling also serves to aid the examinee in test taking. Placing the same headings, item numbers and arrows on the answer sheet will reduce the possibility of errors when transferring the letters corresponding to the answers chosen from the test booklet to the answer sheet. (This point is discussed further in the following discussion on recording answers.)

The Answer Sheet

Employing a separate answer sheet greatly benefits the test administrators by reducing costs and facilitating scoring and key punching. That using separate answer sheets does not complicate the examinees' test-taking behavior is evidenced by the fact that many widely employed achievement tests use separate answer sheets with little or no difficulty, even at the primary school level.

There are two general types of answer sheets in current use. One is adapted for hand scoring, using a type of overlay stencil with holes punched corresponding to the correct responses. The other type is for machine scoring and is specially designed to be read by a test-scoring machine or optical scanner. There is no advantage in using machine scored sheets unless: a) machine scoring facilities are readily available; b) the machines can accommodate individual subtest scoring and; c) the number of tests to be processed warrants the expense of machine scoring. A more technical discussion of this mode of scoring is beyond the scope of this Manual. The reader considering machine scoring should discuss his specific needs with the suppliers of such services when he is planning the layout of items.

The design of the answer sheet to be hand scored or coded for key punching must be simple so that it provides an unambiguous task both for the examinee and the scorer. Systematic errors either from incorrect recording of responses on the answer sheet or from inaccuracies in scoring will confound the analysis based upon resulting scores.

The layout of the answer sheet can be based on a variety of formats. The most appropriate format is one which takes into account the test booklet design as well as some general rules of test-taking.

1. The general layout of the answer sheet should correspond to the item format in the test booklet. Item subgroups should be labeled by set number; items within each set should be numbered consecutively; beginning with number 1, and there should be separations between subsets with indicators (arrows) showing the correct progression of items. The labels and arrows, appearing as they do on both the test booklet and answer sheet, reduce the possibility of error in marking answers. Also, the item responses are serially displayed on the answer sheet so that simple hand scoring can be easily and quickly carried out. A sample answer sheet, designed for a 75 item test (with 5 item sets), is shown in Figure 2.

2. Complete instructions for recording answers should be stated on both the test booklet and answer sheet; sample items should also be provided in the booklet with space for responding to them set aside on the answer sheet. This will allow the examinee (before beginning the actual test) to become familiar with the item formats and with the procedure for recording answers on the separate sheet. (Sample items will not be scored.)

3. Examinees should be instructed to cross out (i.e., place an "X" through) the letter corresponding to the answer selected for each item. An "X" will readily appear through the holes of the scoring stencil while "circles" around the letter may not, resulting in failure to credit a correctly answered item.

FIGURE 2

MODEL ANSWER SHEET

TRAINEE ID _____

DATE _____

Instructions: Place an "X" through the correct, or most correct, answer alongside the same item number as in the booklet.

Sample Items
1. a b c d
2. a b c d

SET 1

1. a b c d
2. a b c d
3. a b c d
4. a b c d
5. a b c d
6. a b c d
7. a b c d
8. a b c d
9. a b c d
10. a b c d
11. a b c d
12. a b c d
13. a b c d
14. a b c d
15. a b c d

↓

SET 2

1. a b c d
2. a b c d
3. a b c d
4. a b c d

(set 2 cont'd)

5. a b c d
6. a b c d
7. a b c d
8. a b c d
9. a b c d
10. a b c d
11. a b c d
12. a b c d
13. a b c d
14. a b c d
15. a b c d

↓

SET 3

1. a b c d
2. a b c d
3. a b c d
4. a b c d
5. a b c d
6. a b c d
7. a b c d
8. a b c d

(set 3 cont'd)

9. a b c d
10. a b c d
11. a b c d
12. a b c d
13. a b c d
14. a b c d
15. a b c d

↓

SET 4

1. a b c d
2. a b c d
3. a b c d
4. a b c d
5. a b c d
6. a b c d
7. a b c d
8. a b c d
9. a b c d
10. a b c d
11. a b c d
12. a b c d

(set 4 cont'd)

13. a b c d
14. a b c d
15. a b c d

↓

SET 5

1. a b c d
2. a b c d
3. a b c d
4. a b c d
5. a b c d
6. a b c d
7. a b c d
8. a b c d
9. a b c d
10. a b c d
11. a b c d
12. a b c d
13. a b c d
14. a b c d
15. a b c d

– END –

SCORE SUMMARY	
SET 1	
SET 2	
SET 3	
SET 4	
SET 5	
TOTAL SCORE	

ADMINISTERING THE INSTRUMENT

The achievement measures derived from the test instrument described here are based on responses to a set of structured stimuli. The test items themselves are only one component of the total stimulus situation, the other integral part being the test instructions. The interaction of these two components gives the test situation its structure and provides a standardized basis for measurement. That is, what the completed instrument measures is not only a function of the types of items employed but also a function of the examinee's conception of the test's purpose and what he is required to do. Since test instructions and administration procedures have an influence upon the measurement obtained, their formulation should be an integral part of the planning and development of the instrument. Their development should parallel item construction and the design of the physical layout.

The testing situation must be structured so as to reduce the influence on the examinee's test performance of factors other than those related directly to training outcome.

Extraneous Factors

In order to underscore the importance of the guidelines discussed in this section, it is necessary to review those extraneous factors that relate to the mechanics of test-taking and the testing situation.

1) The purpose of the test, depending upon how it is interpreted by the examinees, can positively or negatively influence test performance. The way in which the test is presented will have direct effects on such factors as attitude toward the test, test-taking motivation, and arousal of test anxiety.

2) Examinees in any structured testing situation vary widely with respect to prior experience with objective-type tests. Differential sophistication in objective test taking, unless controlled for, will be reflected in the variation among subsequent scores.

3) How the test constructor deals with the problem of "response-guessing" will influence test performance. Imposing penalties for guessing can introduce extraneous score variability due to influence of non-cognitive,

personality factors. For example, when penalties are assessed for wrong answers ("assumed guessing"), it is likely that a part of the variation among scores will be due to individual differences in risk-taking behavior.

4) Time constraints can have an effect on measurement outcome. There are subject areas in which speed of response is an integral component of achievement (e.g., measuring proficiency in such areas as typing, shorthand, telegraphy, airplane navigation, etc.). However, the objective of this proposed instrument is to measure both the range of subject information acquired during training and the ability to adapt and apply this learning to new situations and not to measure how fast examinees can respond to items correctly. In addition, timed testing conditions introduce the risk of operation of personality factors (e.g., risk-taking behavior, test anxiety, etc.) which, although not directly correlated with what is being measured, could affect the measurement outcome. The test should be administered as a "Work-Limit Test" of the type described by Ebel (9), the objective of which is to determine how much the examinee can do, regardless of how fast or how slowly he works.

The above factors are of primary concern in any testing situation where objective-type items are employed. The need for such controls is increased when measuring a learning outcome under a Test/Retest design. Every attempt must be made to keep the structure of the testing situation as consistent as possible from the Test to Retest. The following guidelines should be employed by the test constructor at the appropriate stages.

Test Instructions

In order to ensure uniformity of task orientation for every examinee, it is necessary to observe the following rule: regardless of any references made to the nature of the test prior to actual testing, make the assumption that the examinees know nothing about the nature of the test or the mechanics of test-taking. The test instructions (in written form) should structure the testing environment. Test performance must reflect only the behavior the instrument was designed to measure, not the examinee's ability to decode instructions.

Test administration procedures should provide a testing environment that is uniform for all examinees. A number of points

should be covered in the introductory period preceding the test-taking. These discussion points can be grouped under several topic areas.

1. Test Introduction

   a. The Pre-Test

      1) In general, examinees should be informed that they will be given a test that will provide the training staff with information for use in (1) adjusting the level of instruction to meet the needs and demands of the current trainee group (see pp. 61-64; Utility of the Pre-Test) and (2) defining program strengths and weaknesses when revising it for a future presentation. They should be informed that another test, to be given at the end of the course, will be an integral part of this evaluation. (They should not, for obvious reasons, be told that the two tests will be the same.)

      2) While the major emphasis should be on the importance of the test in helping to improve training, trainees should also be informed that the information obtained will help determine how much they have individually learned as a result of the sequence of instruction.

   b. The Post-Test

      The reintroduction of test purpose should, for the most part, be exactly the same in structure and content as at its pre-training counterpart. The exception is part (a) which should be revised so as to reintroduce the test as the second part of the evaluation procedure that was initiated at the beginning of the course.

   Note: The test description and instructions should be printed on the first pages of the test booklet. This is to ensure that all the introductory material developed by the test designer is presented in a standardized manner on both administrations of the test.

2. Procedures For Responding and Recording Responses

a. The method for selecting the correct response should be printed on the first page of the item booklet.* (This is necessary since the instrument is made up of a number of subtests, each of which will contain items varying in format.) In addition, the instructions should be followed by examples of the types of items covered, together with the correct answers.

b. Trainees should be requested to read the directions silently while they are being read aloud by the examiner. (This is suggested since examinees do not always read introductory material carefully and understanding such material is necessary for establishing the optimal "mental set" for test-taking.)

c. The examiner should demonstrate how to record response choices on the separate answer sheet. A separate section for recording answers to sample items (provided in the instructions on the test booklet) should be printed on the answer sheet.

d. Trainees should then be instructed to answer the sample items, recording their choices on the separate sheet. These items and the correct answers should be covered by the examiner to correct any errors in the mechanics of recording.

Note: It is important, so that the test items function with maximum effectiveness, that the examinees understand the instructions completely before proceeding beyond this point.

3. Instructions On Time Limits

A printed statement should reflect the fact that the testing session will not be timed; that all examinees will have adequate time to attempt all items. They should be instructed, however, not to spend more than (X)** minutes average per item, at first to go through the entire test once and then return to any unanswered items.

---

* Except for directions that should precede the more unique, complex items (e.g., interpretive exercises).

** It is the task of the item writer to determine the time required to answer the average item, depending upon such factors as item difficulty, total number of test items, etc.

4. Directions For Guessing

> Instructions should state that scores will be calculated as the number of items answered correctly, with no penalty for guessing. Examinees should be strongly encouraged (since unanswered items will be counted as incorrect) to respond to every item regardless of whether or not they are completely sure of the correct answer. (They might also be told that each response given is an important bit of information which will work to facilitate the evaluation being conducted.)

While formulation of the exact wording is the responsibility of the test designer(s), the four factors described above should be included in simple and unambiguous language to ensure that test-taking is as uniform as possible for all examinees.

A model set of instructions (with Pre-Test Introduction), incorporating the guidelines covered in this chapter, is illustrated in Appendix D. The model is flexible enough so that the reader can employ the same format when developing his own test instrument. (The model is complete for the Pre-Test; it can be adapted for the Post-Test by making the recommended changes in the introductory section.)

Note: The introductory material will vary from Test to Retest while the same group of items is administered each time. Therefore, in order to minimize the cost of duplicating test materials, the test booklet should be designed so that the Pre-Test and, later, the Post-Test Introduction section can be affixed to the same set of test items. That is, after initial testing the Pre-Test Introduction section can be removed from the test booklet and replaced with its Post-Test counterpart.

Guidelines For Administering The Instrument

The timing of test administration (specifically of the Pre-Test) can have an unwanted influence on the examinee's performance. While the Post-Test should definitely be administered during one of the last formal training sessions, the time of initial testing should be based, in part, on the nature of the training group.

1. In a situation where the training program brings individuals from different backgrounds together as a formal group for the first time, it is best to administer

the initial test several days after formal training sessions have begun. This will allow the participants to become acclimated to the training environment. Since they will have totally adapted to the training situation at the time of retesting, providing a period of adjustment prior to initial testing will help equalize the conditions underlying the two administrations of the instrument. When testing is scheduled in this way, it will be necessary for the test designer to exclude items assessing subject material covered in the sessions before the Pre-Test. The added control over testing conditions should more than compensate for the reduction in total course content coverage.

In situations where participants have had time to familiarize themselves with the training environment prior to the first formal sessions (e.g., when training courses are preceded by an orientation program), the Pre-Test can be administered during the first formal training session.

2. The introductory format (except for the minor changes suggested on page 43) should be exactly the same for both administrations of the test. Statements of purpose and instructions should be repeated orally by the examiner on the Post-Test exactly as they were for the Pre-Test. A good rule to follow is: treat the administration of the Post-Test as if it were being given to the examinees for the first time.

3. The same examiner should administer both the Pre-Test and the Post-Test whenever possible.

4. When the test examiner is someone other than the test designer, he should become thoroughly familiar with the test materials prior to the first testing session. He should study the test booklet, instructions and answer sheet thoroughly so that he can administer the test in exactly the way it was conceived by the test designer and be capable of answering questions and dealing with problems should they arise during the course of testing. It should be obvious that examinees are placed at a disadvantage when the examiner is not familiar with the test materials.

5. The examiner should respond to all questions related to test-taking mechanics. He is cautioned, however, not to attempt to answer questions relating to individual test items.

6. The examiner should scan each answer sheet as it is turned in and check that each has proper examinee identification. The final administrative task for the examiner is to make sure that all test materials have been returned, especially the test item booklet.

7. At the time of pre-testing it is important not to reveal the fact that the same test will be readministered at the end of the training course. If requests for copies of test items and/or answers are made at the end of the testing period, the examiner should make a statement to the effect that items and answers cannot be distributed to examinees since they are a standard part of the training program and their effectiveness in future use requires that they remain confidential. At the same time, examinees should be assured (when it is feasible to do so) that the results of the evaluation will be made available to them some time after the course has been completed.

CHAPTER III

CODING AND PREPARATION OF RESPONSE DATA FOR STATISTICAL ANALYSIS

This chapter discusses procedures for preparing the data on the answer sheet for hand scoring (using scoring stencils) and for computer-processed scoring (using punched card input).

The reader with ready access to machine scoring facilities will require special answer sheets designed for that purpose. This section is therefore optional for those intending to utilize machine scoring since the preparation of answer sheets can be left to the personnel in charge of such scoring services.

MANUAL/MECHANICAL PROCESSING

The Scoring Stencil

The most reliable method for hand scoring separate answer sheets is to employ a scoring stencil -- a simple answer sheet overlay with holes punched corresponding to the correct answers. The stencil, which can be easily constructed, reduces scoring to a standardized, mechanical procedure that is less subject to errors than are the more "free-style" hand-scoring methods.

1. An easy-to-design and inexpensive stencil can be constructed by reproducing a copy of the answer sheet and punching out spaces corresponding to the position of the correct answers. A sample punched stencil for use with the answer sheet (provided on p. 40) is shown in Figure 3. (The dark circles correspond to punched-out areas, designating answer positions; the dark rectangle when cut out will reveal the score summary section on the answer sheet.)

2. An alternative to this type of stencil, which is punched according to a specific set of pre-selected answers, is a more flexible, reusable type of scoring overlay, similar to the one shown in Figure 4, which was also designed for the sample answer form. (The darkened areas correspond to punched-out windows.) This type is very easy to use.

   a) Letters indicating correct answers are written in the spaces, between the parentheses, which correspond to their respective item numbers within item subsets. (The letters when pencilled-in can be erased and the stencil used again, provided that the same number of items is used with the same format.)

FIGURE 3

MODEL SCORING STENCIL I

FP/MCH Program-2
4/74-7/74

TRAINEE ID _____

DATE _____

Instructions: Place an "X" through the correct, or most correct, answer alongside the same item number as in the booklet.

Sample Items
1. a b c d
2. a b c d

SET 1
1. a b c ●
2. a ● c d
3. a b c ●
4. ● b c d
5. a b ● d
6. a b ● d
7. a ● c d
8. a b c ●
9. a b ● d
10. a b ● d
11. a ● c d
12. ● b c d
13. ● b c d
14. ● b c d
15. a ● c d

SET 2
1. a b ● d
2. a b c ●
3. a b ● d
4. a ● c d

(set 2 cont'd)
5. ● b c d
6. a ● c d
7. a b ● d
8. a b ● d
9. ● b c d
10. a b c ●
11. a b c ●
12. ● b c d
13. a b ● d
14. a ● c d
15. ● b c d

SET 3
1. a b c ●
2. a b ● d
3. a b c ●
4. a ● c d
5. a b c ●
6. ● b c d
7. a ● c d
8. a b ● d

(set 3 cont'd)
9. a ● c d
10. a ● c d
11. ● b c d
12. a ● c d
13. a b ● d
14. a b c ●
15. a b ● d

SET 4
1. ● b c d
2. ● b c d
3. a ● c d
4. a b c ●
5. a b c ●
6. a b ● d
7. a b c ●
8. a b c ●
9. ● b c d
10. a b ● d
11. a ● c d
12. a b c ●

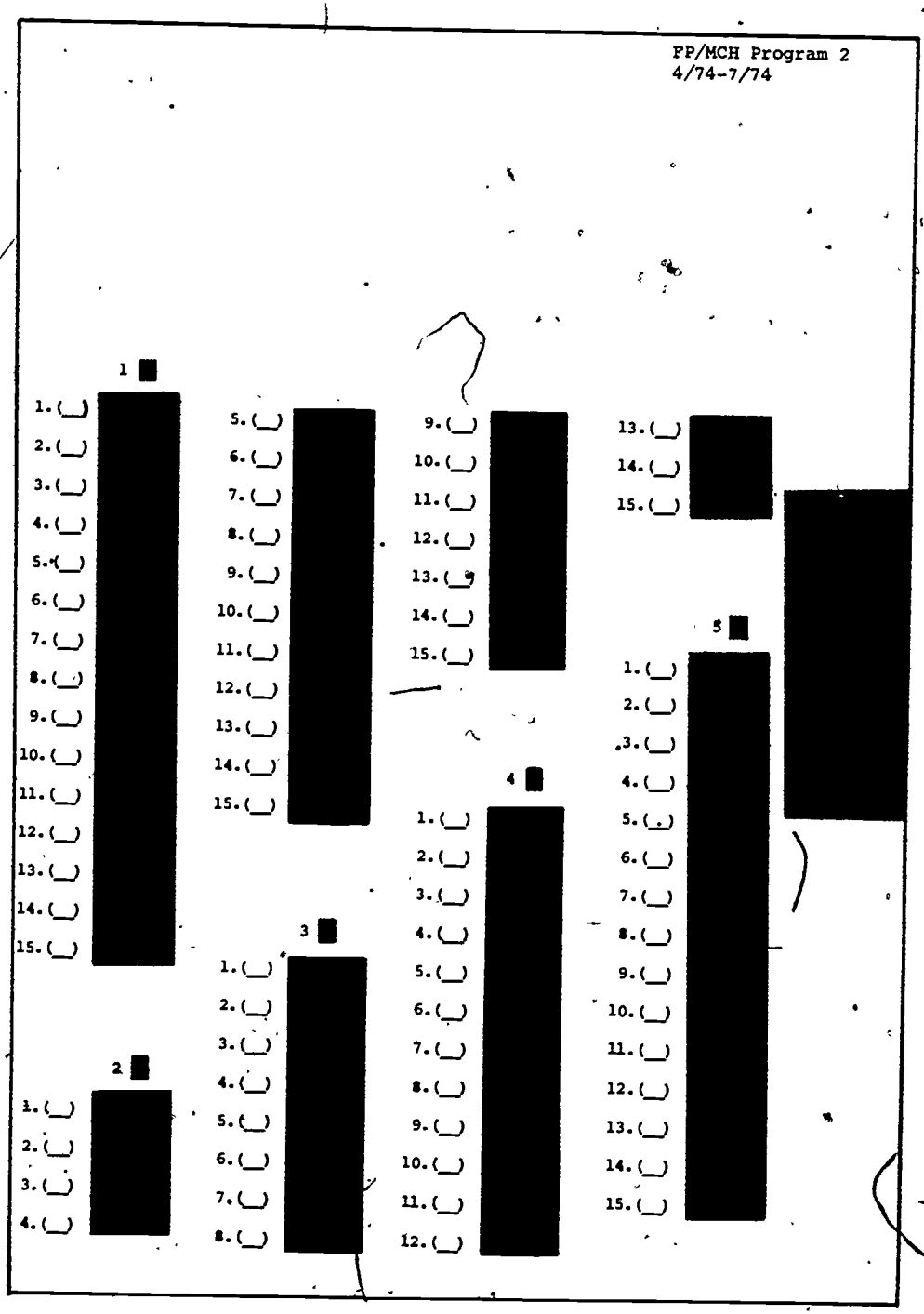
(set 4 cont'd)
13. a ● c d
14. a ● c d
15. a ● c d

SET 5
1. a b c ●
2. ● b c d
3. a ● c d
4. a ● c d
5. a ● c d
6. a b c ●
7. a b ● d
8. a ● c d
9. ● b c d
10. a ● c d
11. a b ● d
12. a ● c d
13. a b ● d
14. a b ● d
15. ● b c d

- END -

FIGURE 4

MODEL SCORING STENCIL II



b) The stencil is placed over the answer sheet so that subset numbers from the two forms line up and the item choices on the answer sheet appear next to their item numbers on the stencil.

c) Subset and composite scores are then calculated and transferred to the score summary column provided on the answer sheet. These scores will subsequently be transferred to a profile sheet which will display the scores for all examinees.

The Score Profile

Two sets of data will provide the input for the statistical analysis. The first set is made up of the responses of each examinee to all the items on both administrations of the test. This data in the form of an Item/Examinee data matrix is obtained from the individual answer sheets.

The second data set consists of the Test and Retest scores, both composite and item subsets. To facilitate the non-computer analysis of score data, some type of score summary or profile form should be employed. This is simply a large sheet containing the score distributions for each examinee (identified by an ID code). Having all scores for all examinees displayed on such a form eliminates the necessity of having to manipulate a large number of answer sheets (i.e., 2 X the number of examinees) when analyzing score data, a situation conducive to quantitative error and unnecessary expenditure of time and labor.

Score profile sheets are not difficult to construct and can be based upon any number of formats. An example of one type of profile, with sample test scores, is shown in Figure 5. The value of using a score profile such as the one in this example will become apparent in a later discussion on non-computer analysis of score data. It should be noted that the value of such a device is reduced if errors occur in transferring scores from the answer sheets to the profile. Whenever data are being transferred from one form to another, the forms should be checked by another individual to ensure reliability of the data.

Processing For Hand Scoring

1. The first step in processing is to assign identification codes to the response sheets. The response sheet for <u>each</u> examinee should have the following basic information:

FIGURE 5

SCORE PROFILE WITH TEST/RETEST SCORES
FOR 31 EXAMINEES ON A 113 ITEM TEST

FP/MCH Program 1 (11/73-12/73)

EXAMINEE ID	ITEM SET 1 T1	R1	ITEM SET 2 T2	R2	ITEM SET 3 T3	R3	COMPOSITE ΣT	ΣR
01	28	34	23	32	9	12	60	78
02	30	42	23	29	4	6	57	77
03	13	19	14	17	3	2	30	38
04	26	43	25	36	9	9	60	88
05	23	26	29	33	5	11	57	80
06	29	37	30	36	7	7	60	80
07	33	33	25	34	7	12	65	79
08	23	31	21	28	6	6	50	65
09	28	32	22	24	9	8	59	64
10	29	34	24	30	11	15	64	79
11	25	34	21	28	6	6	52	68
12	13	29	28	29	4	9	45	67
13	18	35	27	30	10	9	55	74
14	22	35	21	30	6	9	49	74
15	24	36	26	29	6	8	56	73
16	16	36	19	30	6	10	41	76
17	31	41	28	39	8	10	67	90
18	20	36	25	30	4	9	49	75
19	19	36	23	32	6	12	48	80
20	21	41	23	31	7	6	51	78
21	29	37	24	33	9	13	62	83
22	8	31	14	29	6	8	28	68
23	27	35	16	33	7	9	50	77
24	29	36	25	32	7	9	61	77
25	28	38	27	30	10	12	65	80
26	23	38	25	33	8	12	56	83
27	27	29	28	30	7	8	62	67
28	29	39	29	38	6	12	65	89
29	18	38	24	28	6	8	48	74
30	22	35	23	33	5	7	50	75
31	26	31	28	33	5	9	59	73

a. An exclusive, two-digit (or more) ID number to identify the trainee.

b. A code to designate whether the responses are from the the Pre-Test or Post-Test (e.g., 1 for Pre- and 2 for Post-Test).

c. The date of administration to identify the specific course attended by the trainee. (If more than one course is being conducted during the same time period, each course should be assigned a code number and the appropriate one written on the sheet.)

In certain situations it may be necessary to add further identification information to the response sheets. For example, the training staff may wish to evaluate the effectiveness of instruction on various training subgroups (e.g., professional vs. paraprofessional health workers). In this case, a code which classifies the trainee into one or another subgroup would also be entered on the response sheet.

2. Each answer sheet should be visually scanned to detect any items with more than one answer marked. This is necessary prior to scoring with some stencils, since only the correct answer will appear through the punched hole. Multiple-answered items are to be regarded as incorrect and should be eliminated from scoring. A colored line drawn across the response choices for those items will show through the stencil, indicating to the scorer that the item is to be discounted.

HAND SCORING PROCEDURES

General Scoring: Item Set and Composite Scores

When deriving the item set and composite scores, the hand tabulating method used should result in error-free scores with a minimum amount of staff time and effort.

The employment of a scoring stencil and separate answer sheets has been suggested to facilitate scoring with reduced errors. In a further effort to reduce the chances of error and confusion when processing individual item response protocols, a systematic procedure should be used in tabulating and recording scores. The approach recommended is as follows:

1. Tabulate each item set score separately and record the

score in the place provided on the answer sheet (see Figure 2, p. 40).

2. Sum all item set scores to derive the total (composite) score and record it on the answer sheet.

3. When all scores have been tabulated and checked, transfer each one to its appropriate space on a score profile form.

4. Repeat steps 1 - 3 for each trainee's answer sheet.

The Test score data that are to be used (along with the Re-test scores) in the sample analysis runs illustrated in Chapter VII are displayed on the Score Profile Form in Figure 5. With scores displayed in this manner, summary statistics (i.e., means and standard deviations) can be easily computed for the Pre-Test scores (and later for the Post-Test scores). These statistics can be useful when discussing the trainees test performance, since they summarize the masses of score data and provide a concise description of over all pre-training (and post-training) levels of subject matter competence. The importance of carrying out separate analyses of Pre- and Post-Test data, immediately after their respective administrations, in addition to the combined Pre-/Post-Test data analysis, will be discussed further in subsequent sections of the Manual.

Scoring by Trainee Subgroup: Item Set and Composite Scores

It should be noted that on the Score Profile (Figure 5), the 31 trainees were considered as a single group when the item set, composite scores and summary statistics were computed. An alternative approach is to conduct the score analysis on subgroups of trainees, determined on the basis of one or more relevant parameters. Some factors that could be used to define subgroups are the results of Pre-Test scores (high, medium, and low scoring subgroups), or the educational/experiential backgrounds of trainees (e.g., professional vs. paraprofessional groupings). The purpose of such groupings would be to determine whether the training has greater or lesser effectiveness with one group than with others. That is, does a specific factor such as initial competence or professional background have some influence on the degree to which the training instruction is successful in meeting its objective of increasing subject competence? The decision to subdivide trainees and the selection of factors that would underlie any subdivisions will be made by the training administrators according to training objectives and composition of the group.

If the decision is made to put trainees in subgroups, the end of instruction Pre-/Post-Test analysis of item set and composite score distributions should be conducted separately for each subgroup so that inter-group comparisons can be made. In such cases, it is recommended that separate profile sheets be used for each trainee subgroup, using the same format as the Profile in Figure 5, and employing distinct ID codes to identify the various subgroup profiles.

COMPUTER PROCESSING

Editing

1. Same as processing step 1 for hand scoring (p. 53).

2. Items with more than one answer marked are to be eliminated from scoring by drawing a line across the item on the answer sheet. (When punching, the card columns corresponding to these items will remain blank.*)

3. All item responses should be number coded and transferred to 80 column key punch coding forms. (See Figure 6 for a sample coding form.) Punching should be done from these coding forms and not directly from the answer sheets, to reduce the possibility of error.

Coding

Systematic coding procedures must be employed since considerable data manipulation is involved, presenting numerous possibilities for error. The following sequence outlines the procedures required to convert the raw data (individual item responses) to coded data for use in punch card processing.

The format described is required for data that is to be computer-analyzed under the specific system of scoring and analysis programs presented in the Manual. (See Appendix E).

1. a) When the response alternatives are listed alphabetically on the answer sheet, it is necessary to convert

---

* Unless a standard Item Analysis of the Response Data is to be conducted (see pp. 121-127) most Item Analysis procedures require a code for differentiating between the types of possible responses given to an item (i.e., the correct, incorrect and multiple response and no response.) A suggested code is provided on page 56.

responses to their numeric counterparts (a=1, b=2, etc.). The converted numbers designating an answer to an item should be written in to the right of that item. (Note: For Item Analysis purposes (see pp. 121-127), no response given should be coded "0" and multiple responses, which are to be considered incorrect, should be coded "9".)

b) When the conversions have been completed for all answer sheets, each of the sheets should be checked for conversion errors. If possible, checking should be done by a second person.

2. The response choice numbers should then be transferred to 80-column coding forms. It is assumed that those planning to use key punch and computer facilities will have ready access to standard coding forms. If this is not the case, a simple form can be constructed using the sample form displayed in Figure 6 as a model.

Punch Card Data Format

In order to run the data with the programs provided in Appendix E, it is necessary to transfer the data to the coding forms according to a specific punch card format. That is, data are to be displayed on the coding forms so that after punching there is a card (or cards, depending on the number of items) for each trainee containing his responses to all items.* The first card for each trainee must contain an ID number. When more than one card per trainee is required, these cards should also contain ID numbers to identify each trainee's cards should the data deck become disordered.

The data transfer from answer sheet to coding form should proceed according to the following punch card format (each 80-column row on the coding form equals one punch card)

---

* A more complete description of the computer input data formats will be found in the documentation for Program COMSCOR in Appendix E.

Data card format (for each trainee):

Card 1
Columns 1-2 : trainee ID number (required).
Columns 3-80 : item responses. If total number of items exceeds 78, extra cards will be needed.

Cards 2-4
Columns 1-2 : trainee ID number.
Columns 3-80 : item responses.

After the last trainee's set of responses has been recorded on the coding form, each set should be checked for possible errors. (Again, it is suggested that cards be checked by a different person, if possible.) A sample coding form displaying the responses of four trainees to a 113 item test (from the Rennes Assessment Study) is shown in Figure 6.

Keypunching

The only requirement for punching the response data is that a keypunch comparable to the IBM 029, which employs an EBCDIC code, be used. The FORTRAN IV programs presented in Appendix E require that data decks be punched according to this code. It is suggested that this requirement be discussed with available computer personnel before the data is punched.

Since all of the numeric data for the complete statistical analysis will be contained in the Pre-Test and Post-Test item response decks, all appropriate procedures for data verification should be employed during the punching stage.

In order to avoid confusion during the analysis stage, each response deck should be designated as either Pre-Test or Post-Test. Such identification can be written (with an ink marker) either on the front and back data cards or on the top and bottom edges of the card deck.

COMPUTER SCORING PROCEDURES

General Scoring: Item Set and Composite Scores

The computer scoring of item responses by separate item sets and composite scores will be carried out by Program COMSCOR, using as input the punched card data decks, the preparation of which was described above. The program will compute scores and summary statistics for each item set as well as for the item composite, treating the trainees as a group. A complete description of the requirements for and capabilities of Program COMSCOR is provided in the documentation for the program,

DATA CA

Health and Family Protection - 11/8/73.

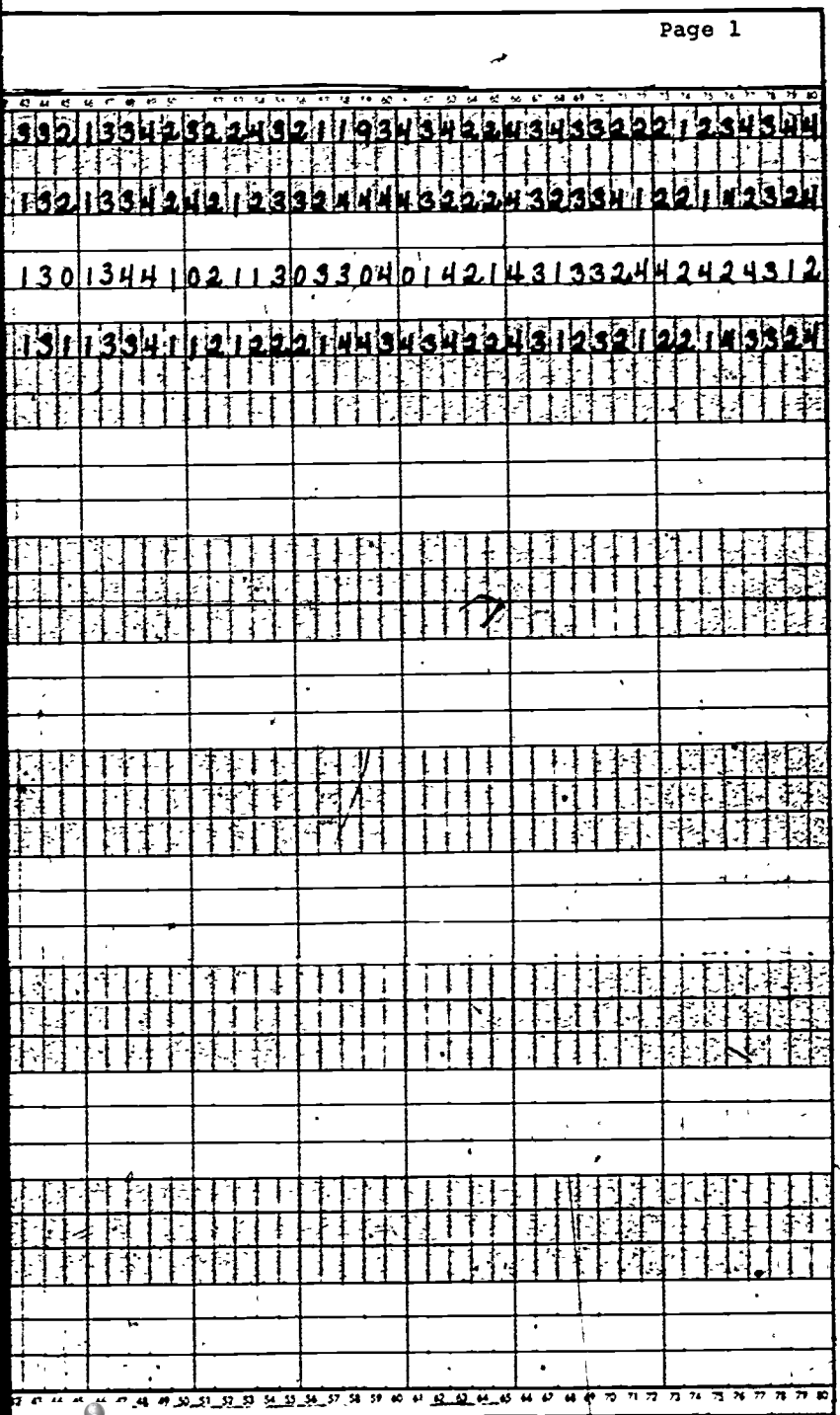
Pre-Test Item Responses (31 trainees x 113 items)

0	1324	4	2311	21134	44234	32343	21213	33333	2423
0	1322	4	3313	42213	31123	42123	42211	44343	24
0	2423	4	2311	21134	24214	34233	21413	33323	2123
0	2242	4	3313	41214	22244	42123	11332	11412	62
0	3403	4	4412	31320	01204	43331	11410	04333	4434
0	3011	2	3303	42112	33140	44332	11444	21466	52
0	4313	4	2312	21134	24214	43333	23413	33323	2123
0	4221	4	3313	42214	32124	22123	41312	44414	34

6

DING FORM

Page 1



in Appendix E.

Scoring By Trainee Subgroup: Item Set and Composite Scores

A variation in the computer scoring procedure will be necessary if the trainees have been divided into subgroups, and inter-group scores and item response analyses are required. The input data deck would be divided into subdecks, each subdeck corresponding to a subgroup of trainees, and serving as data input for analysis of subgroups. For example, with high, medium, and low score subgroups, the program (COMSCOR) would be run three times, resulting in three separate item set and composite score printouts. With this data available, it would be possible to make intergroup comparisons to check for any differential effects of instruction due to differences in group characteristics. Separate subgroup runs for all other computer programs comprising the analysis package would also be carried out.

Timing of the Scoring: the Pre-Test

Although it is possible to defer the scoring of the responses to the Pre-Test until the end of the training sequence, running both Pre- and Post-Test analyses together, it is recommended that the Pre-Test be scored as soon as possible after its administration. While a comprehensive discussion of this issue is presented in the next chapter, it should be noted at this point that the scoring data derived from the Pre-Test can be valuable to training administrators and instructors, by giving them information to help guide the course of instruction, to make it more responsive to the needs and demands of the current trainee group.

CHAPTER IV

UTILITY OF THE PRE-TEST

In terms of assessing the effect of training on subject matter competence, the Pre-Test is used to provide pre-instruction baseline levels of competence against which the post-training levels are to be compared. However, the Pre-Test also has independent utility as a means by which the proposed course of instruction can be assessed and modified (if necessary) to meet certain trainee needs and demands.

Before becoming involved in a discussion of the first stage of data analysis, it is necessary to consider the three basic assumptions which underlie the evaluative design.

The importance of establishing the validity of these assumptions rests on the fact that the degree of confidence ascribed to any inferences concerning the effectiveness of training is based upon the degree to which the assumptions are considered tenable.

Assumption 1. The test item content provides a representative sampling of the subject content comprising the course of instruction.

Assumption 2. All items were properly constructed and critically reviewed (with the necessary revisions made) in accordance with the rules and guidelines presented on pp. 28-31 and Appendix C.

Assumption 3. The Pre-Test scores of the trainees should tend to be quite low across subject areas. That is, the incoming trainees should not, as a group, be highly competent in the subject material comprising the training. There would be little value in constructing a training program with students who were already competent in the material to be covered. (The following section will discuss situations in which this assumption does not hold.)

One way in which an analysis of the Pre-Test scores can be of value is as an indicator of whether or not the course material (as outlined in the curriculum plan and as represented by the test item content) is appropriate (or adequate) for the current group of trainees. Consider, for example, a situation in which a majority of the trainees score very high (e.g., 80%+) on one item set and do as expected (see assumption 3, above)

on the remainder of the test. Since the pre-instruction competence in one of the subject areas to be covered is already quite high, it would be a waste of time, both for the training staff as well as the trainees, to present the course with the same subject material as outlined in the original curriculum plan. It would be necessary to effect some revisions in the proposed curriculum to make the instruction more relevant to the demands of the group. This can be easily done by adopting one of the following procedures:

1. Consider the trainee group as competent in the high score area and drop the subject area completely from the curriculum, devoting all of the training time to a coverage of the low-score areas. This might be done in situations where the scores in one subject area are so high that the instructors feel that little more could be learned by the trainees in that subject area.

2. Keep the subject area in the curriculum, but shift emphasis away from the high score area toward a more intensive coverage of the areas in which the trainees displayed low pre-instruction competence.

3. Keep the subject area in the curriculum, but upgrade the instruction to a more advanced (and possibly more difficult) coverage of the subject area. That is, a segment of instruction would be devoted to coverage of subject material in an area that required an already basic level of competence. (This is similar to the practice of using the introductory course or an advanced placement exam as the prerequisite to the intermediate and/or advanced level courses in a specific area.) In some subject areas, it might not be possible to upgrade the level of instruction so that it would be better to drop the area from the curriculum and concentrate on the other areas (see option 1).

While it is possible for the trainees as a group to score very high in certain subject areas and low in others (the type of situation described above), another type of situation is more likely to occur. This would be one in which different individual trainees or trainee subgroups displayed varying degrees of competence in different subject areas. Such a possibility might exist since (in many training situations) the trainees have dissimilar educations, professions, or experience. Such heterogeneity is quite likely to manifest itself on an achievement test through variations in scores both for the individual trainee as well as for one

or more trainee subgroups. When the instrument is administered, as a Pre-Test, to such a heterogeneous group, the picture that is likely to emerge is one in which a few trainees obtain consistently high scores, some get consistently low scores and the majority displays wide variations in the pattern of high and low scores. Such a high-low profile of score "scatter" can be used to determine which individuals or groups of trainees require special instructional attention and the subject areas in which such attention is required. This can be done by considering the scores for each segment of instruction (i.e., as represented by the separate item sets) independently. On the basis of each group of item set scores, the trainees can be grouped (using some pre-defined set of cut-off scores) into 'high,' 'medium' or 'low' scoring categories. Special attention could then be provided to those trainees in the 'high' and those in the 'low' score categories. This special attention, commensurate with competence level of each of the subgroups, could take the form of small group and, when necessary, individual trainee tutorials as well as outside supplementary readings and/or projects.

The grouping of trainees, in addition to helping adapt the level of instruction to the needs of the trainee subgroups, can also serve an important evaluative function by adding another level of evaluation to the already existing stages. Up to this point, emphasis has been upon an analysis of test data for the total group and for the individual trainee. With the total group now subdivided into 'high,' 'medium' and 'low' pre-scoring categories, it will be possible to make a quantitative assessment of the relative impact of instruction in terms of the trainees' incoming level of competence. Since the possibility exists that a trainee's relative increase in competence is not only a function of the effectiveness of training, but also a function of the trainee's initial (i.e., pre-course) level of competence, an analysis of scores by trainee subgroups might prove valuable.

In order to carry out the analysis at this level it would be necessary to rank-order the trainees on the basis of Pre-Test scores from highest to lowest score. The trainees would then be split into 3 groupings of equal (or near equal) size and the groups labelled as 'high,' 'medium' and 'low' scorers. The Test/Retest data for each subgroup would then be subjected (separately) to the type of analysis described in Chapter VII. The final assessment of training effectiveness would be based on the pooled findings of the analysis of data for each subgroup together with a series of inter-group comparisons of trends.

It should be noted at this point that subdividing the trainees into scoring categories for purposes of individual subgroup analysis should not be done unless the pre-score data warrant such a breakdown. For example, if after ranking, only a few score points separate the highest from the lowest score, then it would not be of value to conduct the analysis of subgroups. It is suggested that separate subgroup analyses be conducted only when the range is wide enough (e.g., 20 or more points, with the trainees more or less equally distributed along that score range) to help ensure a moderate amount of score difference between subgroups. A close grouping of Pre-Test scores would not permit the subdivision of trainees into subgroups that display discriminable differences in pre-instruction levels of competence.

Another situation that could occur is one in which the trainee group obtains an extremely low mean score (i.e., one approaching zero) on the Pre-Test. The question that arises is whether or not the curriculum level is unreasonably high and beyond the capabilities of this group, given the amount of instruction time available. Opinions in this respect would be influenced by indications that the trainee group has (or does not have) a basic understanding of prerequisite terminology, concepts, history, etc. The decision would then have to be made as to whether or not the course level of instruction is to be lowered and, if so, to what degree. A similar question and decision would pertain also to any trainee subgroup who fell markedly below the Pre-Test score levels of the others.

NOTE: If, on the basis of data provided by the Pre-Test, the curriculum and, thus, the course of instruction is revised, it will also be necessary to make revisions in the content and/or structure of the test instrument and the subsequent Test/Retest analysis. The effect on the evaluation of training developing out of these revisions will be taken up in the discussion on the utility of the Post-Test (see Chapter VI).

CHAPTER V

THE CURRICULUM AUDIT

Definition and Purpose

On pages 28-31 (and Appendix C), guidelines are given for facilitating the construction of test items (prior to the beginning of instruction) using only the information provided by the curriculum plan and the list of specified learning objectives. It was shown that the learning outcome being sampled by a test item is defined by the specific grammatical and semantic structure of that item. Thus it is possible to ensure, at the time of test construction, that the items represent a valid sampling of the types of behavioral objectives (i.e., specific learning outcomes) that the training program attempts to achieve. However, it is not possible to determine during the test development stage whether or not the item content is a true representative sample of the types of subject material actually covered in training. Therefore the issue of content validity requires further consideration.

In the Test/Retest design, the testing instrument must be constructed before the sequence of instruction has actually been carried out. Unlike the tests of learning common to the school and college classroom where test items are usually written after the subject material has already been covered, the Pre-/Post-Test instrument consists of items developed exclusively from the curriculum study plan. The validity of this procedure rests on the assumption that what is outlined in the study plan will actually be presented to the trainees during the course of instruction. This assumption is warranted with certain types of formal instruction: an example is the basic or introductory course where there is a well-defined core of specific material that must be covered (e.g., algebra, descriptive statistics, English grammar, etc.). In this case the course content generally remains the same, regardless of the teaching approach used or the instructors involved.

Although there probably are training programs of long standing whose curriculum plans are well-established and accurately reflect the structure and content of the actual course of instruction, this is not generally the case. Training programs are directed toward preparing individuals for some specific vocational objective. Once training begins, it is not unusual for the instructor(s) to alter the subject content (as outlined in the curriculum plans) to meet the needs and interests of the trainee group and to conform to the general level of subject competence as indicated in a Pre-Test.

Further, unlike well-established courses where both the subject content and the criteria required for subject competence are highly structured, the training program curriculum is often less well-defined and usually undergoes some modification each time the sequence of instruction is presented. Regardless of the cause, there is a high probability of disparity between subject topics proposed for study as described in the curriculum plan, and the actual areas covered during instruction. This discrepancy may result in lower content validity of test items developed prior to the training, thus weakening the inferential powers of the test results. Therefore, the conclusions drawn from the statistical analysis of scores must be weighed against an estimate of the degree of concordance between the content of instruction and the content of the test items. This estimate is best obtained from a detailed study of the content of instruction as it is actually presented, through a Curriculum Audit.*

The basic requirement for such an audit is a systematic procedure for rating the degree of coverage given during each training session with relation to the specific subject areas sampled by each test item. This is best accomplished by the use of a content checklist, a listing of test items classified by subject content. An explicit statement defining each item's content can be provided by the test constructor from his test blueprint (see pp. 12-16). A checklist is then made by constructing a two-way grid with the item content list on the vertical axis and the horizontal axis labeled with the degrees of coverage (i.e., complete/partial/not covered). An example of this type of checklist is provided in Figure 7. Although it is designed for a hypothetical statistics training sequence, the basic form can easily be adapted for any type of training. (The checklist is incomplete since the numbers and types of items required to provide adequate coverage of such a training sequence are too numerous to illustrate here.)

Concurrent and Retrospective Auditing

There are essentially two ways to conduct a curriculum Audit; namely, concurrently or retrospectively.

1. Concurrent Auditing as the name implies, is an assessment of concordance conducted on a session-by-session basis while the training is in progress. This approach

---

* Developed by S.M. Wishik for the evaluation of Maternal and Child Health courses he conducted at the University of Pittsburg.

FIGURE 7

SAMPLE ITEM COVERAGE CHECKLIST

quant. meth. prog. (stat.) 10/72-2/73		DEGREE OF COURSE COVERAGE		
ITEM #	CONTENT AREA	COMPLETE	PARTIAL	NOT COVERED
1	nominal, ordinal, interval & ratio measurement scales			
2	grouped frequency distributions			
3	cumulative frequencies & distributions			
4	percentiles			
5	percentile ranks			
6	measures of central tendency: mean, median & mode			
7	selecting the appropriate measure of central tendency			
8	relationship between central tendency measures & the shapes of distributions			
9	measures of variability			
10	mathematical operations with the variance & standard deviation			
11	the concept of the random sample			
12	characteristics of a sampling distribution			
13.	sample statistics as estimators of population parameters: mean & standard deviation			
14	normal distributions			
15	use of the tables of the normal distribution			
16	statistical hypotheses			
17	the problem of error in hypothesis testing			
18	statistical inference: selecting the appropri... statistical ...			

requires that an auditor sit in on every training session. (The auditor should be someone other than the instructor who is familiar with the nature of the subject material.) There are two approaches to conducting the concurrent audit:

a) Employing the Item Coverage Checklist

The auditor should check the training syllabus (i.e., the curriculum plan) for each session to identify the specific items on the checklist whose subject areas are scheduled for discussion. (The items should be listed in the sequence in which the subject areas they sample are to be covered during instruction.) During the session he will record on the checklist, next to those items, whether or not those subjects were, in fact, presented and the degree to which they were covered.

Complete and/or no coverage of a content area can be designated by placing check marks in the appropriate cells adjoining that area on the checklist. A brief sketch of what was and was not covered should be recorded, however, when partial coverage of a content area is to be noted.

This procedure is conducted for each of the training sessions. At the end of training there will be a record of the degree of coverage of subject areas sampled by all of the test items. The data from the checklist can then be reduced to a table of summary tabulations (see example, p. 73) for use in the interpretation of the statistical analysis of test results.

b) An Alternative to the Item Checklist

The training seminar is designed as a structured learning experience for the participants. Seminars are conducted according to predetermined curriculum plans which structure each sequence of instruction in terms of the specific subject content to be covered as well as the sequence and mode of presentation of that content.

In order to maximize the learning opportunities, it is often desirable to encourage the two-way (instructor $\rightleftharpoons$ participant) over the one-way (instructor $\rightarrow$ participant) instruction model. When this type of interactive flexibility is emphasized, it will not always be possible to maintain high concordance between the subject content of the seminar

and the content outlined in the curriculum plan. When the sequence of instruction has been altered, the subject material reflected by the test items may not necessarily be covered in the session planned, but at a later (and possibly unspecified) time. When there is a strong probability that such alterations in sequence will occur, a variation in the Concurrent Audit as described above, should be implemented.

An auditor would still attend each session but the Item Coverage Checklist would not be used. Instead, a brief "subject coverage-by-session" table would be constructed by the attending auditor. That is, instead of employing a pre-constructed Item Checklist (as described above) the auditor would compile a content outline for each session attended.

The auditor would list, in sequence, the major subject areas (by topic and subtopics) together with a brief description of the content subsumed under each major heading. Then, for each major topic, he would describe the degree of content coverage in terms of substantive variables, such as amount of time devoted to topic, list of concepts discussed, types of examples presented and/or computations carried out, nature and number of student exercises provided, etc.

A table, derived from a statistics seminar, which displays a sample entry for a training session covering measures of central tendency is in Figure 8.

Each description of topic and coverage should be brief and concise so that the auditor can spend the major portion of each session attending to the subject presentation, not filling in the tables. In most cases, a few key words can be used to describe the topic areas together with a concise paragraph summarizing the coverage activities. However, enough information should be provided so that the tables are self contained (i.e., self-explaining) summaries of each training session both in terms of topics and instructor/participant activities.

Upon completion of the final session of the seminar, these tabular summaries would be compared with the items comprising the test instrument. The content of each item would be compared with the relevant topic listed in the summary to assess the degree of concordance between subject content of the item and

FIGURE 8

SAMPLE SUBJECT COVERAGE-BY-SESSION TABLE

statistics training seminar - 11/16/72)	
topic by session	degree of coverage
	exercises — How to analyze data yielded by the various measurement scales
SESSION 3 Measures of Central Tendency * Descriptive and operational definitions of mean, median + mode * Computational requirements for each measure — grouped + ungrouped data * How and when to apply each measure.	(120 min.) Complete description of each measure — computational formulas with calculations of each measure provided (both for ungrouped + grouped frequency distributions). Training exercises conducted — selecting appropriate measure for different types of data with justifications for selection; sample computations of each measure. Exhaustive question + answer period — complete coverage of the topic area.
SESSION 4 Measures of Dispersion * Description and operational definitions of range, semi-interquartile range, mean deviation, variance and standard deviation. * How and when to apply each measure. * Mathematical operations	(120 min.) Complete description of each measure — applications and limitations of the range, mean deviation, semi-int.

the coverage provided that subject during the session. This adequacy-of-coverage procedure would be conducted on an item-by-item basis. (A broad COMPLETE/PARTIAL/NOT COVERED scale will be adequate for classifying each item on degree of coverage.) This data can then be reduced to a frequency table (see p. 73) for later use in the interpretation of statistical results.

NOTE: Assigned outside readings providing relevant subject material not covered during the training sessions should also be noted during the audit. Simply recording on the audit form used the specific readings assigned, with a brief description of content, will be sufficient. Test items covering proposed outside material can later be checked with the audit to determine if the material was actually assigned.

2. Retrospective Auditing When it is not feasible to conduct the audit during the course of training, it can be done after the final session on a retrospective basis. This approach is, essentially, a content analysis of the instructor's training log describing the proceedings of each training session. Many instructors maintain detailed accounts of subject coverage (usually recorded after each session) throughout the course of instruction. If this is not the case, the instructor should be requested to do so, if only for auditing purposes. (The auditors can increase the effectiveness of the audit by providing the instructor with an outline of the types of data they will require.)

   It is possible to employ several staff members to serve as auditors since a retrospective assessment is conducted at a single point in time. The procedures are basically the same as those for the Concurrent Audit except that the analysis is based upon written accounts of each session rather than on direct observation of instruction by the auditor. The auditors check each item on the checklist against the content data detailed in the log to determine the degree of coverage given to the subject area sampled by the item. A table summarizing the results is then constructed (see p. 73) for use in the subsequent test analysis.

Which approach to employ will depend on conditions specific to each training program. Where funds and available staff time permit, a concurrent approach should be used since the auditor can base his assessment on direct observation rather than having

to depend on the written accounts of the instructor. When this is not feasible the Retrospective Audit can be used. With the latter approach, the maintenance of detailed, written accounts of each training session conducted must be made an integral part of the instructor's training schedule.

Using the Results

Regardless of the approach selected, the resulting data will provide information concerning the degree to which the test measures a representative sample of the subject matter content under consideration. While the results of the audit do not enter into the formal statistical analysis of test data, the findings should be considered when interpreting the data in terms of trainee achievement and training effectiveness. For example, if the audit shows that many of the items were only partially covered or not covered at all, there is reason to conclude that the test results are invalid for the subject areas being assessed. If, on the other hand, the audit finds a high degree of overlap between items and actual course content, it can be concluded that the test is assessing those subject areas it was designed to assess.

An illustration of the way in which the results of an actual Curriculum Audit were used in the assessment of content validity is presented below.

Example

The specific training program involved was a 7 week Seminar-Workshop in Family Planning/Population Program Management. An Evaluation Inventory, composed of 85 test items, was constructed to measure trainee competence in three major subject areas. The items were grouped into 3 separate item sets.

> The type of audit employed was retrospective. The subject content of each of the 85 items was checked against the detailed descriptions of subject coverage for each daily training session as reported in the curriculum logbook. The auditing was conducted by staff members who were actually involved in instruction and were familiar with the subject material. Each item was discussed until all the auditors were in agreement as to its "degree of coverage" rating. The resulting data, tabulated from the checklist, are reproduced in the following table:

	DEGREE OF COVERAGE		
	COMPLETE	PARTIAL	NOT COVERED
SET 1-ITEM Nos.	1-9; 12; 14-20; 23-28; 30-32	10; 22	11; 13; 21; 29
SET 2-ITEM Nos.	1-28; 30-34	——	29
SET 3-ITEM Nos.	1-12; 14-19	13	——

TOTAL - COMPLETE: 77
PARTIAL : 3
NOT COVERED: 5

Less than 10% (.094) of the 85 items were rated as not or partially covered by the course instruction. The results justified a subjective assessment of moderate to high content validity. It was thus concluded that the Inventory items provided a representative sample of the relevant subject areas covered in the Seminar-Workshop.

The "not covered" items were omitted from the subsequent scoring and the analysis of test results. This procedure was acceptable as the number of uncovered items was small and the total number of items sufficiently large. It is suggested that when the audit indicates a moderate to high degree of validity, the "not covered" items should be omitted since they will contribute little to an assessment of what was learned in the training program.

If a large number of items is in the "not covered" or "partially covered" categories, it will probably be necessary to abandon the Test/Retest design and write a new test to be administered at the end of instruction. This test can be made using the Curriculum Audit itself as the subject area checklist and drawing upon it for new items. (See pp. 74-80 for a discussion of the Post-Test, and its uses apart from those specific to the Test/Retest assessment).

CHAPTER VI

UTILITY OF THE POST-TEST

The Post-Test, like its Pre-Test counterpart, serves a function beyond that of assessing the effect of instruction on the levels of trainee competence. The Post-Test data, taken alone, provides information that is similar to the data derived from the type of test given in most academic situations. That is, it will indicate an individual's level of competence in one or a number of specified subject areas.

If a group of trainees is preparing to assume job positions where high levels of competence in the subject matter covered in training will contribute greatly to their "on the job" success, then the trainees' performances on an "end of instruction" test will provide one indication of their job readiness.

Like most "end of instruction" testing situations, the Post-Test results alone will indicate only how competent each of the trainees is in each of the subject areas being assessed. Without the baseline data derived from the Pre-Test, the relative contributions of the sequence of instruction and various pre-training experiences to the competence demonstrated by the performance of the trainees on the Post-Test, cannot be assessed.

Although the use of the Post-Test alone is not recommended when the objective is to assess the impact of instruction on levels of competence, certain situations will arise (to be discussed below) in which the only objective data available are those which were derived from administration of the test instrument at the end of instruction.

Previously (pp. 61-64), the point was made that it will sometimes be necessary to make revisions in the proposed sequence of instruction. Such revisions will be called for when the results of the Pre-Test indicate that the proposed curriculum is not appropriate for the current group of trainees.

Since changes in the course of instruction involve changes in subject matter coverage (see pp. 61-64), the assessment of trainee competence might be directly affected. This is due to the relationship between the course material and the objective test instrument. That is, there is a high probability that since the test items were constructed from and reflect the content of the proposed sequence of instruction (as

originally outlined in the curriculum plan), any change in emphasis on the material actually covered will create a disparity between the item content and the course content. The degree of disparity will determine the extent to which the testing instrument (administered under a Pre-/Post-Test design) is, or is not, valid for assessing changes in trainee competence in the subject matter of instruction.

Situations Requiring Post-Test Revision

An illustration of the type of testing situation that can arise which necessitates varying degrees of revision in the curriculum will help clarify this important issue. Consider the following premise:

> On a test instrument assessing competence in 4 subject matter areas, a group of trainees obtained mean Pre-Test scores of 94% on Item Set 1 and 15, 23, and 31% on the other three, respectively. The assumptions (stated on p. 61) are held to be valid in this case. Thus, there is evidence for concluding that the high Item Set 1 mean score indicates a very high level of competence among the total trainee group in the subject material being assessed by that item set.

Based upon the above premise, some representative guidelines for modifying the assessment design in light of necessary curriculum revisions can be provided (taking into account the general suggestions for making curriculum revisions stated in an earlier discussion, see pp. 61-64). Since the Pre-Test items cannot be revised, the emphasis will be upon changes in the Post-Test items and/or revisions in the overall analysis of the Pre-/Post-Test data.

1. The content of instruction in the subject area assessed by Item Set 1, when possible, can be upgraded to a more advanced, complex level of coverage (see option 3, p. 62). If this is done, it will be necessary to construct new Set 1 test items to cover this more advanced material. These items would then replace those items that were answered correctly by most of the trainees on the Pre-Test. These new items, together with those original items which most of the trainees missed on the Pre-Test would constitute a <u>new</u> Item Set 1 to be administered on the Post-Test.

   In terms of analysis, the entire set of Test/Retest statistical procedures, as outlined in

Chapter VII would be applied only to the trainee responses to Item Sets 2-4. Since the Item Set 1 of the Pre-Test will now differ from that of the Post-Test, no Test/Retest analysis would be conducted for the responses to Item Set 1 (or for the Composite scores since the total Pre-Test and Post-Test do not consist of completely comparable items). Instead, the final analysis of the responses to Item Set 1 would be conducted on the Post-Test administration only. The analysis of Item Set 1 data would consist of an assessment of the number of items correct out of the total number of items, both for the individual trainees and for the trainee group. Since there will be no comparable Pre-Test baseline data for Item Set 1, the analysis will focus only upon assessing the trainees' levels of competence with the more advanced subject material without relating these current levels to the effects of instruction.

Although the Pre-Test and Post-Test responses for Set 1 are analyzed separately and will provide little direct evidence of the impact of instruction on competence levels, both pieces of information can be useful to the overall assessment in terms of trainee achievement. That is, the Post-Test data will show how competent the trainees are in advanced subject matter while the Pre-Test results (for Set 1) display their competence in subject material at the level which administrators initially considered appropriate for the trainee group.

Thus, a two step analysis is being effected. The Test/Retest analysis of responses for Item Sets 2-4 will provide the assessment of training effectiveness (and job readiness) while the separate analysis of Pre-Test and Post-Test responses for Item Set 1 will provide the assessment of trainee achievement and job readiness with respect to competence in a specific subject matter area.

2. It will not always be possible to elevate the content of instruction to a more advanced, complex level in subject areas where the Pre-Test results indicate existing high levels of (pre-instruction) trainee competence. This is because the content of instruction assessed by a specific item set might have been initially set at a fairly comprehensive and exhaustive level, making an upgrading of the subject material difficult, if not impossible.

   In certain situations it might be more appropriate either to (1) drop the subject area completely from the curriculum and concentrate the efforts of instruction on coverage of the low score areas (option 1, p. 62); or (2) retain the subject area but give it less coverage than originally proposed; place greater emphasis on the subject areas in which there were low pre-instruction levels of competence (option 2, p. 62).

   A change in emphasis on the subject matter of instruction may or may not necessitate changes in the item content of the test instrument. This will be determined by the nature of the change in instructional content. This is illustrated below employing the premise as stated on p. 75.

   a. If the subject material tested by Item Set 1, the high score area, is either dropped from the curriculum or given less coverage than originally proposed, more time can be devoted to coverage of the subject materials tested by Item Sets 2-4 without adding new content to the curriculum. Without the addition of new subject matter it would not be necessary to construct new and additional test items for the Post-Test Item Sets 2-4.

      For purposes of analysis, the Pre-Test and Post-Test items would be the same for Item Sets 2-4. The items of Set 1 can either be completely deleted from the Post-Test or reduced so that the Post-Test Item Set 1 consists only of the items in that set which were missed by most of the trainees on the pre-testing.*

---

*An alternative procedure is to retain all of the original items of Set 1 and administer them in the Post-Test. This option will be valuable when the subject matter being tested by Set 1 is dropped completely from the curriculum. In

Since the competence of the trainees in the subject material tested under Item Set 1 was shown to be quite high on the Pre-Test, these scores will indicate the level of competence among trainees without relating these levels to the effects of instruction.

The Pre-/Post-Test analysis would be conducted for each of the Item Sets 2-4 and with the 3 sets combined into a Composite score to assess the impact of the course of instruction on competence.

b. The open instruction time which results when subject material in high score areas (i.e., Item Set 1) is dropped or given reduced coverage, can be employed in a coverage of new subject material added to supplement the original subject content in the areas being tested by Item Sets 2-4. Then, as was done with the content of the original curriculum, objective items would be constructed to test the trainees' competence with the new material.

The original items, together with the new items would be administered as the Post-Test. The original items are to be analyzed separately from the items constructed to cover the new subject material. Therefore, the Post-Test item booklet should be constructed so that the original items are administered in their original sequence and the new items are presented at the end, listed according to item set.

Since the new items will be administered only at the end of instruction, there can be no Pre-Test baseline data for these items. Thus, the analysis of trainee performance with the new items would be limited to a frequency count of the total correct out of the total

---

this situation, the Item Set 1 would then serve as a control to determine the changes that occur in responses to items administered under the Test/Retest design without the influence of a mediating instructional sequence. A separate statistical testing of the Test/Retest data for Item Set 1 would determine if the score change that occurred was, or was not, significant.

number possible on the Post-Test. From this certain conclusions can be drawn concerning the competence of the trainees with the new material without, however, the capacity to relate the levels of competence to a specific learning experience (which in this case would be the sequence of instruction).

The item responses for each of the Item Sets 2-4 would be subjected to the standard Pre-/Post-Test analysis, both as individual item sets as well as a combined, composite test. The results of this analysis would indicate the effectiveness of the training program in raising the levels of competence among the trainees.

The importance of the Curriculum Audit has been illustrated in a previous discussion. When revisions in the original course of instruction are required, the Audit will prove particularly valuable for monitoring changes in the relationship (i.e., the discrepancy) between the course subject material and the item content. The original Audit should be updated with newly constructed items and a detailed assessment conducted to determine the degree to which the items cover the subject matter they were designed to cover. With the deletions, additions, and reduced coverage of subject matter that may occur, it is very important to determine the degree to which the course content presented is actually covered by the items comprising the Post-Test instrument so as to maintain the integrity of the assessment study.

NOTE: The results of the administration of the Pre-Test will not be the only factor to suggest needed change in a proposed course of instruction. Another potential source of change can be found in the structure of the training experience.

The training experience should not consist of a series of instructor's lectures on fixed topics with the trainees as passive assimilators of relevant information. The instructor should be as responsive as possible to the specific educational needs and demands of the trainees.

A responsive educational experience should consist of an interactive balance among lectures, recitations, discussions and question and answer sessions. The various contributions of the trainees can, however, have an effect on the sequence and content of instruction. That is, discussion and questions on the part of the trainees may direct the course of instruction away from the proposed subject area coverage to a more peripheral, but relevant, topic. (Negative factors such as

trainee boredom and disinterest may also dictate a change in subject matter.) Such flexibility in training structure can, in most cases, enhance the effectiveness of the educational experience.

When content changes occur they should be noted and assessed during the Audit and, depending upon the type of revision made (i.e., deletion, addition and/or reduced coverage of subject matter), the test instrument and subsequent analyses procedures should be revised according to the guidelines provided in this chapter.

CHAPTER VII

ANALYSIS AND INTERPRETATION OF RESPONSE DATA

General Considerations

The statistical analysis of testing data can be conducted either by manual/mechanical methods, or by computer, employing the programs provided in Appendix E. The statistical procedures and computations suggested here are not complex and can be carried out easily, using prestructured worksheets and a standard desk calculator. It should be emphasized, however, that hand tabulations and mechanical computations are subject to numerical errors which often go undetected, and can be time consuming, especially when the number of trainees or test items is large.

For those with ready access to computer facilities and personnel, a programmed analysis should be considered. Such an analysis requires less staff time and minimizes the probability of computational error. Regardless of the system employed, the quantitative procedures comprising the analysis will be the same.

Overview

This chapter considers the statistical analysis and interpretive assessment of data obtained from the combined pre-/post-instruction administrations of the test instrument. There are three levels, or stages, of this analysis, each level determined by whether the focus is on subject area scores (item set and composite), individual trainee scores, or individual item responses. The discussion of the analysis process follows this stepwise progression.

For the stages involved with assessing the significance of score changes from Pre- to Post-Test (Stages 1 & 2), several hypotheses are presented, derived from a set of questions which the analysis purportedly will answer. The specific statistical tests appropriate to hypothesis testing in the first stages, as well as the quantitative procedures employed in later stages of analysis, are provided here, together with some basic analytic assumptions and underlying statistical theory. The discussion of each stage will be supplemented with procedural examples employing the Pre-/Post-Test data derived from the second field application of the methodology, conducted at the national

School of Public Health in Rennes, France.* The computational formulas, statistical tables and worksheets, and stepwise procedural guidelines, together with sample results employing the Rennes data, are presented in Appendix F.

The computer programs are fully documented in Appendix E, and therefore will not be given extensive discussion in this chapter. However, as each analysis stage is presented, references to the appropriate programs will be made. A procedural flowchart for a computer-run analysis is given in Figure 9. This flowchart is provided to indicate the temporal sequence in which each of the programs is to be run, as well as to serve as an illustrative overview of the analytical design. Therefore, although developed for computer purposes, the flowchart can also serve as a general analysis plan to guide those employing non-computer means.

A Note on Computer Usage: It will be necessary to determine if a specific computer system can accommodate the FORTRAN IV programs as written or if changes in the programs will be required to provide the recommended output. When considering a computer-run analysis, it is suggested that the programming requirements described in Appendix E be discussed with personnel familiar with the capabilities of the computer system to be used.

ANALYSIS OF DATA

Application of Statistical Tests for the Significance of Pre- to Post-Test Score Increases

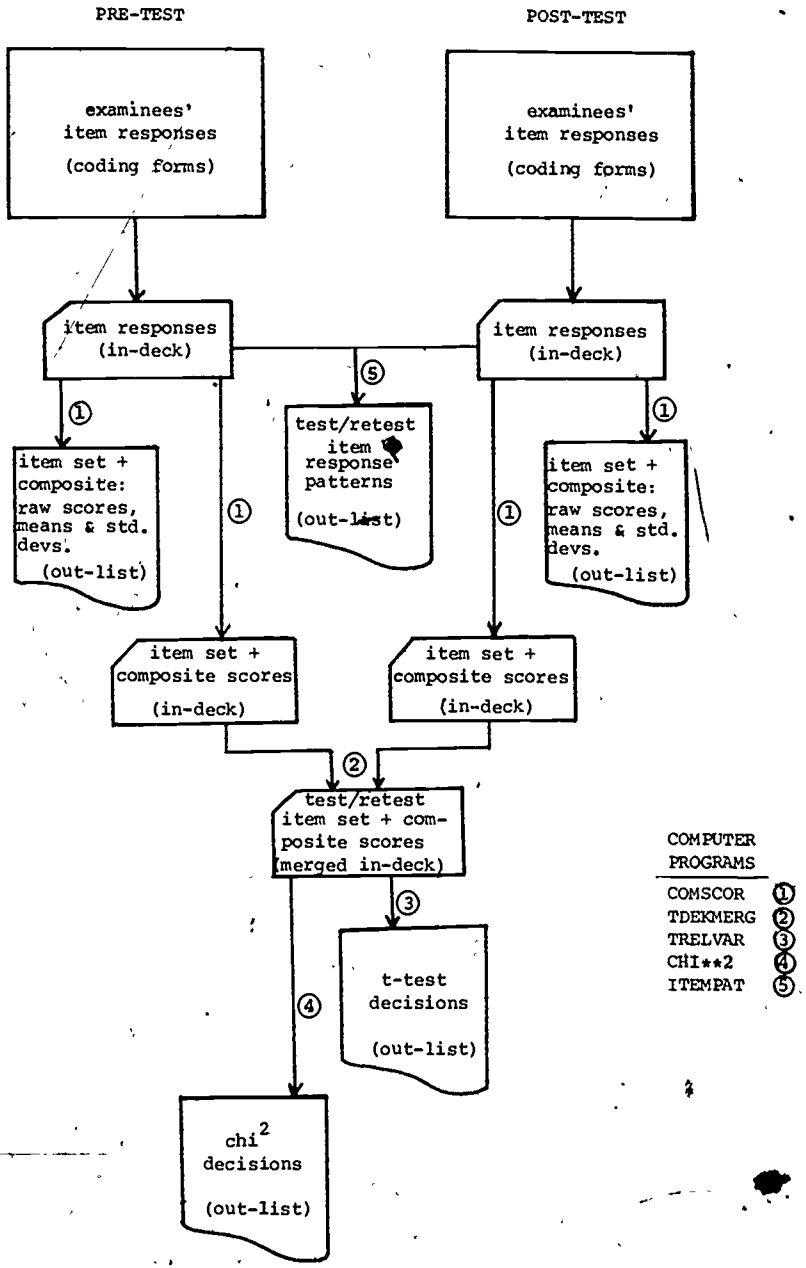
General Considerations The primary questions, the solution of which comprises the first two stages of the analysis, involve whether or not the Post-Test scores have increased in magnitude over their Pre-Test counterparts (on item sets, both for the trainees individually and as a group), and whether or not any of the score increases are statistically significant.

---

* The score and item response data illustrated here were derived from a 113 item subject competence assessment instrument administered under a Test/Retest design to 31 health professionals and paraprofessionals attending a seven week training program in Health and Family Protection. The composite test was subgrouped into three sets of 51, 45, and 17 items. Scoring was on a 1 point per item basis.

FIGURE 9

PROGRAM FLOWCHART FOR COMPUTER ANALYSIS



These major queries can be translated into a general hypothesis that in turn can be partitioned into a set of sub-hypotheses. These are tested by the application of appropriate tests of statistical significance.

In order to make the discussion of specific statistical tests more comprehensible to those not familiar with hypothesis testing and statistical induction, or inference, several important concepts are introduced here. (Those with experience in the application of tests of significance can skip directly to the Stage 1 level of analysis, on p. 89.

1. Statistical Inference A set of trainee test scores (on both Pre- and Post-Test) should be viewed as a sample drawn from the population of all possible scores that would be obtained if the same trainees take the same test an infinite number of times (assuming that on each test administration, the trainees knew only what they knew the first time the test was taken). The mean computed for this hypothetical infinite set of scores is the population mean. Correspondingly, for each of the infinite number of sample scores, of which the actual scores found are one example, a sample mean can be computed.

   While any of the sample means can have the same value as the population mean, only occasionally will this happen. Chance factors such as variations in the success of random guessing occurring across samples will contribute to differences in sample means. Consider now the situation where examinees are administered the same test instrument before and immediately after a sequence of instruction. In all likelihood, whether the instruction had an influence on the Post-Test performances or not, the test scores for the two administrations would have different means. The question to be answered is if the differences in Test and Retest score means are attributable to the intervening effects of instruction or to chance fluctuations of sample means about some common population mean.

   The application of tests of statistical significance will attempt to determine if the observed differences between Pre-Test and Post-Test score means reflect actual differences in the levels of subject competence underlying Pre-Test and Post-Test performances or simply random score fluctuations.

   The specific statistical test employed takes random chance factors into account. If the test results show that there is a low probability of a difference as large as, or larger than, the one observed occurring by chance, then the difference is said to be statistically signficant.

Depending on the type of research hypothesis that was set up as an alternative to the null hypothesis, the rejection of the null hypothesis will provide the basis for inferences concerning the effect of instruction on levels of subject matter competence.

2. The Null and Research Hypotheses The null hypothesis (H_0) is what is tested by the application of tests of statistical significance. In order to determine whether the Pre- and Post-Test means are significantly different from each other, the strategy employed is to test the hypothesis that the means are derived from two random samples drawn from the same population.

   For the first stage of analysis (see p. 89), the null hypothesis will state that training efforts had no effect on increasing cognition competence. The hypothesis will be tested by comparing the Pre- and Post-Test score means.

   For the second stage (see p. 90), the null hypothesis states that the individual trainee's level of competence was not increased as a result of training. (Operationally, the hypothesis being tested is that the distribution of correct and incorrect responses for each trainee on each section of the test is independent of the time of test administration.)

   The research hypothesis (H_1) is set up as an alternative to the null hypothesis. It is formulated to provide for a definite decision when the null hypothesis is rejected and is based on some outcome predicted by a theory or by the research interests of the investigator (see example below).

3. Operationalized Hypotheses The null hypothesis, as described above, will be stated in a general form: that instruction produces no effect on levels of subject matter competence. In order to test the hypothesis, it must be stated in operational form. For example, the first null hypothesis to be posited for testing is that the mean Pre-Test score is equal to the mean Post-Test score for Item Set 1 (i.e., $H_o: \mu_{T1} = \mu_{R1}$). The research hypothesis will be so stated that, if the null hypothesis is rejected, evidence will be provided to support an outcome predicted (or desired) by the investigator.

   With the type of analysis to be described here, the direction of the desired mean differences is posited in advance and is important to the evaluation objectives for which the statistical tests were initially called into use. Assuming that all the proper controls, discussed

in Chapter II, have been employed, significant increases in Post-Test scores are a powerful indicator that the sequence of instruction was effective in raising the trainees' levels of subject matter competence. Therefore, the research hypothesis will be directional.

In the example cited above, the research hypothesis would state that training does increase the trainees' levels of subject matter competence. Operationally, H_1 would be that the Post-Test mean is significantly higher than that of the Pre-Test for Item Set 1 (i.e., $H_1: \mu_{T1} < \mu_{R1}$). Since the direction of the score deviation can be hypothesized in advance, the test employed should be one-sided. That is, the alternative hypothesis should declare not only that one score mean will be significantly greater than the other, but it should also specify which score mean it is (in this case, the mean for the Post-Test).*

The table of t-values provided in Figure F2 of Appendix F considers only the one-tailed test.

4. Interpreting the Statistical Test Results. The question to be answered is whether or not the testing outcome justifies the rejection of the null hypothesis in favor of the alternative hypothesis. The conclusion drawn from the test results is referred to as a judgment of significance. Such a judgment is based upon the probability of obtaining, by chance, results as discrepant as those actually observed, when in fact, the null hypothesis is true. In the case of Test/Retest scores, the judgment is based upon the probability of getting a mean (or individual trainee) score difference as large as the difference actually observed. If the probability is low enough, then the null hypothesis that such large score differences were achieved by pure chance is not accepted.

---

* Although the one-tailed test is appropriate for the analysis proposed here, it should be noted that another approach is sometimes employed. That is, the alternative hypothesis could simply state that the training does influence levels of subject competence without suggesting the direction of that influence. Operationally, the H_1 would state that the Item Set 1 mean scores are not equal -- i.e., $H_1: \mu_{T1} \neq \mu_{R1}$. A statistically significant difference in the means, whether larger or smaller, from Test to Retest would reject the null hypothesis in favor of H_1. This is an example of employing a non-directional, two-tailed test.

It can therefore be inferred (if the alternative hypothesis is directional in favor of Post-Score gains) that the sequence of instruction contributed to the magnitude of the score differences.

After both the null and alternative hypothesis have been posited, the next step is to compute the probability of obtaining score differences as great as (or greater than) those observed on the assumption that the null hypothesis is true and that whatever differences were observed were due to chance. For Stage 1, this probability will be derived from the application of the t-Test to the mean Test and Retest scores (the probability value being the value of the computed t-statistic).

For Stage 2, the probability will be the value of the χ^2 statistic computed from the application of the χ^2 Test to the individual trainee Test/Retest scores.

The question that now arises is "how small a probability is necessary in order to consider the score differences significant"? The probability of a chance occurrence can be of any magnitude between 0 and 1 (i.e., 0 indicating that an outcome as discrepant as that observed could not happen if the null hypothesis were true, and 1 indicating that an outcome as discrepant would be certain to happen on the basis of the null hypothesis). It is therefore necessary for the evaluators to decide upon some particular probability value (called the level of significance) which they consider is small enough so that they will feel confident in inferring a highly effective training sequence (at the level of subject matter competence) when the computed probability is less than this selected value. The decision as to which particular probability value to employ will be somewhat arbitrary and can vary from one evaluator to the next.

One evaluator may feel that if the observed Test/Retest score differences would occur by chance with a probability of 1 in 10, then he is justified in inferring that the training is effective. Another might accept only a probability as low as 1 in 20 in order that the score differences be accepted as significant. Still another may only accept a probability as low as 1 in 100; and some may require a probability as low as 1 in 1000.

The position taken in this Manual is that instruction should have a strong impact on post-competence levels, in terms of large increases in Post-Test scores before such data can serve as evidence of training effectiveness. Therefore, the minimum standards for a statistical judgment should be as follows:

Refer to statistical test results with a probability greater than .05 as not significant; to those with a probability equal to or less than .05 (i.e., 1 in 20) but greater than .01 as significant, or as significant at the 5% level; and to those with a probability equal to or less than .01 (i.e., 1 in 100) as highly significant, or as significant at the 1% level.

NOTE: The above discussion is only a cursory presentation of the concepts of statistical inference as they relate to the types of analyses described in this chapter. Although the topic will be referred to throughout this chapter, many of the assumptions and principles underlying the specific statistical tests and procedures employed in the analysis are beyond the scope of the Manual. For a more comprehensive coverage of the area of statistical testing, the interested reader should consult one of the statistical texts cited in the Bibliography.

STAGE 1: SUBJECT AREA SCORES

The first area of analysis is the Pre-/Post-Test performance of the total trainee group on the test as a whole and on item sets. The assumption underlying this stage of analysis is that significant increases in levels of competence among trainees as a result of training should first manifest themselves as significant increases in Post-Test scores over Pre-Test scores, for individual item sets and the composite test. The logical question to pose at this stage then is whether or not the increased competence involves all subject matter covered during the course, or only selected areas.

The analysis question is whether or not the item set and composite Post-Test scores show statistically significant increases over their Pre-Test counterparts. Since in a Pre-/Post-Test situation there may be factors other than the effects of instruction that could cause score increases (such as random score fluctuations), it is necessary to apply a selected statistical test to determine the significance of any observed score increases.

Consider, for example, the test scores illustrated in Figure 10. As can be observed, all subject area scores show positive mean score gains. Whether or not these gains represent significant score increases or simply a function of chance factors can be determined by the application of the appropriate test for statistical significance.

Figure 10

MEAN PRE-/POST-TEST SCORES FOR ITEM SETS AND COMPOSITE TEST

Subject Area	Pre-Test	Post-Test	% Gain	Maximum Possible Score
Item Set 1	23.77	35.06	47.4	51
Item Set 2	23.87	30.94	29.6	45
Item Set 3	6.74	9.13	35.4	17
Composite	59.39	75.13	38.1	113

The statistical test selected for this level of analysis is the t-Test for Related Variables.

(A variation of the standard t-Test is included in Appendix F. It was designed by Dr. David Wolfers, of the Institute

staff, specifically for application to score data derived under a Test Retest administration design. Since it is a new and, as yet, untested statistical technique, it was not included for the purpose of replacing the standard t-Test in the analysis. However, it is included with the intention that those with experience in quantitative assessment of test score data might employ and assess this new technique and communicate their findings to the authors. This variation of the standard test is considered by the authors as potentially valuable to the area of Test Retest assessment and it is believed that collective experience of its application will prove its utility and validity. A further discussion of this new technique, including the derivation of the formula, and computational data also provided in Appendix E.

The score distributions recorded on the Score Summary Profile (see Figure 8) will provide the data input for this stage of analysis. For each item set and for the composite scores, the hypothesis to be tested is that the Pre-Test mean is equal to the Post-Test mean. This will be tested, in each case, against the alternative that the Post-Test mean is significantly greater than the Pre-Test mean.

The computational formulas for the standard t-Test and the criteria for significance, together with a sample run, are given in Appendix F.

When all t-testing has been completed, the decisions should be recorded on an Analysis Summary Profile (see Figure 13) for use in subsequent reporting of the assessment.

** PROGRAMS TDEKMERG & TRELVAR ARE USED FOR THIS STAGE OF THE ANALYSIS (see Appendix E)**

STAGE 2: INDIVIDUAL TRAINEE SCORES

When the objective is to determine if individual trainees have increased their levels of competence from the pre- to the post-training period, the statistical procedure employed is the Chi Square Test of Independence. This test is applied to item set and composite score pairs for each trainee. The score data can be expressed in a 2 X 2 Contingency Table. An example of this type of table, containing the Pre- and Post-Composite scores for trainee #01 (see Figure 11), is shown below. (The computational formula and criteria for significance for the Chi Square Test, using data from the Table are provided in Appendix F.)

ITEMS

TESTING	Correct	Incorrect
Pre-Test	41	53
Post-Test	78	35

The hypothesis being tested here is that for any trainee the distribution of correct and incorrect responses is independent of the time of testing -- i.e., independent of the effect of instruction. This will be tested against an alternative that the distribution of correct and incorrect responses depends on the time of testing. The data input for this analysis stage is the individual trainee's Pre-Test and Post-Test scores (both item set and composite) provided on the Score Summary Profile (Figure 5).

Since the Chi Square test is applied repeatedly to each set of Pre- and Post-Test scores for all trainees, it is suggested that each computation be checked by an independent evaluator.

When all testing is completed, the conclusions (significant or non-significant) should be recorded with the other data on the Analysis Summary Profile, sheet (Figure 12).

** PROGRAM CHI**2 IS USED FOR THIS STAGE OF DATA ANALYSIS. (see Appendix E)**

STAGE 3: ITEM RESPONSE PATTERNS

The first two stages of analysis involve the application of statistical procedures for determining the significance of Test to Retest score changes by subject area and for individual trainees. The objective of the final stage is to isolate the factors which contributed to these changes. The data must be analyzed item-by-item, categorizing each item according to the type of response pattern that occurred from Test to Retest. There are four possible patterns:

(1) Correct in both Pre-Test and Post-Test (C→C);
(2) Incorrect in both Pre-Test and Post-Test (I→I);
(3) Correct in Pre-Test and Incorrect in Post-Test (C→I);
(4) Incorrect in Pre-Test and Correct in Post-Test (I→C).

Categories 1 and 2 are "stable" item response patterns. C→C items may have involved subject material that trainees were

competent with prior to training and that was positively supported during instruction. I→I items may represent subject material that the trainees were not competent with prior to training and that was not adequately learned during instruction. It should be pointed out here that underlying all these patterns is the assumption of the contribution of both successful and unsuccessful guessing.

The I→C and C→I items will be the primary focus of study. Regardless of whether the analysis is of subject area or individual trainee scores, the number of I→C items represents the total amount of Pre- to Post-Test score increase. Since there is no basis for assuming that the degree of correct guessing was greater on the Post-Test than on the Pre-Test, the I→C items provide the quantitative measure of the magnitude of the increase in subject matter competence. This measure indicates the amount of significant learning that occurred as a result of training and is, thus, an indicator of the training program's effectiveness in attaining one of its primary objectives.

The overall magnitude of the score increases is reduced by the C→I items. While this type of pattern does occur in Test-Retest situations, it is difficult to determine the cause. While the I→C shift can be considered a measure of learning, it cannot be assumed the C→I changes represent "unlearning". Some assumptions can be made however. It is possible that the pre-correct responses resulted from blind guessing or naive reasoning rather than from true competence with the subject material. The C→I items may contain ambiguities or irregularities in structure or content. When faulty items are included in a test, variations in response to these items from one testing to the next are highly probable. The new learning acquired during training may produce confusion, especially on faulty items. That is, attempts to apply new information to the questions or problems posed by the Post-Test items may result in a greater number of incorrect responses than would be obtained by guessing, with little or no understanding, of the subject material. The C→I pattern can be the result of the common problem associated with objective-type items, of "knowning too much" to respond correctly or of over-analysis of items of questionable structure and intent.

In order to conduct an adequate analysis of item response patterns, the item response data should be displayed in tabular form for each item set and the composite test. The types of tables required are those shown in Figures 11 and 12. The data in these tables are tabulations of the Test-Retest responses of the 31 trainees to the 51 items of Set 1. (While the discussion of the analysis will center on one item subset, the procedures involved are the same for all item sets and the composite test.)

To construct such a table, work with the responses on the

FIGURE 11

ITEM RESPONSE PATTERNS
BY INDIVIDUAL TRAINEE

ITEM SET 1 — FP/MCH Program 1 (11/73-12/73)

ID	(1) TOTAL CORRECT PRE	(1) TOTAL CORRECT POST	PRE----→ POST C-->C	PRE----→ POST C-->I	(2) TOTAL INCORRECT PRE	(2) TOTAL INCORRECT POST	PRE---→ POST I-->I	PRE---→ POST I-->C	TOTAL ITEMS (1+2)
1	28	34	24	4	23	17	13	10	51
2	30	42	26	4	21	9	5	16	51
3	13	19	6	7	38	32	25	13	51
4	26	43	25	1	25	8	7	18	51
5	23	36	19	4	28	15	11	17	51
6	29	37	28	1	22	14	13	9	51
7	33	33	28	5	18	18	13	5	51
8	23	31	16	7	28	20	13	15	51
9	28	32	22	6	23	19	13	10	51
10	29	34	25	4	22	17	13	9	51
11	25	34	21	4	26	17	13	13	51
12	13	29	12	1	38	22	21	17	51
13	18	35	15	3	33	16	13	20	51
14	22	35	17	5	29	16	11	18	51
15	24	36	20	4	27	15	11	16	51
16	16	36	11	5	35	15	10	25	51
17	31	41	30	1	20	10	9	11	51
18	20	36	17	3	31	15	12	19	51
19	19	36	16	3	32	15	12	20	51
20	21	41	20	1	30	10	9	21	51
21	29	37	25	4	22	14	10	12	51
22	8	31	5	3	43	20	17	26	51
23	27	35	22	5	24	16	11	13	51
24	29	36	23	6	22	15	9	13	51
25	28	38	26	2	23	13	11	12	51
26	23	38	21	2	28	13	11	17	51
27	27	29	15	12	24	22	10	14	51
28	29	39	25	4	22	12	8	14	51
29	18	38	16	2	33	13	11	22	51
30	22	35	16	6	29	16	10	19	51
31	26	31	23	3	25	20	17	8	51
TOTAL:	737	1087	615	122	844	494	372	472	1581

46.6% - - - - - - - - - - 53.4% - - - - - - - - - - →100%
68.8% - - - - - - - - - - 31.2% - - - - - - - - - - →100%
100%← - - - - - - 83.4% - - -16.6%   100%← - - - - - -44.1% - - -55.9%

original Test and Retest answer sheets. Transferring the data from the answer sheets to the final table can be facilitated by means of an item response worksheet. A section of the worksheet used to construct the tables in Figures 11 and 12 as well as the steps involved in constrcuting these tables from the worksheet are fully described in Appendix F.

It should be stated at this point that although the data used are quantitative, the analysis of item pattern responses is essentially qualitative and subjective in nature. It results in tentative assumptions rather than substantiated conclusions derived from rigid statistical testing.

In order to illustrate how the analysis should be conducted, an actual situation is described. The following is a summary of the analysis conducted on the data in Figure 11. In this situation, application of tests of statistical significance found that the Post-Test scores displayed significant gains over their Pre-Test counterparts. An item response pattern analysis was conducted in an attempt to isolate the factors which may have contributed to this significance.

> The percentage of total test items having correct Pre-Test responses is 46.6% with 53.4% being incorrect Pre-Test items. Pre- to Post-Test response patterns were analyzed by comparing changes occurring among pre-correct items with those pre-incorrect.
>
> 83.4% of the pre-correct items were also correct on the Post-Test, compared with 44.1% of the pre-incorrect items that were also incorrect on the Post-Test. This shows a greater degree of response stability for the pre-correct than for the pre-incorrect from Test to Retest. That is, the number of pre-incorect items that change from Test to Retest is greater than those initially correct.
>
> The percentage of total pre-incorrect responses that became correct on the Post-Test is 55.9% which represents the total amount of Pre- to Post-Test Score increase. This increase is reduced, however, by a downward shift of 16.6% of the total pre-correct to post-incorrect responses.*

---

* The influence of these opposing patterns on relative Test/Retest score change can best be illustrated by considering an individual case (see Figure 11). For example, on Item Set 1 trainee No. 1 displays Pre- and Post-Test scores of 28 and 34, respectively. The score gain of 6 is the result of two opposing patterns of Test to Retest item response shift. (The number of I→C items is 10 and the number of C→I items is 4; from this the net score increased is calculated as 6 -- i.e., 10-4.)

This C→I shift is probably to a large extent some type of testing artifact which had a negative effect on overall trainee performance on Item Set 1. While the available data do not permit a determination of the cause(s) of this artifact, the fact remains that it does reduce the overall score gain in Item Set 1 to a certain degree. The net increase in Post-Test scores was of sufficient magnitude, however, to prove significant when subjected to statistical testing.

Such an analysis should be conducted for each item set and the composite test. In terms of the data provided, the breakdown of Pre- to Post-Test response behavior into discrete response pattern categories will be relevant to the assessment-of-training study being proposed. The analysis of item patterns provides not only a quantitative indicator of the amount of learning that has occurred from the pre- to post-instruction period (i.e., the magnitude of the I→C shift), but also data concerning those factors which contribute, both positively and negatively, to the net score gains, and that portion of response behavior where no Test to Retest change occurred.

An item set and composite response pattern analysis is even more relevant when the Post-Test scores fail to display significant increases over their Pre-Test counterparts. In this situation, the analysis would attempt to isolate those item response factors which contributed to the lack of significant score increase. The assumption underlying the analysis in this case is that non-significance of score gain may be the result of negative factors related to test performance and not only to the failure of the trainees to increase their levels of subject competence between the pre- and post-instruction periods.

In addition to the distribution of response patterns for total item groups, the data in that same table (in Figure 11) provide item response patterns for each individual trainee. An analysis of the relative contribution of each response pattern to the trainee's post-scores will, like the analysis of item groups, help isolate those test factors that contributed to the trainee's achievement of (or failure to achieve) significant score gains.

Test/Retest item response patterns can also be assessed by individual item. A table of response patterns generated for each of the 51 items of Set 1 is illustrated in Figure 12. For each item, comparisons of the frequencies in each response pattern category will indicate the relative contribution of that item to the Pre- to Post-Test score change. Again, since the focus is on score increases, the item patterns involving change (I→C and C→I) would be examined. Those items contributing most to the Post-Test score increase will be those with the highest frequency of response change in the I→C category. For example, Item No. 27 (with C→I and I→C frequencies of

FIGURE 12

ITEM RESPONSE PATTERNS
BY INDIVIDUAL ITEM

FP/MCH Program 1 (11/73-12/73)

ITEM SET 1	(1) TOTAL CORRECT		PRE----→ POST		(2) TOTAL INCORRECT		PRE---→ POST		TOTAL RESPONSES (1+2)
ITEM	PRE	POST	C-->C	C-->I	PRE	POST	I-->I	I-->C	
1	8	11	3	5	23	20	15	8	31
2	10	11	6	4	21	20	16	5	31
3	7	16	6	1	24	15	14	10	31
4	11	27	10	1	20	4	3	17	31
5	22	27	21	1	9	4	3	6	31
6	14	26	13	1	17	5	4	13	31
7	26	30	25	1	5	1	0	5	31
8	23	25	18	5	8	6	1	7	31
9	14	9	4	10	17	22	12	5	31
10	21	22	14	7	10	9	2	8	31
11	21	27	19	2	10	4	2	8	31
12	15	26	13	2	16	5	3	13	31
13	24	29	23	1	7	2	1	6	31
14	19	27	18	1	12	4	3	9	31
15	22	21	18	4	9	10	6	3	31
16	20	29	18	2	11	2	0	11	31
17	26	22	20	6	5	9	3	2	31
18	25	30	24	1	6	1	0	6	31
19	4	3	1	3	27	28	25	2	31
20	3	13	0	3	28	18	15	13	31
21	7	1	1	6	24	30	24	0	31
22	6	21	6	0	25	10	10	15	31
23	20	21	16	4	11	10	6	5	31

" and I" respectively" makes a greater relative contribution to the Test Retest score increase than Item No. 9 (with a C→I frequency of 11 and an I→C of 5). A comparision of each item's response patterns with those of every other item in the table would be time consuming and unnecessary since the most relevant item response pattern assessments are those at the item group (item sets and composite test) and individual trainee levels. However, the evaluator should examine data of the type generated for the item tables (i.e., the table in Figure 11) to obtain a more complete picture of Test/Retest response behavior down to and including the individual item level.

It will also be useful, if staff time permits, to compare the results of the Curriculum Audit (see Chapter 7) with the data provided in the individual item tables. The judgments of degree of course coverage given the content assessed by individual items can be compared with the item pattern response frequencies to determine if variations in item content coverage are reflected in variations in the response patter frequencies. For example, one hypothesis that might be evaluated is that the more complete a test item's content is judged to have been covered during instruction, the higher will be the frequency of the item's I→C category and the lower the frequency in the C→I category.

Looking at the response patterns of individual items is also helpful in determining the quality of items. Items which have high frequency of C→I or I→I responses should be carefully reviewed. Having this information available for individual items greatly assists the item analysis discussed on pages 121-122.

The validity of the item pattern analysis depends upon how well the items themselves were constructed. If it can be assumed that the items are adequately designed and actually measure what they were designed to measure, then such an analysis combined with the statistical test results of Stages 1 and 2 will provide most of the data which will contribute to decisions concerning training effectiveness and trainee achievement at the level of subject matter competence.

** PROGRAM ITEMPAT WILL GENERATE THE ITEM PATTERN TABLES OF THE TYPE ILLUSTRATED IN FIGURES 11 AND 12 (see Appendix E)**

USING THE ANALYSIS RESULTS: EVALUATING TRAINING

The final results of the statistical analysis of score data (stages 1 and 2) should be recorded on an analysis summary table of the type shown in Figure 13*. In this way the staff will have (all on one form) the score distributions with their respective means and standard deviations together with a summary of results from the application of tests of significance to the trainee group as well as individual trainee Test/Retest scores. The score data is grouped by separate item subsets (e.g., Item Sets 1-3) as well as by total (composite) test. These data, when combined with the results of the response pattern analysis and Curriculum Audit provide a comprehensive profile of a trainee group's performance in a pre-/post-instruction testing situation. (NOTE: When data have been analyzed by trainee subgroup, the score data for each grouping should be presented separately -- to facilitate inter-group comparisons -- either on the same or, when the number of trainees is large, on separate summary sheets.)

Valid inferences concerning training effectiveness and trainee achievement in the area of subject matter competence can be drawn from this body of data if the test instrument which generated these data was constructed according to the structured guidelines provided. In addition, certain technical factors relating to statistical testing and achievement criteria, when taken into consideration, increase the confidence with which such inferences can be drawn.

Statistical Test Results When interpreting the results of repeated application of tests of significance to the item group and individual trainee score gains, only a large number of "significant" differences should be accepted as providing evidence for the effectiveness of training. This is due to the fact that when a large number of statistical tests is applied to a body of score data, a small number of significant results can be expected to occur by chance alone.** One or a few marginally "significant" results (from a large number of test applications) can, therefore, be misleading, but the consistency of a large number of "significant" differences can serve as a valid indication of the positive impact of instruction on levels of subject competence. Therefore, even when

---

* "ns" indicates non-significant score gains, while $p < .05$ and $p < .01$ indicate score increases found to be statistically significant at or beyond the 5% and 1% levels, respectively.

** Both Kish (10) and Selvin (11) discuss this factor in their assessment of the application and misuse of tests of significance in research; computational formulas are provided for estimating the number of significant results to be expected by chance when X number of statistical tests are applied.

FIGURE 13

ANALYSIS SUMMARY PROFILE

FP/MCH Program 1 (11/73-12/73)

TEST/RETEST DATA ANALYSIS SUMMARY

EXAMINEE	ITEM SET 1			ITEM SET 2			ITEM SET 3			COMPOSITE		
ID	T1	R1	CHI^2	T2	R2	CHI^2	T3	R3	CHI^2	ΣT	ΣR	CHI^2
1	28	34	ns	23	32	ns	9	12	ns	60	78	p<.05
2	30	42	p<.05	23	29	ns	4	6	ns	57	77	p<.05
3	13	19	ns	14	17	ns	3	2	ns	30	38	ns
4	26	43	p<.01	25	36	p<.05	9	9	ns	60	88	p<.01
5	23	36	p<.05	29	33	ns	5	11	ns	57	80	p<.01
6	29	37	ns	30	36	ns	7	7	ns	66	80	ns
7	33	33	ns	25	34	ns	7	12	ns	65	79	ns
8	23	31	ns	21	28	ns	6	6	ns	50	65	ns
9	28	32	ns	22	24	ns	9	8	ns	59	64	ns
10	29	34	ns	21	28	ns	6	6	ns	52	68	p<.05
11	25	34	ns	21	28	ns	6	6	ns	52	68	p<.05
12	13	29	p<.01	28	29	ns	4	9	ns	45	67	p<.01
13	18	35	p<.01	27	30	ns	10	9	ns	55	74	p<.05
14	22	35	p<.01	21	30	ns	6	9	ns	49	74	p<.01
15	24	36	p<.05	26	29	ns	6	8	ns	56	73	p<.05

16	16	36	p<.01	19	30	p<.05	6	10	ns	41	76	p<.01
17	31	41	p<.05	28	39	p<.05	8	10	ns	67	90	p<.01
18	20	36	p<.01	25	30	ns	4	9	ns	49	75	p<.01
19	19	36	p<.01	23	32	ns	6	12	ns	48	80	p<.01
20	21	41	p<.01	23	31	ns	7	6	ns	51	78	p<.01
21	29	37	ns	24	33	ns	9	13	ns	62	83	p<.01
22	8	31	p<.01	14	29	p<.01	6	8	ns	28	68	p<.01
23	27	35	ns	16	33	p<.01	7	9	ns	50	77	p<.01
24	29	36	ns	25	32	ns	7	9	ns	61	77	p<.05
25	28	38	ns	27	30	ns	10	12	ns	65	80	ns
26	23	38	p<.01	25	33	ns	8	12	ns	56	83	p<.01
27	27	29	ns	28	30	ns	7	8	ns	62	67	ns
28	29	29	ns	29	38	ns	6	12	ns	64	89	p<.01
29	18	38	p<.01	24	28	ns	6	8	ns	48	74	p<.01
30	22	35	p<.05	23	33	p<.05	5	7	ns	50	75	p<.01
31	26	31	ns	28	33	ns	5	9	ns	59	73	ns
Mean =	23.77	35.06		23.87	30.94		6.74	9.13		54.39	75.13	
Std.Dev.=	5.94	4.60		4.09	4.06		1.95	2.64		9.59	9.50	
t-Test =	p<.01			p<.01			p<.01			p<.01		

the item set and composite scores are found to be significant, the total number of "significant" Test/Retest score differences within item groups must be considered and weighed as evidence. For example, in Figure 13, the large number of significant decisions among trainees for Item Set 1 (i.e., 16 out of 21) and the composite test (23 out of 31) provides a strong indication of the positive impact of instruction on competence, both with the subject matter sampled by Set 1 and with the overall subject matter. However, only 6 of the 31 trainees displayed significant score gains in Set 2 and none was displayed in Set 3. Given the high significance of the mean score gains in these two areas (i.e., $p < .01$), the small number of within-set "significant" decisions indicates that the training was less effective in raising levels of competence in these two areas than with the Set 1 material and the overall material. Reporting (in the final write-up) both the overall results of testing for the item sets and composite together with number of "significant" results out of total tests applied within item groups will allow a more adequate and reliable assessment of training effectiveness in general and will allow the evaluators to assess the relative effectiveness of training by subject areas and by individual trainees within subject areas.

One final consideration regarding statistical test results: although both the 5% and 1% levels are suggested here as acceptable significance levels when assessing score gain, the training staff might want to consider only the more substantial score gains as providing evidence of training effectiveness. In such cases, the staff might only accept as evidence those score gains significant at the 1% level or higher (e.g., the .005 or .001 levels). The higher the level of significance for a particular Test/Retest score gain, the more confident will be the inferences drawn concerning the impact of instruction in affecting competence levels. (The reader can set significance levels different from those suggested here and obtain the sampling distributions for these levels -- like those provided in Figure F2 for the 5% and 1% levels -- from any standard statistics text.)

Criteria of Achievement Another technical factor that must be considered when interpreting the testing results is establishing criteria of achievement, i.e., defining standards of what constitutes high competence in a particular subject matter area.

It will not usually be possible for a training staff to set meaningful, pre-determined score values which the trainees must attain or exceed in order to be considered highly competent in one or another subject area. The reason for this lies with the nature of the subject matter covered during instruction. According to Ebel (12),

"Course content is usually selected on the basis of subjective decisions, often by individual instructors. As such, it hardly possesses the characteristics of an absolute standard of achievement. Nor is it ordinarily possible for the constructor of an objective test to gauge the difficulty of his items precisely enough to define a fixed standard of achievement with respect to that content."

If it is not possible to set meaningful cut-off scores for determining high competence in the post-instruction period, another indicator, or set of indicators, of training effectiveness must be employed.

The first indicator to consider is the results of the application of tests of statistical significance. However, even when there are highly significant Pre- to Post-Test score increases, these may or may not be acceptable to the staff as an indication of training effectiveness. Consider, for example, a test with a maximum score of 100. A mean composite score increase (all trainees) from a Pre-Test of 20 to 45 on the Post-Test is found to be statistically significant. Given the significant finding, the training staff is faced with the task of assessing how effective the training sequence was (and how competent the trainees are) when an average of 55% (i.e., 100-45) of the subject material sampled by the test items was not learned. In this situation, the significance of the score gain is not sufficient evidence for any definitive evaluative decision. This, combined with the usual absence of established criteria against which trainees' scores can be compared, makes it necessary to carry out a further assessment of testing results.

Level and Magnitude of Score Movement

Along with the magnitude of the Pre- to Post-Test score movement, the level at which that movement occurs must also be considered when drawing inferences concerning training effectiveness. That is, trainees may be distinguished according to subject competence reflected by the magnitude of the Post-Test scores and/or their achievement which reflects their performance on the Post-Test compared with their performance on the Pre-Test. Training administrators and the trainee supervisors (in the post-training job situation) will be interested in both these parameters.

The amount of competence a trainee displays will be of obvious interest to those who are responsible for placing him in a job or assigning him job duties. The training administrators will also be interested in seeing that the trainees demonstrate high levels of competence, as indicative of the effectiveness of their training efforts.

The amount of achievement displayed, however, is also an important piece of information for all concerned. To take an obvious case: if two trainees complete a training sequence with Post-Test scores of 90%, it can be assumed that they are, more or less, equally competent with the subject matter being assessed. However, if one of these trainees started out with a Pre-Test score of 40% and the other with a Pre-Test score of 85%, the large amount of achievement demonstrated by the first trainee may be said to show: a) that the sequence of instruction was highly effective in increasing his subject matter competence and b) that he has a great capacity for self-improvement, which could be an important quality to take into account when deciding what job to assign him.

In addition to the amount of achievement (i.e., the magnitude of the Pre- to Post-Test score movement), the level at which the score movement takes place within the possible range of score change will be of importance. In order to consider both parameters of achievement, it is recommended that an additional score be computed for all groups, subgroups and individual elements; that is, for the entire trainee group, for the total item group, for trainee and item subgroups and for individual trainees and items. Such a score would consist of the sum of a competence score (i.e., the Post-Test score) and an achievement score (i.e., the Post-Test score minus the Pre-Test score). Giving both scores equal weight in computing the new combined score allows for a trainee who has made a great score gain to show well in the overall picture, even if his final competence score may be somewhat lower than that of another trainee.

A sample set of combined scores of the type described above is displayed in Figure 14. In the sample, the combined Achievement/Competence scores (A+C) are grouped from highest (190) to lowest (30) by intervals of 10 score units. It should be noted that the listing is not exhaustive of all possible combinations of competence scores and achievement scores, but only a small sample of possible combinations that serves to illustrate the necessity of considering not only the magnitude of the score increase but also the level at which the increase occurred.

As can be seen, the highest possible A+C score is 190, based on a Test score of 10 and a perfect Retest score of 100 while the lowest score possible is 30 based on Test and Retest scores of 10 and 20, respectively.*

* Actually the highest A+C score possible is 200 based on a Test score of 0 and Retest score of 100 (where C = 100 & A = 100) and the lowest possible is 0 where the Test and Retest scores are both 0. For purposes of illustration however, not all possible A+C scores are necessary.

FIGURE 14

ACHIEVEMENT/COMPETENCE SCORES (UNWEIGHTED)

Highest to lowest possible, intervals of 10 points:

Pre → Post	= A+C
10 → 100	= 190
20 → 100	= 180
10 → 90	= 170
30 → 100	= 170
20 → 90	= 160
40 → 100	= 160
10 → 80	= 150
30 → 90	= 150
50 → 100	= 150
20 → 80	= 140
40 → 90	= 140
60 → 100	= 140
10 → 70	= 130
50 → 90	= 130
70 → 100	= 130
20 → 70	= 120
40 → 80	= 120
60 → 90	= 120
80 → 100	= 120
10 → 60	= 110
30 → 70	= 110
50 → 80	= 110
70 → 90	= 110
90 → 100	= 110
20 → 60	= 100
40 → 70	= 100
60 → 80	= 100
80 → 90	= 100
10 → 50	= 90
30 → 60	= 90
50 → 70	= 90
70 → 80	= 90
20 → 50	= 80
40 → 60	= 80
60 → 70	= 80
30 → 50	= 70
50 → 60	= 70
20 → 40	= 60
40 → 50	= 60
10 → 30	= 50
30 → 40	= 50
20 → 30	= 40
10 → 20	= 30

What does the A+C score tell the evaluator? It is quite clear that upper scores (140-190) reflect both a high degree of competence as well as a high degree of achievement. One could infer from these data a highly effective training program. That is, the trainees started off with very low Pre-Test scores and displayed large score gains (i.e., high achievement) with a resulting high Post-Test score (i.e., high competence). Similarly, low A+C scores (i.e., 30-60) indicate a situation where trainees began with low Pre-Test scores, made low to moderate score gains, and completed a sequence of instruction with relatively low levels of competence in the subject area under assessment. The need to revise the curriculum would be indicated on the basis of these low A+C scores.

Consider, however, the combinations of Test and Retest scores which provide an A+C score of 110. (Refer to the data in Figure 14.) As can be seen, both a low Pre-Test score of 10 and a high Pre-Test score of 70 can result in an A+C score of 110. Here, a low Pre-Test level of 10 combined with a moderately large score gain of 50, and a high Pre-Test level of 70 coupled with a relatively low gain of 20 result in the same A+C score. It is clear from the data that the first score situation (i.e., 10 → 60) displays a greater magnitude of achievement than the second situation (i.e., 70 → 90). However, the 70 → 90 score change reflects a higher level of post-training competence than does the 10 → 60 change. What inferences can be drawn concerning the impact of the training experience based upon these two testing situations? It would seem that one cannot readily draw any definite conclusions.

Since competence and achievement are given equal weight, it cannot be readily determined which of the two score situations represents the more effective training experience. Such conclusions will be based upon the criterion for effectiveness selected by the evaluator(s). That is, if the highest level of competence attained is the effectiveness criterion, then the second testing situation (i.e., 70 → 90) will reflect a more effective training sequence than the 10 → 60 situation.

However, if the effectiveness criterion is the magnitude of achievement, than a different picture emerges. In this case, the first score situation (10 → 60) displays a greater degree of achievement than the 70 → 90) situation. Based upon the achievement criterion, then, the first score situation reflects a more effective training sequence than the second.

Since competence and achievement are afforded equal weight in computing the A+C scores, the evaluator might decide to consider any combination of achievement and competence scores resulting in similar A+C scores as reflecting equal levels of effectiveness (or ineffectiveness, as the case may be). Based upon this

reasoning, the scores situations 10→60, 30→70, 50→80, 70→90 and 90→100, since they all have an A+C score of 110, would be considered by the evaluator as equal levels of training effectiveness.

The use of unweighted A+C scores of the type described above is based upon the assumption that equal gains of raw-score points (anywhere along the entire range of potential scores) represent equal increments in competence or achievement. This means, for example, that it is as difficult to increase a Pre-Test score from 20→40 on the Post-Test as it is to increase a score from 60→80, both of which involve equal increments of 20 raw-score points.

If evaluators are willing to accept this assumption, then the use of unweighted A+C scores can be used to draw inferences concerning training effectiveness and trainee achievement.

However, the strong possibility exists that the difficulty in increasing a score a certain number of raw-points varies at different levels along the range of possible scores. That is, equal gains of score points may not correspond to equal increments in competence (or achievement) at all points along the score range. Some evidence to support this contention is provided by Diederich (13) in an article describing the results of Pre- and Post-Tests given to 1,400 college students. In a discussion of the difficulty in translating gains of raw-score points into increments of ability (or achievement), he states that "it is harder to get from the mean (score) up to plus one standard deviation than it is to get from minus one standard deviation up to the mean." (Of course, this finding cannot be generalized to all test situations, with a high degree of confidence, without further study.)

If the evaluators accept this assumption as true in their particular testing situation, then it is necessary when deriving the A+C scores to use weighted values. That is, the scores should be weighted to reflect the assumption that as you increase the Pre-Test score level, it becomes increasingly difficult to increase that score (on the Post-Test) by X number of raw-score points. For example, the weighted A+C score will show that it is more difficult to increase a pre-score of 50 to a post-score of 70 than it is to raise a pre-score of 30 to a post-score of 50.

The relatively higher level of achievement attained in increasing

a score from, say, 50 to 70 compared to that of increasing a score from 30 to 50 will also be reflected by the weighted A+C scores.

The weighted A+C score curves of Figure 15 and the table in Figure 15A displaying both Test and Retest scores and the derived weighted A+C values will serve to illustrate the use of the weighted score in the assessment of the effectiveness of a sequence of training. (The mathematical equation and the parameter values employed in generating the series of curves from which the A+C score values are derived are provided in APPENDIX G*.)

The A+C curves and the Test, Retest and weighted A+C score values are, like the unweighted score data in Figure 14, for illustrative purposes and represent only a small sample of possible score curves and combinations of score values that would result from a large number of Test/Retest administrations of the same instrument.

Before discussing how the weighted scores might be employed in the assessment process, the procedure for deriving the weighted values for any series of Test and Retest scores will be considered, using the data in Figures 15 and 15A.

The weighted A+C score values for a specific set of Pre- and Post-Test scores can be obtained from the score curves in the following manner:

> Consider the case in which the Pre-Test score (Test) is 10 and the Post-Test score (Retest) is 20. Locate the Test score value along the vertical axis and the Retest score along the horizontal axis. The point on the graph at which these two values intersect -- the coordinates labelled (10, 20) in Figure 15 -- defines the A+C score value. In this case the A+C score associated with those coordinates is 10 since the A+C curve that passes through coordinates (10, 20) is 10 (the circled numbers next to each of the curves are the weighted score values).
>
> In cases where the A+C curve does not pass directly through the intersection of a pair of Test/Retest coordinates, the A+C value for those coordinates can be obtained by interpolating between two adjacent score

---

* The mathematical equation and the resulting family of weighted A+C score curves were developed by Dr. David Wolfers of the Institute staff specifically for use in this Manual.

FIGURE 15

WEIGHTED ACHIEVEMENT/COMPETENCE SCORE CURVES

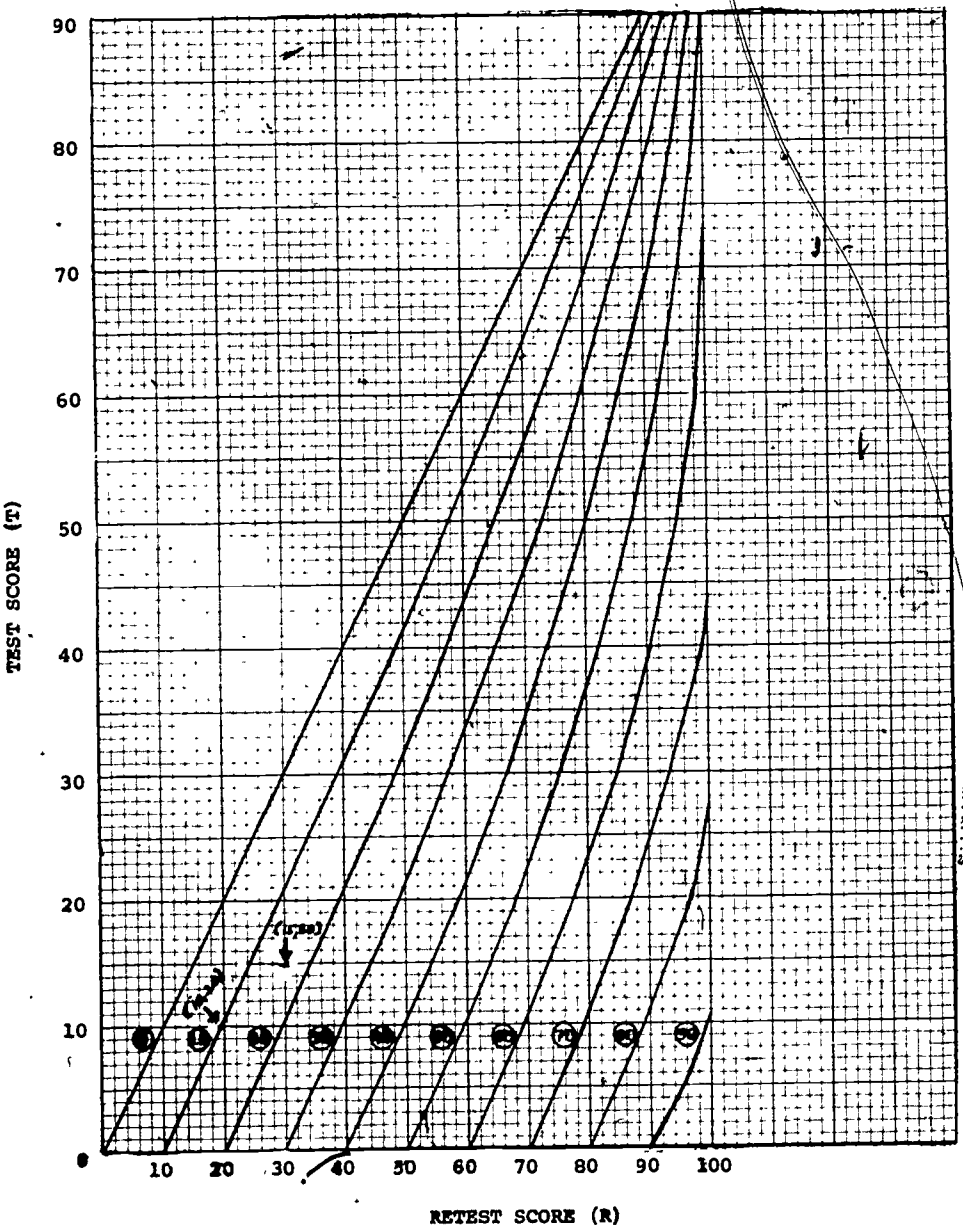


FIGURE 15(A)

WEIGHTED ACHIEVEMENT/COMPETENCE SCORES*

Grouped according to amount of achievement (10 point intervals):

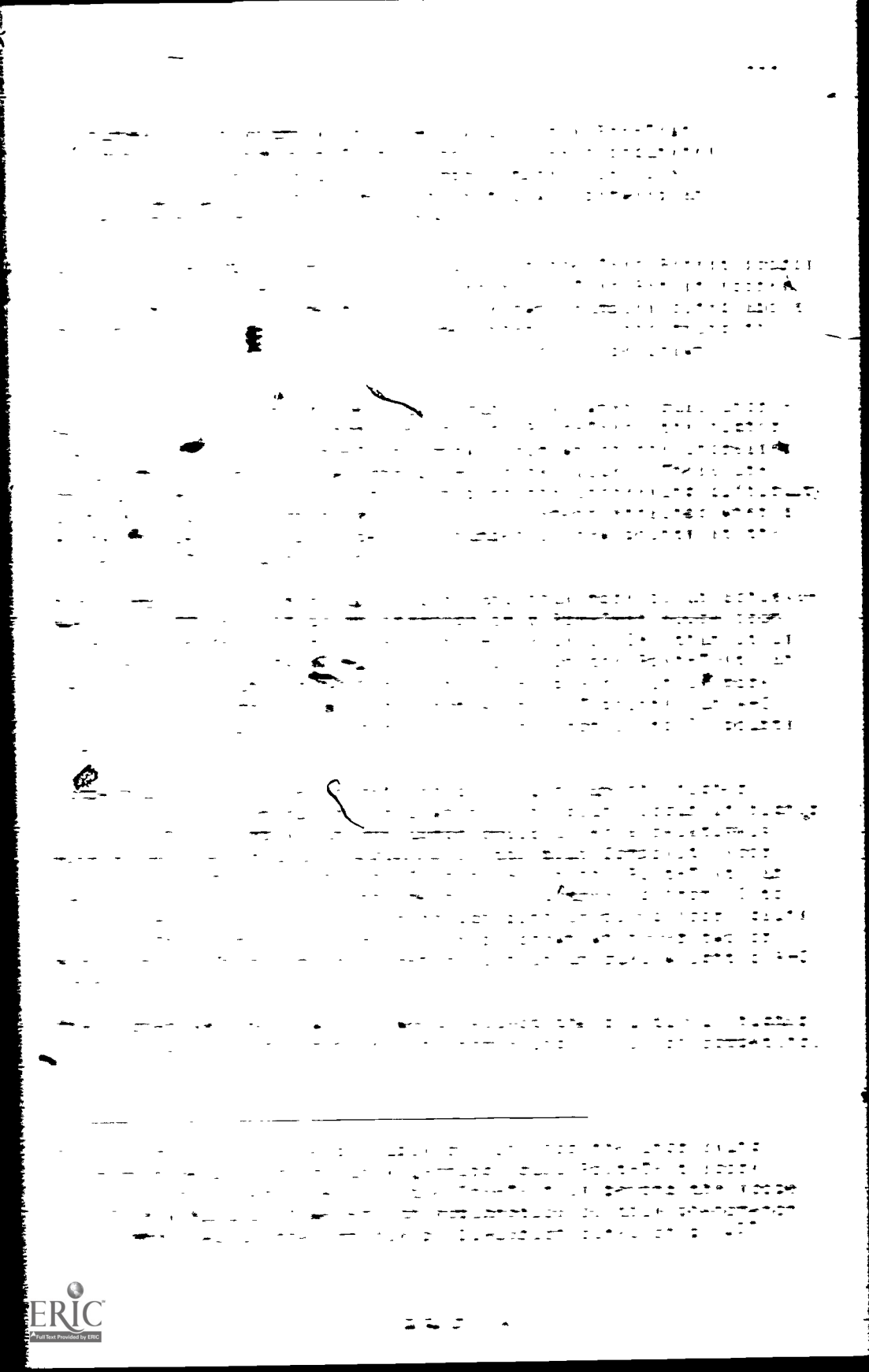
Pre → Post = weighted A+C Score

10 Points	20 Points	30 Points
10 → 20 = 10	10 → 30 = 20	10 → 40 = 31
20 → 30 = 10	20 → 40 = 20	20 → 50 = 32
30 → 40 = 12	30 → 50 = 22	30 → 60 = 33
40 → 50 = 12	40 → 60 = 25	40 → 70 = 36
50 → 60 = 13	50 → 70 = 28	50 → 80 = 40
60 → 70 = 14	60 → 80 = 32	60 → 90 = 56
70 → 80 = 18	70 → 90 = 49	70 → 100 = 59
80 → 90 = 30	80 → 100 = 55	
90 → 100 = 50		

40 Points	50 Points	60 Points
10 → 50 = 41	10 → 60 = 50	10 → 70 = 61
20 → 60 = 42	20 → 70 = 53	20 → 80 = 62
30 → 70 = 44	30 → 80 = 56	30 → 90 = 66
40 → 80 = 47	40 → 90 = 66	40 → 100 = 75
50 → 90 = 54	50 → 100 = 69	
60 → 100 = 65		

70 Points	80 Points	90 Points
10 → 80 = 71	10 → 90 = 81	10 → 100 = 90+
20 → 90 = 74	20 → 100 = 89	
30 → 100 = 79		

* See Appendix G for derivation formula.



compared to a lower A-C value. The relatively higher level of difficulty can be the basis for inferences of higher levels of effectiveness. Such inferences will hold whether the Test/Retest comparisons are being made between:

1) item sets
2) individual trainees
3) trainee subgroups
4) different training sequences (provided that the same test instrument is administered to both groups of trainees

It should be clear from the above discussion that it is necessary to consider not only the statistical significance of the Test/Retest score gains that occur but also the levels at which those gains occur when drawing inferences concerning the impact of the training experience on levels of subject matter competence.

NOTE: Whether weighted or unweighted scores are more valid indicators of levels of achievement (when raising scores from various Pre-Test levels) cannot, at this time, be determined for each situation in which a test instrument is administered under a Test/Retest design. At the risk of complicating matters further, it needs to be stated that very little is known concerning comparative difficulty at different levels along a score range. One can hypothesize a difficulty curve in the form of a parabola reflecting the extreme difficulty in raising a score at both the lower and upper ends of the score range, with the mid-range being the area where score increases are easiest to effect. This question is being studied by analysis of a series of test data. For example, it is interesting to note that in Figure 13, trainees with Pre-Test scores below 50 gained an average of 26.7 raw percentage points compared with an average of only 18.6 among those whose Pre-Test scores were 50 or higher. However, any definitive conclusions concerning score level and difficulty will have to await more extensive research in this area. Thus, the discussion of weighted and unweighted scores, even though there is evidence cited for the validity of the former, was presented to provide possible approaches to the assessment of level and magnitude of score change. While the approaches employed may differ, the important point is that the interaction of the level of score change and the magnitude of that change be considered an integral part of the analysis of training effectiveness. Collective experience in training evaluation of the type described in this Manual may well show that the use of the Achievement/Competence score (especially when weighted) is an important refinement of testing methodology.

In terms of standards for trainee post-instruction competence, however, the basic criterion will not be (as with the case of assessing training impact) their Test/Retest score changes. The ultimate indicator of levels of competence will be the Post-Test scores. A trainee who begins with a Pre-Test score of 10% and obtains a Post-Test score of 75% may have learned more (attained a higher level of achievement) and have demonstrated higher motivation and other desirable qualities than one who shows an increase from 75 to 90%. But the 90% score must still stand as an indicator of a higher level of post-instruction competence. From the point of view of the supervisor who will assign the trainee to a specific job in the post-training period, the higher level of competence attained by the second trainee may be more important than the amount of achievement displayed by the first trainee. However, the selection of one trainee over another is a job specific decision that would have to be made by supervisors considering both the requirements of the job and the qualifications of the trainees being considered. It is, therefore, important that the maximum amount of assessment data for each individual be made available in the post-instruction period.

Summary

The tables provided in Figures 16 and 16A are designed to provide summary information about the test instrument, the training, and the trainees. They illustrate what the test instrument will show when the analysis has been completed.

1. Part I (Table 16) The figures entered in column A (Pre-Test Level) are the total percent of items answered correctly on the total test and on subsets of items; and for individual items, the percent of times they were answered correctly. (For example, if a test of 100 items was taken by 50 trainees, there is a potential of 5,000 correct answers. The percentage of those 5,000 which were actually correct on the Pre-Test is the number that is entered at the top of this column in the row labeled "Total Subject Matter".) Comparable entries apply to the Subset rows. For individual items, however, the percentage of trainees who answered each one correctly is entered lower in this same column.

The values in column A have been given the caption of "Difficulty", on the assumption that the easier items and subsets are answered correctly more often.

The figures entered in column B of Table 16 contain similar information about the Post-Test — the percent of items answered correctly on the total test and item subsets; and for individual items, the percent of times they were answered correctly.

The difference between the Pre-Test and Post-Test score levels is entered in column C (Amount of Change). The column values reflect the effectiveness of the training sequence -- how much of the subject material has been transmitted to the trainees.

Column D (Direction of Change) is divided into four parts, showing the direction of the change between Pre- and Post-Test. It shows what % of total test subsets and individual items remained correct or incorrect on both tests, which went from incorrect to correct, and which went from correct to incorrect. This tells something about the quality and consistency of the test material.

2. Part II (Table 16A) In this part, column A contains the mean Pre-Test scores of the total trainee group and of trainee subgroups, and the scores of the individual trainees on the Pre-Test. This column represents how prepared the trainees were at the beginning of training (i.e., how competent they already were with the subject material to be covered).

FIGURE 16

ANALYSIS SUMMARY PROFILE II

PART I	A	B	C	D			
	PRE-TEST LEVEL (1)	POST-TEST LEVEL (2)	AMOUNT OF CHANGE (2-1)	DIRECTION OF CHANGE			
				C → C	C → I	I → I	I → C
	% correct answers of total possible	% correct answers of total possible	no. of % points (plus or minus)	%	%	%	%
TOTAL SUBJECT MATTER	40	78	+38	38	2	20	40
SUBJECT ITEM SET 1	35	80	+45	35	0	20	45
ITEMS	% trainees correct	% trainees correct	no. of % points (plus or minus)	%	%	%	%
Item 01	30	70	+40	24	3	14	43
Item 02	28	86	+58	28	0	14	58
Item 24	42	66	+24	38	4	30	28
Item 25	51	79	+28	50	1	20	29
SUBJECT ITEM SET 2	45	82	+37	46	1	17	36
ITEMS	% trainees correct	% trainees correct	no. of % points (plus or minus)	%	%	%	%
Item 01	40	65	+25	35	5	30	30
Item 02	51	90	+39	51	0	10	39
Item 24	57	87	+30	57	0	13	30
Item 25	62	90	+28	62	0	10	28
SUBJECT ITEM SET 3	32	82	+50	31	1	17	51
ITEMS	% trainees correct	% trainees correct	no. of % points (plus or minus)	%	%	%	%
Item 01	10	50	+40	2	8	42	48
Item 02	25	65	+40	20	5	35	45

FIGURE 16A

ANALYSIS SUMMARY PROFILE II (cont'd)

PART II	A PRE-TEST LEVEL (1)	B POST-TEST LEVEL (2)	C AMOUNT OF CHANGE (2-1)	D AMOUNT & LEVEL OF CHANGE
	mean score among trainees	mean score among trainees	mean no. % points (plus or minus)	weighted A+C values (see Figure 15A)
TOTAL TRAINEE GROUP	32	84	+ 52	58
TRAINEE SUBGROUP #1	14	70	+ 56	57
INDIVIDUALS	SCORE	SCORE	no. of % points (plus or minus)	
Trainee 01	12	60	+48	49
Trainee 02	18	72	+54	55
Trainee 29	10	40	+30	30
Trainee 30	22	85	+63	67
TRAINEE SUBGROUP #2	40	89	+ 49	60
INDIVIDUALS	SCORE	SCORE	no. of % points (plus or minus)	
Trainee 01	56	92	+36	54
Trainee 02	65	89	+ 24	44
Trainee 29	37	87	+50	59
Trainee 30	49	76	+27	35
TRAINEE SUBGROUP #3	52	76	+ 24	34
INDIVIDUALS	SCORE	SCORE	no. of % points (plus or minus)	
Trainee 01	43	59	+16	26
Trainee 02	41	78	+37	46

The mean score of the total training group and trainee subgroups, and individual trainee scores on the Post-Test are entered in column B. These figures show the amount of competence displayed by trainees at the end of the course, collectively and individually.

The differences between the scores on Pre- and Post-Tests are given in column C of Part II. This shows the amount of achievement the trainees displayed -- how much they learned during the training.

The Amount and Level of Change. In Part II of the table, column D shows the amount and level of the change from Pre- to Post-Test scores, as represented by a score value derived from the curves like those shown in Figure 15A. This represents the amount of competence and achievement displayed by trainees.

Note: Part I of the table can be repeated for trainee subgroups, and even for individual trainees, if such detail is deemed useful. Similarly, Part 2 of the table can be repeated for each subject subset, and for individual items.

PRESENTATION OF DATA: THE EVALUATION REPORT

When the entire Test/Retest procedure is completed, including analysis, it will be necessary to gather all of the information about it into some form of report -- either for program administrators, funding agencies or government departments, or simply for the record. While a number of charts and tables will have been generated during the course of designing and implementing the instrument, it is best to summarize the data from these sources in an easily accessible form. Such a summary should state clearly what was learned, and what implications can be drawn from what happened.

It is recommended that a report should contain the following:

1. Description of the Test Instrument. This should be a simple statement, describing the instrument by the number of items it containd, how the items were grouped (if they were), with any additional information that might be pertinent.

2. Description of the Training Group. Information on the number of trainees and any special characteristics by which they were grouped should be stated.

3. Dates of Applications. The dates should be noted, together with a brief description of the circumstances where necessary. If any radical differences between the two administrations occurred, these should also be noted.

4. The Curriculum Audit. A brief description of the type of audit conducted, concurrent or retrospective and, if the latter, who conducted it, should be provided. The rating form employed should be attached to the write-up to supplement the report.

5. Reporting Trainee Analysis Results. The results of the assessment will be described in three ways.

A. Results for the total training group: Include percent or number correct for Test and Retest, percent and direction of movement from one administration to the next, and whether that movement is statistically significant. The level of that movement should also be noted.

B. Results for trainee subgroups: The same information should be given here.

C. Results for individual trainees: If the trainee group was small, these results could be listed here. In most cases, however, it would only be necessary to refer the reader to the appropriate table, and note here any trends.

D. Interpretation of trainee results: Based on the data presented above, some preliminary interpretations should be offered.

6. Reporting Item Group Analysis Results: Since the instrument is designed to measure the effectiveness of the training as well as the change in competence among trainees, the results will be further broken down as follows:

A. Results of overall test: Number of items correct on the Test and Retest should be listed, with percent of change and whether or not that change is significant. Number of items that were answered correctly on both applications, incorrectly on both, and that went from correct to incorrect and incorrect to correct should be listed.

B. Results for item subgroups: The same information should be given for item subgroups.

C. Results for individual items: Unless the test was extremely short, all the information for individual items need not be reproduced here. However, specific trends should be noted.

D. Interpretation: Again some interpretation should be offered.

7. Level and Magnitude of Score Movement: This information (as described on pp. 103-113) in combination with the results of the application of tests of statistical significance comprise the major data set from which inferences concerning training effectiveness and trainee achievement will be drawn. The data on score levels and magnitude should be presented in the form of tables of the type shown in Figures 14 and 15A. (Actually only one or the other set of tables should be presented depending upon whether weighted or unweighted Achievement/Competence (A+C) scores were computed.) Following the guidelines and examples provided on pp.104-112, tables (displaying Test, Retest and A+C scores) with brief summaries should be presented for

a) the total trainee group
b) the individual trainees
c) trainee subgroups (when available)

8. General Summary Statement and Recommendations: An overall statement of the success, failure, or uncertain performance of the instrument, with regard to its usefulness as an evaluation of trainees and training, should be made.

9. Attachments: It is suggested that the following materials be attached to the report where their reproduction would not involve excessive cost or labor:

1. The Item Specification Table (see Figure 1)
2. The Curriculum Audit Results (see pp. 72-73)
3. The Data Analysis Summary Profile (see Figure 13)
4. The Unweighted (or Weighted) Achievement/Competence (A+C) Score Tables describing the "Level and Magnitude of Test/Retest Score Change" (see Figures 14, 15, and 15A)
5. The Data Analysis Summary Profile II (see Figures 16 and 16A)

CHAPTER VIII

ASSESSING THE TEST INSTRUMENT

An analysis of items in the form of an assessment of Test to Retest response patterns has been suggested in order to provide additional information concerning the relative effectiveness of instruction for each subject area and for each trainee within subject area (see pp. 89-90). This analysis also attempted to identify those items which were relatively ineffective for their intended purpose of assessing levels, and changes in levels, of subject matter competence (pp. 91-98). Thus, some data on the effectiveness of the test instrument itself was provided by the item pattern analysis. However, a more quantitative approach to assessment of test effectiveness can be employed when staff-time permits. This involves analysis of the responses given by the examinee group to one administration of a series of items.

ITEM ANALYSIS

Three kinds of data are derived from the analysis of individual items:

1) The difficulty level of an item (defined as the total percentage of examinees getting the item correct).

2) The discriminating power of an item (defined by the degree to which an item differentiates between high and low scoring examinees).

3) The relative effectiveness of the item's distracters (defined as the degree to which the examinees respond to the item's incorrect alternatives).

Procedural Steps

For purposes of illustration, the coverage of item analysis procedures will refer to the responses of 45 examinees to test items assessing competence in the area of quantitative research methods.

1. Rank the examinees from high to low according to total test score.

2. Select out the upper one-third of the examinees (i.e., 15) and the lower one-third.

3. For each individual test item, tabulate the number of examinees in the upper and lower segments who selected the correct response and each of the distracters. These can be recorded on an item card designed specifically to illustrate each item with its specific characteristics. (see Figure 17)

4. Compute the estimate of item difficulty. The difficulty level of an item is computed by dividing the total number of correct responses to that item (ΣC) by the total number of examinees in both groups (N).

$$\text{Difficulty (D)} = \frac{\Sigma C}{N} \times 100$$

The item difficulty for the sample data (Figure 17) is:

$$D = \frac{15}{30} \times 100 = 50\%$$

Note: The values of D can range from 0 to 100%; the larger the value the easier the item.

5. Compute the estimate of discriminatory power. The discriminatory power of an achievement test item is computed by subtracting the number of correct responses in the lower group (C_L) from the number of correct responses in the upper group (C_U), and dividing the result by the number of examinees in the upper group (N_U).

$$\text{Discriminatory Power (DP)} = \frac{C_U - C_L}{N_U}$$

For the sample data, the estimate of discriminatory power is:

$$DP = \frac{12-3}{15} = .60$$

6. Assess the effectiveness of item distracters. A distracter is considered effective if more examinees in the lower than in the upper group select it as the correct answer. The effectiveness of each distracter of an item can be determined simply by observation of response frequencies for an item. (An example is provided below.)

Interpreting the Item Analysis Data

The general effectiveness of individual items will be judged on the basis of an assessment of all of the item characteristics

FIGURE 17

SAMPLE CARD FORMAT WITH ITEM CHARACTERISTICS

Quantitative Research Methods

Subject Matter: Demographic Analysis
Behavioral Outcome: Ability to Calculate

ITEM: Approximately how long does it take for a population to double if the annual growth rate is equal to .03 (3%)?

a. 15 years
b. 23 years
c. 31 years
d. none of the above

ITEM ANALYSIS RESULTS (12/73)

Alternatives	a	(b)	c	d	no response	multiple response
Upper 1/3 examinees	0	12	0	3	0	0
Lower 1/3 examinees	0	3	12	0	0	0

Difficulty level: 50%
Discrimination Index: .60

(comments:)

provided by item analysis. How this data is employed in rating an item's effectiveness for its intended purpose can be illustrated with the results of the sample analysis for the item in Figure 17.

> The item is in the middle difficulty range as evidenced by the fact that 50% of the examinees (15 out of 30) got the item correct. The item discriminates in a positive direction as shown by the fact that 12 out of 15 examinees in the upper group got the item correct while only 3 out of 15 in the lower group did so. The discrimination index is quite high .61 indicating that for an item of 50% difficulty, it is performing effectively since it distinguishes adequately between the high and low competence groups
>
> There is, however, wide variation in the effectiveness of the item's distracters. Distracter a is completely ineffective since it attracted no examinees from either the upper or lower group. Distracter c is functioning at maximum effectiveness since it attracted no examinee in the upper group and almost all the examinees in the low group. Distracter d is quite ineffective in that it attracted more examinees from the upper than lower competence group. Regardless of the ineffectiveness of two of the distracters the discriminatory power of the item is quite high. If so desired, the discriminatory power could probably be increased by replacing alternatives a and d with more effective distracters (An examination by the staff of why these alternatives were poor distracters should facilitate the design of more effective replacements

The type of item described above would be most appropriate for inclusion into the type of test instrument proposed in this Manual. It is quite effective in distinguishing between high and low competence examinees the criterion for competence being total test score and it is in the medium difficulty range.*

---

* Many experts in the field of educational measurement recommend that for maximum utility, the majority of items comprising the test should be in the middle range of difficulty with some easy items (placed near the beginning of the test) to encourage low ability examinees and some more difficult items to provide a challenge for the more competent examinee (14).

based on the assumption that a sequence of training presenting some, if not all, of the same subject matter will be conducted again in the future. Each item card would be titled by both the subject area and behavioral outcome assessed by that item. In this way, an item card can be cross-referenced and pulled from the file by either content or behavior. (The format for such a card is shown in Figure 1".

When the next training sequence is to be constructed, the appropriate items can be pulled from the file and compiled into a test instrument. Any changes in course content can be accounted for by constructing new items to cover the new material. The new items would then be analyzed in the later item analysis and either added to the item file or discarded.

If it can be assumed that some, or all, of the training material is to be repeated at a future date, then the staff time and effort invested in a post-instruction item analysis will be well spent. Valid and effective items, especially those measuring the more complex behavioral outcomes, are usually quite difficult to construct and are time-consuming. Thus, items shown to be highly effective in assessing competence should be filed away and drawn out when constructing the next test instrument. The large amount of time and effort saved in not having to construct all new items as well as the high quality of items available for test inclusion will more than compensate for the time and effort expended in conducting a comprehensive item analysis.

Doing the item analysis at this time also allows for the incorporation of data on changes in item responses from Pre- to Post-Test to determine which items should be analyzed. Using the table of item response patterns (see Figure 12), it can be seen at a glance which items were most frequently answered correctly and incorrectly on the two applications of the test. Those with a high percentage of C → I would be subjected to scrutiny and probably discarded. Items with a high percentage of I → I should be subjected to analysis to see which alternate (incorrect) answer is being most frequently chosen (on both applications). Items with a high percentage of I → C would clearly be good items, as would items with a high percentage of C → C (although in this case, if there are a great many items in this category it would probably be a good idea to eliminate some of them as making the test too easy.)

2. <u>During Test Construction</u>. If staff time permits and an appropriate examinee group can be assembled, the test items (constructed according to the guidelines provided) can be given a trial administration employing the same testing format

that will be used when assessing the trainee group.* The responses would then be subjected to item analysis and the appropriate item characteristics computed. Those items judged most effective for purposes of assessing levels of competence (based on the item and validity criteria presented in Chapter II) would be selected for the final instrument. Items suspected of being ambiguous or of containing technical defects can either be revised or discarded and replaced by new items.** Replacement items (either from the item reserve or newly constructed) would assess the same cognitive behaviors and subject matter originally assessed by the discarded items. The structure and content of defective items should be compared with the structure and content of their replacements in order to avoid including new items with deficiencies similar to those found in the items they are to replace.

While item analysis data will greatly enhance the test construction process, the opportunity for conducting a tryout of items depends heavily upon the availability of a sample of examinees that is representative (in terms of education level, professional background, etc.) of the population of individuals who will comprise the trainee group. If such a group can be assembled (e.g., possibly with individuals from a training program going on during the period of test construction) then it is strongly recommended that an item tryout be conducted.

When a preliminary administration of the items is not possible, the procedures for test construction outlined in the text (pp. 8-31, and pp. 35-40) and in Appendix C, if closely followed, should be quite adequate for developing a valid (i.e., appropriate, fair and representative) test instrument for the assessment of changes in levels of subject matter competence.

---

* In addition to item analysis data a tryout of items will provide information concerning such factors as the amount of time required to administer the test of X number of items and the appropriateness and adequacy of test instructions and format.

** Every attempt should be made to correct and revise items with suspected deficiencies; discard them only when it is not possible to upgrade them. This is strongly suggested since replacement items will not have been given a tryout and will be of unknown difficulty and discriminating ability and therefore, of questionable effectiveness.

APPENDICES

FIGURES IN THE APPENDICES

APPENDIX A

HISTORY OF THE DEVELOPMENT OF THE METHODOLOGY

Background

The assessment approach described in this Manual is not new to the field of educational evaluation. The measurement and statistical testing of score differences based on two administrations of the same test as an indicator of the degree of learning that has occurred over time, is a common approach used by instructors in academic situations. The Test/Retest paradigm for assessing achievement probably dates back to the beginning of the formal psychometric testing movement or even before. While the evaluation of educational achievement through the application of objective-type test instruments has been widely discussed in a number of excellent textbooks on educational measurement and assessment (see Bibliography), no step-by-step, Manual-type guide has been available where a training administrator who would like to employ such methods could find the necessary information on the planning, construction and administration of the test instrument together with detailed statistical procedures for the analysis and interpretation of the data that results.

The assessment procedures were originally developed in answer to a request to the Division of Social and Administrative Sciences from the Demographic Association of El Salvador. The Division was asked to assist in evaluating a series of training programs in terms of their effectiveness in reaching a set of pre-defined objectives. The Association was conducting four types of population/family planning/human reproduction training programs, each directed toward a different professional and paraprofessional level. One major objective of the programs was that the participants acquire a comprehensive understanding of new subject material, as well as the abilities to apply this new learning to new problem-solving situations. A major focus of evaluation was, therefore, to determine the degree of relevant learning that occurred during the course of training. It was felt that while this would not be considered a comprehensive evaluation, it would provide a measure of the degree to which the training programs were accomplishing some basic, short-term objectives.

Since the assessment of learning was to involve the measurement of change in levels of substantive knowledge as a function of an intervening educational experience, a baseline level of competence from which to measure change was required. The application of an objective-type achievement instrument administered under a Test/Retest design was selected as the most appropriate approach. Rather than construct the test instrument at the Institute or send staff members to El Salvador to conduct the evaluation, it was decided that it would be more appropriate to develop a set of guidelines for structure and content together with an outline of the statistical procedures required for score analysis, to be used by the program administrators themselves in constructing the test instrument and conducting their own assessment study.

Field Applications

Based upon secondary information feedback from the Salvador training experience, the original guidelines for test construction, administration and analysis were revised, expanded and compiled into a draft Manual of procedures for assessing the acquisition and application of new learning derived from a structured training experience.*

The methodology described, while theoretically and intuitively sound, had not been subjected to controlled field-testing. The lack of first hand field experience left questions concerning the methodology's utility and validity unanswered. It was felt that several field applications of the methodology under varying training conditions would be required.

---

* The original guidelines focussed upon changes in levels of subject knowledge from Test to Retest as the measure of program impact and trainee achievement. It was later realized that assessment of knowledge alone was too limited an area of evaluation since it primarily involves tasks which emphasize remembering, either through recall or recognition (1). The focus was, therefore, enlarged to encompass a greater range of cognitive behaviors.

The first field application was conducted at the invitation of the United States Agency for International Development (AID); in a training situation involving a Government sponsored Population/Family Planning Program Seminar-Workshop in Washington, D.C. The field testing was carried out from September 1972 to January 1973.

A second field testing was carried out at the request of the National School of Public Health, Department of Health and Family Protection, Rennes, France which was planning a seven week training program for French health workers at various professional levels. The field work was conducted from October 1973 to January 1974. The numeric data provided in the section on statistical analysis of response data (including the data in APPENDIX F) were derived from this field testing.

The third field study was also carried out at the National School of Public Health in Rennes. This study involved assessment of changes in levels of competence among health professionals from Francophone Africa who were participants in a four month Family Planning and Maternal/Child Health Training Program being conducted under the sponsorship of the Department of Health and Family Protection. The study was conducted from March to October of 1974.

The content of the Manual, while derived primarily from the field testing experiences and preparation of the guidelines paper drew heavily from the writings of various educational specialists whose major works are cited in the Bibliography. Thus, the methodology described is not so much an innovative contribution to the field of educational assessment as it is a comprehensive synthesis of extant experimental design, qualitative and quantitative guidelines for test instrument planning, construction and administration and statistical analytic techniques into a self-contained reference text for conducting a study to assess changes in levels of subject matter competence as a result of participation in a structured training experience.

APPENDIX B

PSYCHOMETRIC THEORY UNDERLYING THE METHODOLOGY

The purpose here is not to inundate the reader with an exhaustive exposition on the complexities of psychometric theory as it relates to achievement testing. The body of literature concerned with this area is so voluminous as to preclude all but the most elementary and non-technical discussion. The assessment guidelines and methodology comprising this Manual should not be accepted, however, without some understanding of the theory governing their effective use. The objective here, then, is to discuss some of the underlying theoretical principles involved as well as to identify the major problems encountered in measuring learning outcomes*.

General Considerations

Measuring educational achievement requires an objective assessment of what a group of students has learned (i.e., their subject matter competence), in one or more relevant subject areas, through a testing procedure employing a set of subject-related tasks. The testing procedure must be structured so that all examinees interpret the tasks in a similar way (to provide a common basis for assessment), and standardized so that the tasks and procedures for administration and scoring are explicit and fixed (to ensure that the same test procedures are followed each time an assessment is conducted). In order that the procedures conform to an achievement test model, the subject material comprising the tasks should be a representative sampling of the significant subject matter dealt with during the course of instruction. If the content

---

* Procedures for translating "Subject Matter Competence" (the specific learning outcome under study) into operational indices of achievement amenable to objective assessment are discussed on pp. 7-24.

of the tasks adequately reflects the relevant subject content of the course work, then measures of success or failure in dealing with these tasks (when administered under controlled testing conditions) will provide the basis for inferences concerning

(a) the effectiveness of the instructional sequence in achieving a specific training program objective and;

(b) the magnitude of the change in the trainees' levels of subject matter competence.

The use of the same test results to assess both training effectiveness and trainee achievement is not new to the field of educational evaluation. According to Cronbach (2), every time a teacher gives a test he is testing his instruction as much as he is testing the student's efforts and achievements.

Testing Design

In order to relate any change in an individual's level of subject competence directly to a specific training sequence, a testing design is required that will provide a measure of the trainee's level of competence prior to the introduction of instruction (i.e., a quantitative assessment of the degree to which a trainee has already acquired what is to be learned). This pre-instructional baseline level of competence is subsequently compared with a similar measure obtained upon the completion of instruction. A statistical analysis of any increases in competence levels from testing to retesting will help determine whether such increases are significant or simply due to chance. The degree of confidence with which inferences can be made which relate the increases in competence to the direct effects of training instruction will depend upon the type of test instrument administration design selected.

The testing design employed in the Manual is the One-Group Pretest-Posttest Design (3) which is represented graphically by:

O_1 X O_2

Where: X = the introduction of a treatment variable whose effect is to measured;

O_1 = a measurement procedure conducted prior to the introduction of the treatment variable and;

O_2 = a measurement procedure conducted following the application of the treatment variable.

In terms of the level of assessment being proposed here, the "X" represents the structured training sequence (i.e., an educational treatment); "O_1" represents the pre-instruction and "O_2", the post-instruction test performance with tasks sampling cognitive competence. Operationally, then, assessment at this level is essentially a statistical determination of the degree to which training instruction elevates the trainee's initial baseline level of subject competence. This implies that a change will occur as a result of instruction and that the magnitude of the change can be measured quantitatively (and related directly to that instruction).

Factors Affecting Measurement Validity

There are a number of factors* related to the technical/structural aspects of the test instrument that must be acted upon, due to their potential confounding effects on the measurement outcome (to the extent that the test results can be rendered invalid for their intended use in measuring changes in levels of subject matter competence and relating the changes to the impact of instruction).

---

* These factors are covered in greater depth in Chapter II.

1. Test item construction Items must be constructed to ensure that an incorrect response means the examinee has not achieved competence in the subject area sampled by the item and not because the vocabulary was vague or too difficult or the sentence structure too complex.

2. Item content validity Inferences concerning subject matter competence cannot be based on items that provide an inadequate (non-representative) sample of the subject areas and abilities covered in the instructional sequence.

3. Levels of item difficulty Test performance is highly sensitive to and strongly influenced by items which are too easy or too difficult.

4. Test directions and statements of test purpose Effects test performance by shaping the examinee's conception (and perception) of the task and by influencing his level of test-taking motivation.

5. Time limits and guessing penalties Individual differences in non-cognitive functions (not directly related to the test behavior being measured) may enter into the assessment when limits and penalties are imposed.

The above list, while not exhaustive, calls attention to the fact that without proper controls test performance is vulnerable to the subtle and profound influences of factors above and beyond those which the test purports to measure.

Extraneous Variations in Test Performance

In an effort to identify potential sources of extraneous influence, it is necessary to consider the effects of other variables beyond those associated with the technical/structural nature of the instrument itself, which may pose a threat to the internal validity of the assessment. (Internal validity refers to the level of confidence which can be ascribed to findings which infer a causal and direct relationship between the sequence of instruction and the level of subject competence.) These variables can function as "plausible rival hypotheses," offering alternative explanations for the O_1 to O_2 difference (i.e., Pre- to Post-Test

score increases), rival to the inference that "X" (i.e., training) causes the difference (4).

Awareness of the fact that such variables can produce effects confounded with the effects of the training sequence is particularly important given the nature of the One-Group Pretest-Posttest design employed in the assessment. This design, like many employed in educational evaluation, is a quasi-experimental design and, unlike the true experimental design*, is employed "in situ" where necessary controls cannot always be implemented. Further, the practical necessities of training program operations most often preclude the use of a control group (i.e., a group that receives both administrations of the test without the intervening educational treatment) against which to measure the significance of change occurring in the training group.

Campbell and Stanley (5), in presenting this design, discuss a number of threats to valid inference. The following is a list of these factors, together with a judgment as to their potential effect on the type of assessment being proposed here.

a. History. Between the two measurement points (i.e., O_1 & O_2) other change-producing events may have occurred in addition to the educational treatment variable "X".

While extraneous outside influences can produce changes in Test/Retest measurements of certain variables (e.g., attitudes and opinions), their effect on subject matter competence would be minimal. (Any activity occurring outside of the formal training sessions, such as homework assignments and informal student discussion of course-related topics, is an integral part of training and not an extraneous variable).

* A highly structured laboratory-type situation where random assignment of subjects to treatment groups as well as other types of controls are employed to reduce or eliminate the effects of variables other than those being measured.

b. Maturation. All of the biological and psychological processes which systematically vary with the passage of time independent of specific events. For example, in between [illegible] and [illegible] students may have grown older, hungrier, fatigued, bored, etc., and any measured differences may reflect these processes rather than the treatment effect.

Since most training programs are conducted on a short-term basis and are for the most part attended primarily by motivated individuals seeking preparation for a specific vocational objective, this factor should have little influence.

c. Testing. The effect of taking the Pre-Test on Post-Test performance. "On achievement tests there is a tendency for examinees taking the test for a second time, or taking an alternate form of the same test, to do better than those taking the test for the first time."*

A systematic Post-Test score increase as a result of having previously taken the test might be minimized if the examinees are unaware that they will take the same test again, if they receive no information as to scores or correct answers on the Pre-Test, and if the second testing is reasonably far removed in time from first testing. However, it can be argued that pretesting is an educational experience to the extent that it may provoke discussion of the test material among trainees and stimulate some of them into a more intensive study of the subject material covered by the test items. Thus, the Pre-Test itself may act as a stimulus to learn. From this viewpoint, the Pre-Test becomes an integral part of the training process, since the primary purpose of training is to educate.

---

* Research findings concerning the existence of "pre-test effects" on Post-Test performance are equivocal. A discussion of this research is provided in an article by Apter, et al. (6). The findings of their own research show that "pre-testing has no significant, consistent effect on post-test results."

Instrumentation — Autonomous changes that occur in the measuring instrument from one administration to the next must account for O_1 – O_2 differences. This is changes that occur in the standards or criteria of measurement from Test to Retest. For example, when human observation of the child for measurement at two points in time, an individual may be fatigued, bored or more skilled in observation during the second occasion. Also, when essay-type test items are used, the grading standards may shift between one testing and the next.

This factor will be controlled for by employing a standardized testing instrument and objective scoring procedures.

Statistical Regression — Students selected for some specific educational treatment because they obtained significantly low scores on an achievement test (considered here as O_1) will upon retesting average higher on the same test or a parallel test (i.e. O_2). This effect (without elaborating on its source) is independent of treatment effects, Test-Retest practice effects, etc.

The focus of training assessment is to be on subject matter competence. Since it is reasonable to assume that subject competence will at the pre-instruction level be normally distributed among trainees (as measured by the Pre-Test), the effects of statistical regression will be minimal (if any). The selection of low achievement individuals is not common practice in health-related training programs in such areas as family planning, maternal child health, etc.

Practical Considerations for Selecting the Design

It can be concluded from the above discussion that valid inferences concerning the effects of short-term instruction on subject competence can be drawn from assessment procedures employing the One-Group Pretest-Posttest design. It must be admitted, however, that the selection of a quasi-experimental design was based more on necessity than on choice of the most valid design for assessment purposes.

The most valid approach would be to conduct the assessment under conditions representative of true experimental design. That is, individuals would be randomly assigned to one or the other of two groups (i.e., an experimental group to receive instruction and a control group receiving no instruction). Both groups would receive the Pre-Test and Post-Test and the changes that occurred within each group would be compared. The score changes occurring within the control group (reflecting effects on scores that operate in the absence of training) would be statistically partialled out of the score changes in the experimental group and the resulting difference would be attributed to the training sequence. A causal relationship between instruction and significant (experimental group) post-score increases can be inferred with a high degree of confidence since the true experimental design can be considered as actively controlling the extraneous effects of history, maturation, testing, instrumentation, etc. The difference for the experimental group between Pre-Test and Post-Test cannot be explained by main effects of these variables as they are found to effect both the experimental and control groups (7); therefore, the change is attributed to the effects of training.

One major working assumption underlies the incorporation of the One-Group Pretest-Posttest design into the assessment methodology comprising this Manual. This assumption is that the type of training situation where the assessment methodology will most often be employed is one in which the only individuals available for testing are the participants themselves.

Many technical as well as practical considerations preclude the implementation of rigid controls and the use of a student "control group" in most educational settings. This is especially true in training situations where a sponsoring agency conducts a program involving "non-resident" participants.

The training programs involved in the field testing of the assessment methodology are cases in point.

The Government Agency-sponsored Population/Family Planning Training Program conducted in Washington D.C. was attended by health personnel from a number of developing countries throughout the world. Since these individuals were in the country specifically to participate in the training, it would not have been appropriate to divide the group randomly into two subgroups with one to receive training and the other to serve as control. Nor was it possible to secure an independent group of subjects, matched with the trainee group on relevant parameters (e.g., educational level, professional background, English language proficiency, etc.) to serve as the control.

The lack of appropriate individuals to assemble into a comparable control group was also evidenced in the field tests involving both the French and Francophone African Training Programs in Health and Family Protection conducted at the National School of Public Health in Rennes, France.

The test results obtained under a quasi-experimental design can be used to assess training effectiveness if the evaluator is willing to accept certain assumptions about what would have happened to the variable being assessed if the individuals had not been exposed to the sequence of instruction. Essentially, the evaluator assumes that the observed changes were due to the impact of the educational program and that the changes would not have occurred if the trainees had not been exposed to the program (8). For example, in the first Rennes Training Program, where the assessment results display significant increases in levels of competence in the three major subject matter areas, it is an appropriate assumption that the trainees would not have shown such changes in a comparable period of time if they had not participated in the course of training.

The inferential power of the assessment results is greatly enhanced provided that systematic guidelines in test construction and administration are implemented and appropriate statistical tests and procedures are employed in the analysis of resulting test data.

APPENDIX C

GUIDELINES AND RULES FOR CONSTRUCTION OF SPECIFIC FORMS OF OBJECTIVE TEST ITEMS WITH EXAMPLES (9)

Constructing Multiple-Choice Items

The standard multiple-choice item consists of a stem and a set of alternatives or response options. The stem can take the form of either a complete question or an incomplete statement while the alternatives provide possible answers or completions of the statement. (The alternatives will consist of one correct or best response together with two or more misleading options, called distracters.) The following rules, guidelines and suggestions are based on this standard design.

1. A definite problem should be recognized from the item stem. The test taker should be able to tell, from reading the stem of the item, what kind of competence he is expected to demonstrate in answering. An item with an incomplete idea in the stem, meaningless in itself, will be confusing and take more time to figure out. An example:

Developing countries:

a. rarely formulate population policy.
b. have strong conservative elements operating against the adaptation of family planning.
c. must develop population policy in order to set goals and mobilize resources.
d. are among the most interesting places in the world.

Here the test taker is forced to read each response before knowing what information is being looked for. Incomplete ideas in the stem generally make it necessary to write lengthy alternatives, and the alternatives will frequently cover a number of unrelated ideas. It is best to include as much as possible in the stem, to ensure uniformity in the alternatives and reduce reading time. The example would be better if worded as follows:

National population policy is important for developing countries because it will:

a. have a direct effect on the size of the population.
b. set goals for allocation and mobilization of resources.
c. put them in the company of the advanced nations.
d. make them more attractive to visitors.

Here the stem of the item meets a criterion which serves as a useful check on adequacy of initial problem statement: it could be used as a short-answer type item, as "Why is national population policy important for developing countries?"

2. Avoid having to repeat words in each alternative. If such words are included in the item stem, the clarity of the item will be increased and reading time decreased. Thus the item that follows:

A limitation of teaching by external rewards is that:

a. punishment is more effective than external rewards.
b. many students will not be influenced by external rewards.
c. the learner's behavior may not change as a result of external rewards.
d. external rewards may become more important than the act itself.

... might be better worded as below:

One of the disadvantages of the use of external rewards in teaching is that external rewards are likely to:

a. be less effective than punishment.
b. influence only a few of the students.
c. change the learner's behavior.
d. become more important than the learning itself.

3. Avoid negative statements in stems and responses. Unless significant learning outcomes require them, negatives (i.e., no, not, least) are best avoided because they are easily overlooked. While test takers are expected to read items and responses carefully, it is unfair to penalize someone for so obvious an oversight. Also, the learning outcomes should

stress the acquisition of and the ability to use and apply the best or most important methods, principles, facts, theories and not the ability to select the "exceptions to the rule" as measured by the typical "negative" item. If for some reason you must use a negative, underline it: e.g., Which one of the following is not a type of oral contraceptive?

4. Use novel material and situations in formulating problems that aim to measure understanding of or ability to apply principles. As in the case of items taken verbatim from books or lectures, you may end up measuring ability to recognize or remember material (rote memory), rather than ability to use what was learned. Of course, new material must be carefully selected; it should not require knowledge and/or understanding of areas not covered in the course. While the situation must be new to the examinee, try to select material as close to the illustrations used during the course as possible.

5. Be sure no unintentional clues to the correct answer have been written into the item stem. There are many ways in which clues can slip in. Some examples follow:

A) In family planning education programs built around the availability of transistor radios in particular rural areas, one key element should be:

a. scheduling of programs when particular audiences are likely to be listening.
b. talks by university professors.
c. scheduling programs when children are asleep.
d. standardizing the message for all parts of the country.

Here the clue is the word "particular," which appears both in the stem and the correct response. The test taker will be likely to see this association and pick correctly. The best way to deal with this example would be to take "particular" out of the stem, where it doesn't add anything to the meaning anyway.

B) The Ministry of Health has commonly been selected as the principal organization to run population programs because:

a. it usually has responsibility for major activities concerned with population growth.
b. it is always well financed.
c. its clinics can provide services needed for a population program.
d. the medical profession has never failed to initiate and operate new programs effectively.

Here the clue is in the use of the words "usually," "always," and "never" in three of the responses. Answer c is the only unambiguous alternative and thus most likely to be chosen. Ambiguous terms such as these should be avoided in any case, but using them in some responses and not in others will often give the answer away.

C) The net reproduction rate measures an:

a. annual increase of births over deaths.
b. annual rate at which women are replacing themselves on the basis of prevailing fertility, assuming no migration.
c. decennial growth rate of the population.
d. per generation growth rate.

The article "an" can only go with the two alternatives that begin with vowels (a & b), thus reducing the choice to two alternatives. Items should be read over carefully for grammatical matters, particularly for grammatical agreement between the item stem and all the responses.

In addition, the item above gives a further clue in the great difference in length between the correct response and the other alternatives. (Since correct responses usually require qualifications, they tend to be longer than the distracters.) Be sure that you don't give away the answer by trying to squeeze in all the information needed to make it correct, unless you lengthen the other alternatives as well.

D) When demographers refer to the "population pyramid," what are they referring to?

a. A mathematical formula for predicting population trends.
b. A pictorial representation of the distribution of the population by sex.

c. The hierarchy of the staff of a population/ family planning agency.

The answer is partly given away by reference to a "pictorial representation," which easily refers back to the "pyramid" in the stem. To help make this item less easy to guess, either the phrase "pictorial representation" could be taken out, or the item could be reworded as follows:

In demography, the "population pyramid" is a pictorial representation of:

a. a mathematical formula for predicting population trends.
b. the distribution of the population by sex.
c. the hierarchy of a population/family planning program.

6. Avoid responses that overlap or include each other. In the example below, answers b and d include answers a and c:

An average annual growth rate of 2.8% leads to a doubling of the population in:

a. under 15 years.
b. under 25 years.
c. over 50 years.
d. over 100 years.

If the answer was, for example, 7 years, both a and b would be correct. The chances of guessing would be improved.

7. Do not use a pair of opposite statements as alternatives if one of the pair is correct. Most test takers will limit their choice to one of the two opposing statements, thus reducing a four-choice item to a two-choice item, as in the example:

The doubling of the population expected in the next 4½ years is likely to have what effect on the growth rate of total income?

a. The productivity of investment will increase.
b. There will be no change.

c. The rate of savings will increase.
d. The rate of savings will decrease.

This problem can be avoided by employing two pairs of opposites or eliminating the use of opposites altogether.

8. Use the alternative "none of the above" only when required to measure specific learning. Only in cases where a trainee must be able to determine things that do not apply should "none of the above" be used.

Its most appropriate use would be with items requiring numerical computations where the responses can be classified as unequivocally correct or incorrect. If it is used frequently, it must be the wrong answer some of those times. When it is the right answer, the alternatives that do not apply must be plausible, but must also in fact not apply.

9. Avoid the use of the alternative "all of the above". The alternative "all of the above" creates two significant difficulties. First, test takers may recognize the first response as correct and mark it without reading all of the alternatives. Second, a test taker may recognize two of the alternatives as being correct, and not know about the third. He will still get the item correct without complete understanding, however, by marking "all of the above." It is better in these cases to make the alternatives into a list, and then ask the respondent to check which are correct:

a. 1 & 2
b. 1 & 3
c. 2 & 3
d. All of the above

Interpretive Exercises

An interpretive exercise consists of a series of objective items based on a common set of data (written material, tables, charts, graphs, maps or illustrations). Test items are most commonly of the multiple-choice or alternative response type.

Since all test takers are presented with a common set of data, it is possible to measure a variety of complex learning outcomes. Test takers can be asked to apply principles, interpret relationships, recognize and state inferences, recognize relevant information, develop hypotheses, formulate conclusions, recognize assumptions, recognize limitations, state significant problems, and design experimental procedures. All these are indicators of complex achievement.

The most common method of getting students to demonstrate these abilities has been to ask them to write an essay. The main advantage of the interpretive exercise over the essay-type question is derived from the greater structure provided by the interpretive exercise. Test takers cannot redefine the problem, or arrange their answer to demonstrate only those thinking skills in which they are most proficient. The series of objective items forces them to demonstrate the specific mental abilities called for. It also makes it possible to measure separate aspects of problem-solving ability and to use objective scoring procedures.

The validity of exercises measuring intellectual skills may be questionnable in terms of a Test/Retest instrument except in courses specifically designed for the development of such skills. However, it is felt that objective-type exercises can be useful in determining the trainee's ability to apply new learning, or to reason in a subject area with which he has become familiar during the course of instruction. In addition, the amount of factual material given in the exercises or asked to be provided by the pupil can be controlled: definitions of terms, formulas for calculation, and the like, may be either provided or withheld, thus regulating the difficulty of the test item measuring achievement of a specific learning outcome.

Constructing Interpretive Exercises

There are two major tasks involved: selection of appropriate introductory material and constructing a series of dependent items. Special care must be taken to construct test items that require an analysis of the introductory material -- items that simply measure reading skill or rely on general information apart from what is contained in the material are not useful for the purpose for which the exercise has been in-

tended. The following suggested guidelines will aid in constructing valid interpretive exercises.

1. Select introductory material that is in harmony with the objectives of the course. Interpretive exercises, like other testing procedures, should measure the achievement of specific instructional goals. Success in this regard depends to a large extent on the introductory material, since this provides the common setting on which the specific test items are based. If the introductory material is too simple, the exercise may become a measure of general information recognition, recall or simple reading skill. On the other hand, if the material is too complex or unrelated to instructional goals, it may become a measure of general reasoning ability unrelated to specific learning outcomes. Both extremes must be avoided. Ideally, the introductory material should be pertinent to the course content and complex enough to call forth the mental responses specified in the course objectives.

2. Select introductory material that is new to students. In order to measure complex learning outcomes, the content of the introductory material must contain some novelty. Asking students to interpret materials identical to those used in instruction provides no assurance that the exercise is measuring anything other than rote memory. Too much novelty, however, must be avoided. Materials similar to those used during the course but which vary slightly in content or form are most desirable. Such materials can usually be obtained by modifying selections from textbooks, newspapers, news magazines, and various reference materials pertinent to the course content.

3. Select introductory material that is brief but meaningful. One method of minimizing the influence of general reading skill on the measurement of complex learning outcomes is to keep the introductory material as brief as possible. Digests of articles are frequently available and provide good raw material for interpretive exercises. Where digests are unavailable, the summary of an article or a key passage may provide sufficient material. In some cases, the relevant information is summarized more adequately in a table, diagram, or picture.

4. Revise introductory material for clarity, conciseness, and greater interpretive value. Although some materials (for example, graphs) can be used without revision, most selections

require some adaptation for testing purposes. Technical articles frequently contain long, detailed descriptions of events. On the other hand, news reports and digests of articles are brief but frequently present exaggerated reports of events to attract reader interest. While such exaggerated reports provide excellent material for measuring the ability to judge the relevance of arguments, the need for assumptions, the validity of conclusions, and the like, the material must usually be modified to be used effectively.

5. Construct test items which require analysis and interpretation of the introductory material. There are two common errors in the construction of interpretive exercises which invalidate them as a measure of complex achievement. One is to include questions which are answered directly in the introductory material -- that is, asking for factual information which is explicitly stated in the selection. Such questions measure simple reading skill. The second is to include questions which can be answered correctly without reading the introductory material -- that is, requiring answers based on general information in the area. These questions, of course, merely measure simple knowledge outcomes.

If the interpretive exercise is to function as intended, it should include only those test items which require pupils to read the introductory material and to make the desired interpretations. In some instances, the interpretations will require pupils to supply knowledge beyond that presented in the exercise. In others, the interpretations will be limited to the factual information provided. The relative emphasis on knowledge and interpretive skill will be determined by the specific learning outcomes being measured. Regardless of the emphasis, however, the test items should be dependent on the introductory material, while at the same time calling forth mental responses of a higher order than those related to simple reading comprehension.

6. Make the number of test items roughly proportional to the length of the introductory material. It is inefficient to have pupils analyze a long, complex selection of material and answer only one or two questions concerning it. Although it is impossible to specify the exact number of questions which should accompany a given amount of material, the items presented as examples in this section illustrate a desirable

balance. Whenever possible, it is best to use an exercise that has brief introductory material and a relatively large number of test items.

7. In constructing test items for an interpretive exercise, observe all pertinent suggestions for constructing objective items. The form of test item used in the interpretive exercise will determine the suggestions for construction which have greater value. If common forms of the multiple-choice or alternative-response item are used, the specific suggestions for constructing these item types should be observed. Where modified forms are used, suggestions for constructing each of the various types of objective items should be reviewed for their applicability in construction. Freedom from irrelevant clues and technical defects is as important in interpretive exercises as it is in single, independent test items.

SAMPLE INTERPRETIVE EXERCISE

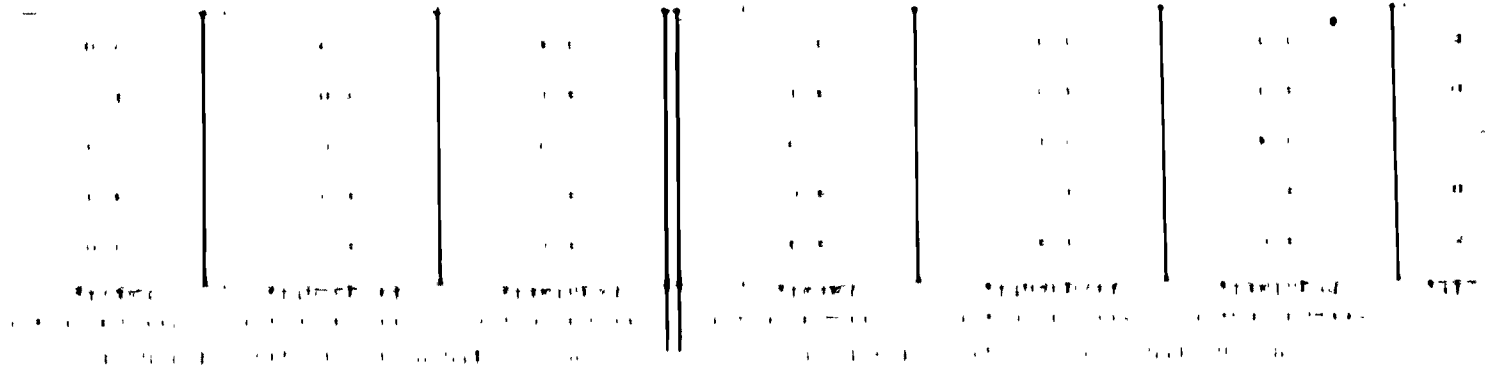
All of the following exercises are to be used with the accompanying chart. See Fig. 21

1. Which one of the following additional pieces of information would be necessary to determine the extent to which the various groups of women in the cities practice family planning methods?

   a. Age of women in each group.
   b. Economic status of women in each group
   c. Availability of contraceptives in each city
   d. Attitude of religious leaders toward family planning

2. Which statement below is supported by the data in the chart?

   a. Women in city "B" have more children than women in any other city, in all groups.
   b. In every case, women with 8 or more years of school have fewer children than women with 7 or less years
   c. In all cases, women in city "C" have fewer children than those in any other city
   d. Women with more education are more likely to practice family planning



[illegible] of the above inferences can be drawn from the data in the chart.

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

If you have any questions concerning what has been covered up to this point, please ask them now.

Do not waste time on items that appear too difficult. You should have to spend no more than [illegible] minutes per item on the average. It is best to go on and attempt all items and then return to the unanswered items. Since the score will be the number of items correct, it is better to make a careful guess than to spend too much time on any one item or leave an item unanswered. Each unanswered item will be counted as incorrect. [illegible] reason for responding to all items is that each item represents one bit of valuable information to the assessment being conducted.

Begin with item [illegible] as soon as you have finished [illegible] [illegible] you have completed all the items. [illegible] although the test must be completed before you leave today's session.

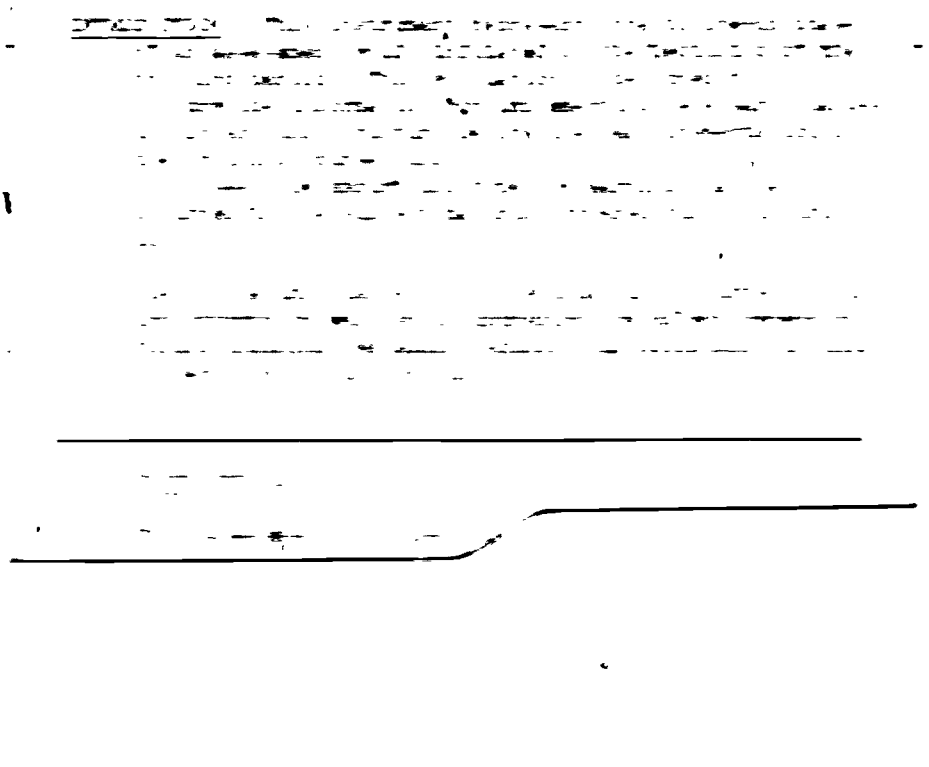
Please return the test booklet together with the answer sheet as soon as you have finished. [illegible]

If there are no further questions, you may begin.

The Post-Test Format

[illegible]

[illegible]



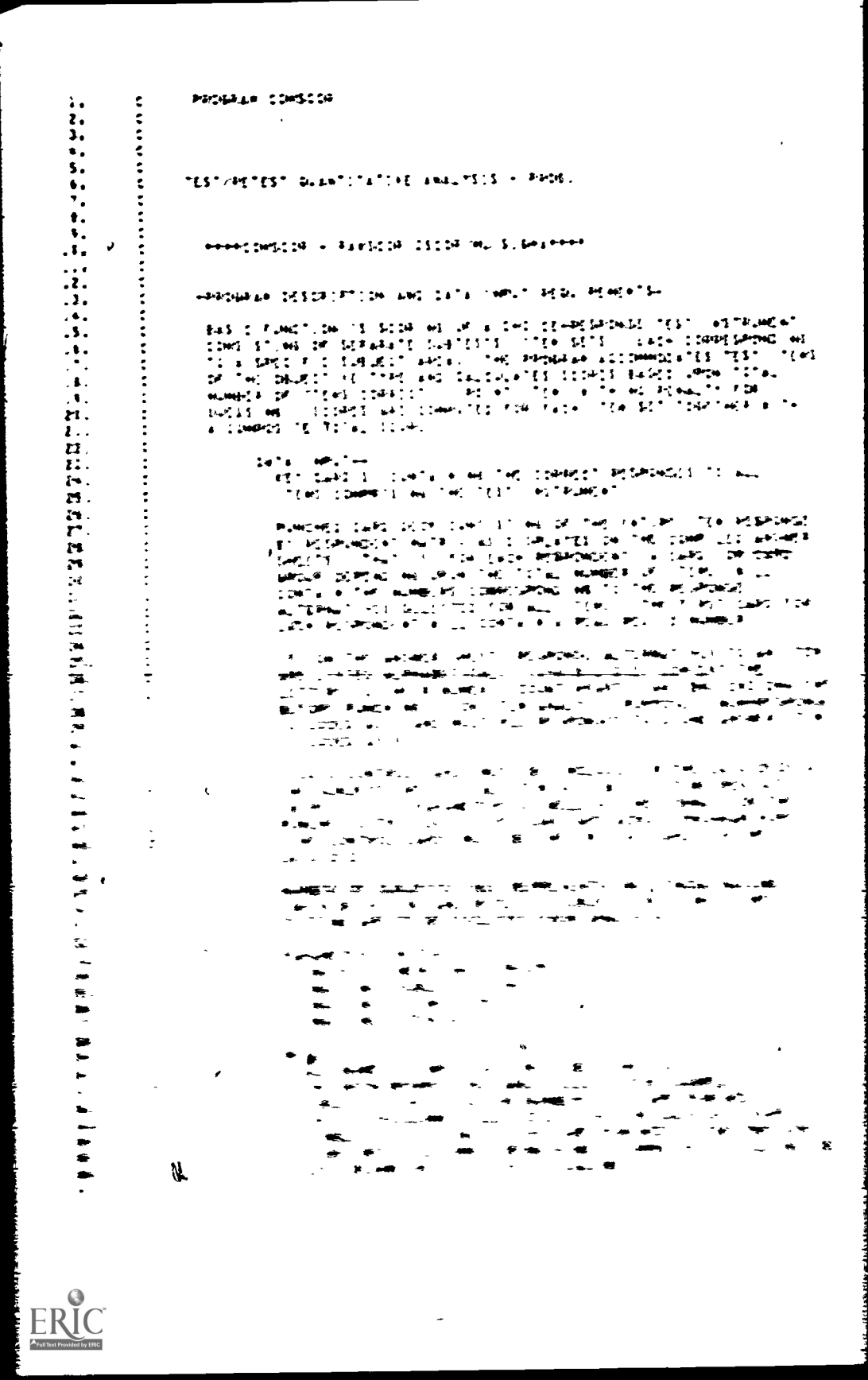
The five FORTRAN IV programs comprising the analysis package were written specifically for the assessment of Test Retest data as described in the main text of the manual.

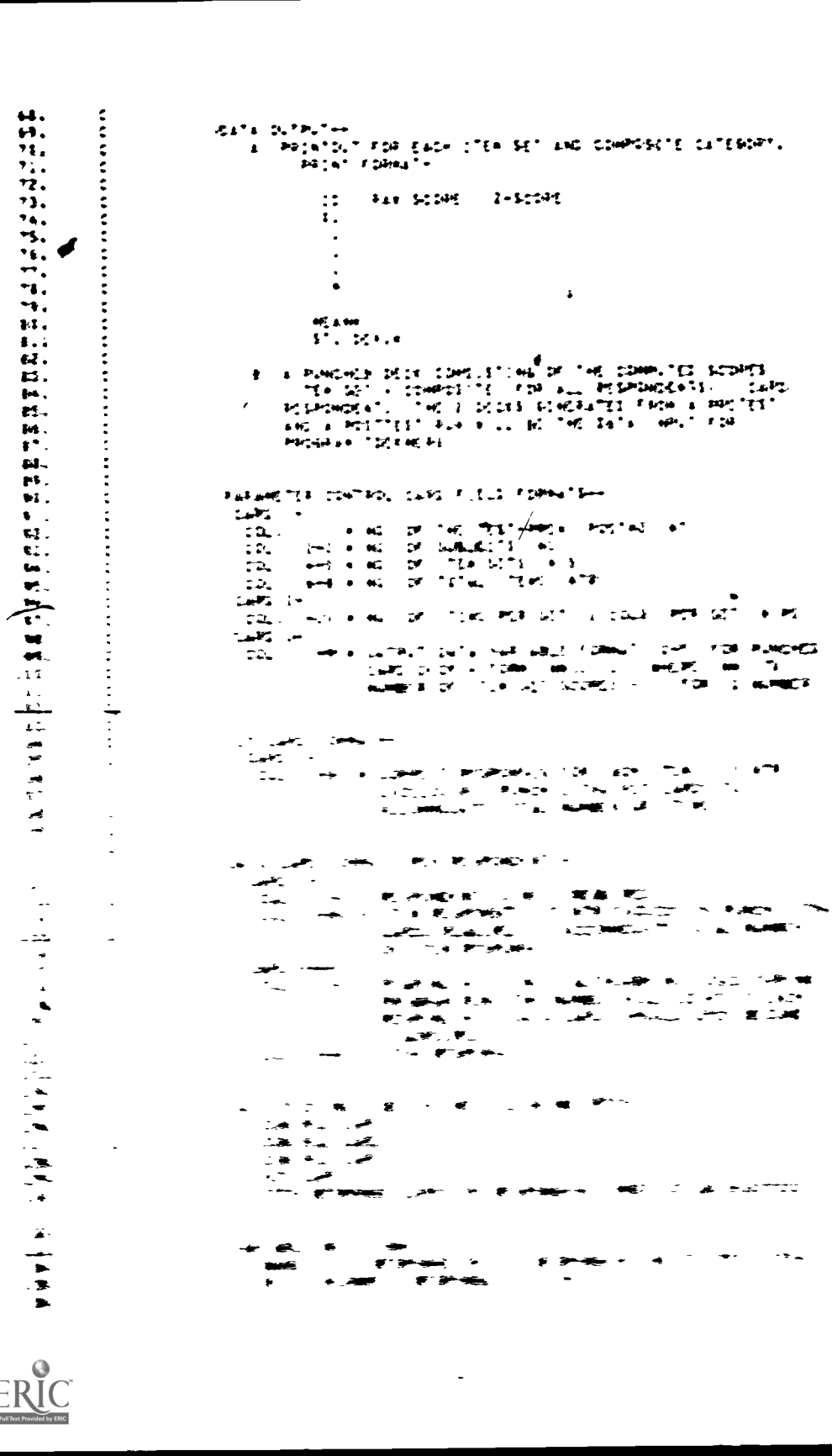
In order to provide the user with the option of [illegible] [illegible]

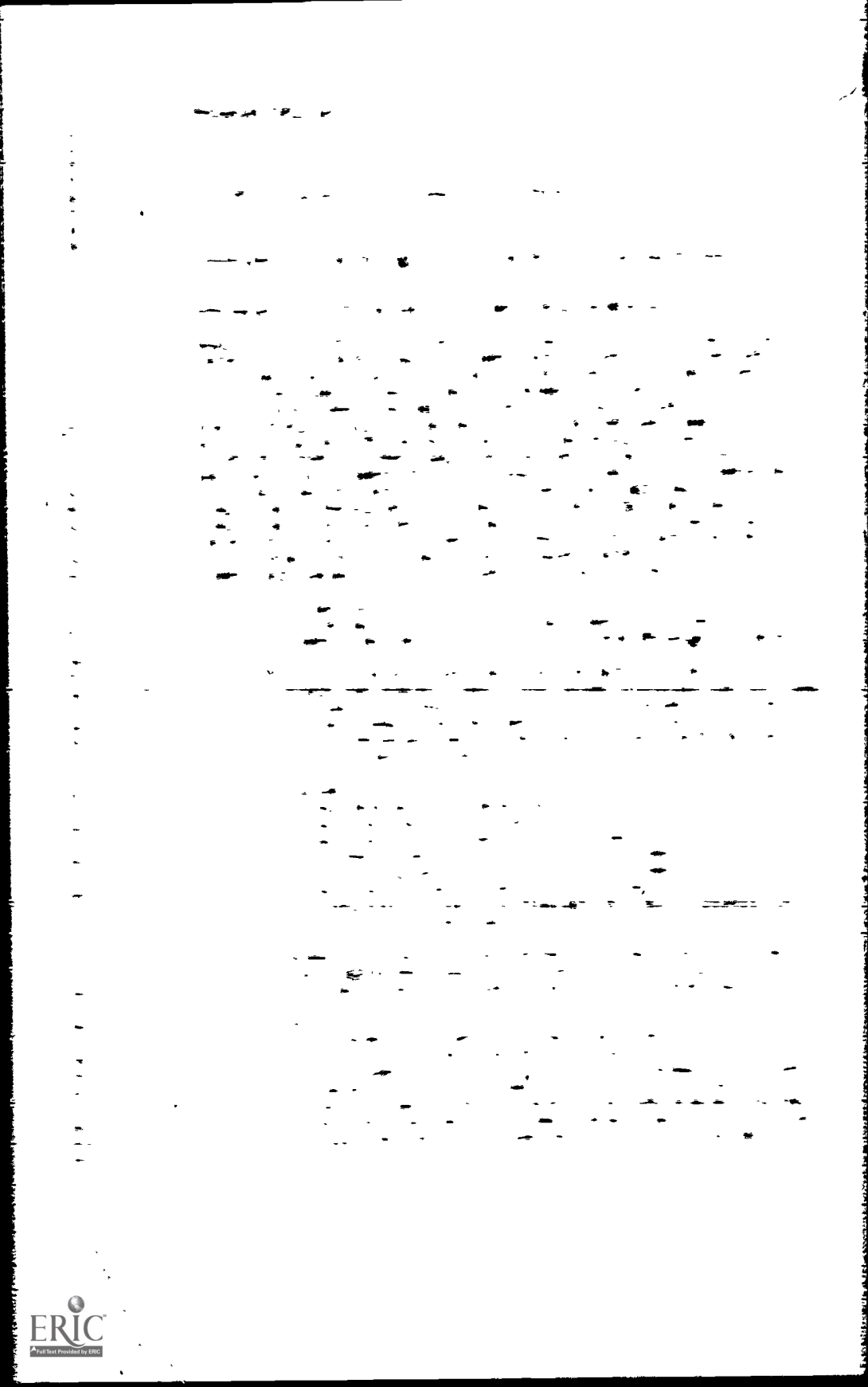
The programs were written for the IBM [illegible] [illegible]

[illegible]

[illegible]







(see pp. 89-90)

Fa

TESTING SIGNIFICANCE OF DIFFERENCE BY APPLICATION OF THE t-TEST FOR RELATED VARIABLES

The computational formula for computing the t-statistic is:

$$t = \frac{M_D}{\sigma_{M_D}}$$

where: M_D = Mean Difference Between Test and Retest Scores

σ_{M_D} = Standard Error of Mean Difference Between Test and Retest Scores.

Computational Procedures

(A sample t-Test run is illustrated in Figure F1. References to the appropriate computations presented in the figure should be made as each successive stage is presented in the discussion.)

1. Compute for each examinee the difference between his Test and Retest scores. This is done in the column labeled D. It makes no difference which score is subtracted from which (i.e., Test from Retest or Retest from Test) as long as the procedure is carried out in the same way for all examinees.

   Comment: The result of Step 1 is the distribution of direct score differences from which all further computations will derive.

2 a. Compute the algebraic Mean of the score difference (M_D). First, sum all positive D values and sum all negative D values. Then subtract the sum of the negative from the sum of the positive D values to obtain ΣD.

   Comment: In the sample run, Retest-Test differences

FIGURE P1

t-TEST WORKSHEET

Item Set 1 - Renneg
11/73-12/73

EXAMINEE ID	ITEM SET 1 T1	R1	D	D^2
01	28	34	06	36
02	30	42	12	144
03	13	19	06	36
04	26	43	17	289
05	23	36	13	169
06	29	37	08	64
07	33	33	0	0
08	23	31	08	64
09	28	32	04	16
10	29	34	05	25
11	25	34	09	81
12	13	29	16	256
13	18	35	17	289
14	22	35	13	169
15	24	36	12	144
16	16	36	20	400
17	01	41	10	100
18	20	36	16	256
19	19	36	17	289
20	21	41	20	400
21	29	37	08	64
22	08	31	23	529
23	27	35	08	64
24	29	36	07	49
25	28	38	10	100
26	23	39	15	256
27	27	29	02	04
28	29	29	0	0
29	18	38	20	400
30	22	35	13	169
31	26	31	05	25
	737	1077		

$\Sigma D(+) = +340$ $\quad \Sigma D^2 = 4856$

$\Sigma D(-) = 0$

$\Sigma D = 340$

$M_D = \frac{340}{31} = 10.97$

$\sigma_D = \sqrt{\frac{4856}{31} - (10.97)^2}$

$\sigma_D = 6.03$

$M_D = \frac{6.03}{\sqrt{30}} = 1.10$

$t = \frac{10.97}{1.10} = 9.97$

$t_{.05} = 1.697$

$t_{.01} = 2.457$

DECISION: $p < .01$

were computed, all of which were positive.

b. Divide the ΣD by N (number of examinees or pairs of raw scores) to compute the M_D:

$$M_D = \frac{\Sigma D}{N}$$

3 a. Compute for each examinee the square of the Test/Retest score difference. This is done by squaring each of the values in the D column and recording them in the column labeled D^2.

b. Compute the ΣD^2 by summing all D^2 values.

4. Compute the standard deviation of the distribution of differences (σ_D). The computational formula is:

$$\sigma_D = \sqrt{\frac{\Sigma D^2}{N} - (M_D)^2}$$

5. Compute the standard error of the mean differences (σ_{M_D}) using the formula:

$$\sigma_{M_D} = \frac{\sigma_D}{\sqrt{N-1}}$$

6. Compute the t-statistic from the t-ratio as follows:

$$t = \frac{M_D}{\sigma_{M_D}}$$

Interpreting the t-Statistic

a. Determine the value of t required for significance at the 5% and/or 1% level and beyond for N-1 degrees of freedom*. These values are provided in Figure F2.

Comment: In the sample run the t values for 30 degrees of freedom (31-1) were used.

b. If the computed t-value is equal to or greater than the value required for significance at the 5% (or 1%) level there is a statistical basis for inferring that the Retest score gain was significant and therefore, that the trainees are significantly more competent with certain subject matter in the post-instruction period than in the pre-instruction period. Further, if proper testing controls are employed as defined in the Manual, such significant score increases (and therefore increases in the levels of competence) can be related to the effects of the training experience.

Conversely, if the derived t-value is less than the value required for significance at a certain level (i.e., either the 5% or 1%), then it can be concluded that there is no evidence for significant increases in levels of competence from initial testing to retesting.

---

* Although both the 5% and 1% levels are provided, it is understood that only one or the other level will be used for the application of tests of significance to a specific body of data. That level should be selected, according to the rules for proper statistical testing, on an "a priori" basis by the evaluator.

Sample t-Test Run

For the data in Figure F1, the t-ratio is 10.97/1.10, giving a t-value = 9.97. When applying the t-Test for Related Variables, the number of degrees of freedom (df), to use when entering the table of t-values for various significance levels (Fig. F2), is N-1 where N is the number of examinees for whom both Test and Retest responses have been obtained (in this case, 31-1=30). With 30 df, the t-statistic is significant beyond the 1% level; therefore, it can be concluded that the trainees increased their levels of competence with the subject matter tested by Item Set 1 to a significant degree from the pre- to post-instruction periods. Furthermore, since all the proper controls were employed during the entire phase of instrument construction and administration, there is no evidence for inferring that the competences increases were due to factors other than the direct effects of instruction.

FIGURE F2

TABLE OF VALUES OF t AT THE 5% & 1% LEVELS OF SIGNIFICANCE

Degrees of freedom (df)	5%	1%
1	6.314	31.821
2	2.920	6.965
3	2.353	4.541
4	2.132	3.747
5	2.015	3.365
6	1.943	3.143
7	1.895	2.998
8	1.860	2.896
9	1.833	2.821
10	1.812	2.764
11	1.796	2.718
12	1.782	2.681
13	1.771	2.650
14	1.761	2.624
15	1.753	2.602
16	1.746	2.583
17	1.740	2.567
18	1.734	2.552
19	1.729	2.539
20	1.725	2.528
21	1.721	2.518
22	1.717	2.508
23	1.714	2.500
24	1.711	2.492
25	1.708	2.485
26	1.706	2.479
27	1.703	2.473
28	1.701	2.467
29	1.699	2.462
30	1.697	2.457
40	1.684	2.423
60	1.671	2.390
120	1.658	2.358

The numeric data in this table are adapted from Table 7 of John T. Roscoe: Fundamental Research Statistics for the Behavioral Sciences, published by Holt, Rinehart and Winston Inc., New York City, 1969, p. 293.

AN ALTERNATIVE TO THE STANDARD t-TEST*

Two types of correct answers contribute to an uncorrected (for guessing; see p. 90) achievement test score: answers guessed correctly and correct answers based upon true competence with the subject matter under assessment. The standard t-Test does not take these two components of a total score into account. Therefore, there might be some question as to whether a significant Test-to-Retest score increase reflects a true increase in levels of competence or an increase in the number of items guessed correctly.

The most appropriate significance test to employ is one which attempts to partial out the contribution of chance factors (guessing correct) to total score.

The statistical test introduced here was designed specifically for application to mean scores derived from the administration of an objective assessment instrument under a Test/Retest design. In contrast to the standard test, the test variation takes into account and attempts to partial out the contribution of chance in order to obtain a more valid evaluation of the changes in levels of competence.

NOTE: The test variation is offered as an alternative to the standard t-Test (rather than recommended outright) because it is a new procedure, its validity yet to be established through repeated application. Therefore, the new test cannot, at this time, be considered as a replacement for the standard t-Test. Nevertheless, we feel that it is a valuable new tech-

---

* The new procedure is referred to as the "z-variation" of the standard t-Test. Although it can be applied to the same test data and is appropriate for small sample testing (i.e., where $N \leq 30$) the z-variation is not a t-Test. The distribution of the z-statistic is normal, unlike the t-statistic which has a Student's distribution.

nique, and one that is probably more appropriate than the standard test in this situation. We would hope that users of this Manual who are familiar with statistical inference will employ this procedure and assess its validity. We would appreciate hearing from those who do apply our procedure to their own testing results. The results of applications and outside comments on the appropriateness and validity of the procedure for its intended purpose would be valuable.

Description of the Procedure

The steps involved in employing the alternative test will differ according to whether the number of response choices is constant or variable across test items. Both situations will be considered.

Situation I (the number of item response alternatives is constant).

a) Definition of Variables:

X1 = number of items Known (i.e., based on subject competence) and scored right on Pre-Test

X2 = number of items known and scored right on Post-Test

Y1 = number of items scored right on Pre-Test (i.e., the sum of items guessed correct and correct items based on competence)

Y2 = number of items scored right on the Post-Test

T = total number of items (i.e., No. examinees X No. items per testing)

N = number of alternatives per item (N in this case is a constant)

The test procedure will determine if an observed score difference (i.e., Y2-Y1) is statistically significant.

b) The Computational Formula

The formula for computing the z-statistic from the raw data is provided below. (The derivation of the formula is provided on p. 177-179.)

$$z = \frac{\sqrt{N}\ (Y2 - Y1)}{\sqrt{2\left(T - \frac{Y1+Y2}{2}\right)}} \qquad (1)$$

c) Computational Procedures

A simple worksheet for computing the z-statistic can be constructed similar to the t-Test worksheet illustrated in Figure F1.

The only raw data required for the computations are the distribution of raw scores by trainees for both Test and Retest (i.e., the data in the columns labeled "Item Set 1 - T1/R1" in Figure F1). All input values for the z-statistic formula are derived from this set of data.

Sample Computations:

Y1 = 737 (Sum of scores in Col. T1)
Y2 = 1077 (Sum of scores in Col. R1)
N = 4 (Number of response alternatives per item)
T = 1581 (Total number of possible responses per Item Set per testing -- 31 examinees x 51 items in Set 1)

Substituting these values in Formula (1), we get

$$z = \frac{\sqrt{4}\ (1077-737)}{\sqrt{2\ (1581 - \frac{1077+737}{2})}}$$

$$= \frac{2\ (340)}{\sqrt{2\ (1581-907)}} = \frac{680}{\sqrt{1348}}$$

$$= \frac{680}{36.71} = \underline{\underline{18.52}}$$

Interpreting the z-Statistic

a) If the derived z-value is greater than or equal to 1.96 but less than 2.58 (i.e., $2.58 > z \geq 1.96$), the difference is significant at the 5% level and evidence exists for inferring a significant increase in the overall trainee group's level of subject matter competence.

b) If the derived z value is greater than or equal to 2.58 (i.e., $z \geq 2.58$), the difference is significant at the 1% level and evidence exists for inferring a significant Pre- to Post- instruction increase in overall levels of subject matter competence.

c) If significance at or beyond other levels are required (e.g., 0.1% level), the critical values can be found in the "Table of Cumulative Normal Probabilities" in any standard textbook of statistical inference.

Situation II (the number of response alternatives is variable across items)

The testing procedure will be the same as for Situation I with the exception of one additional step. Whereas in the first case N is given, here it will be a derived value -- i.e., the average number (harmonic mean) of response choices per item.

In addition to the variables provided above (see "a)" under Situation I), the following variables will be defined:*

a = number of items with n1 choices
b = number of items with n2 choices
c = number of items with n3 choices
t = a+b+c = total number of items in test
x = number of items correct on basis of subject competence
y = number of items guessed correctly
w = number of items guessed wrong

* For illustration purposes the number of response choices range from N1 to N3. More variables (e.g., d = n4, e = n5, etc.) can be added to the harmonic mean formula to accommodate a wider range in the number of response choices provided

a) Based on the assumption that the probability of knowing (i.e., in this case having the subject competence to be able to answer an item correctly) the correct answer to an item is independent of the number of item choices,

$$g = \frac{t - x}{t} \left(\frac{a}{n1} + \frac{b}{n2} + \frac{2}{n3} \right)$$

(the value of g in the case of the same number of choices across items is $\frac{t - x}{n}$)

therefore,

$$\frac{t - x}{n} = \frac{t - x}{t} \left(\frac{a}{n1} + \frac{b}{n2} + \frac{c}{n3} \right)$$

and,

$$\underset{\equiv}{n} = \frac{a+b+c}{\frac{a}{n1} + \frac{b}{n2} + \frac{c}{n3}} \qquad (2)$$

b) Once n is computed using formula (2), the value of n can be substituted for N in formula (1) and the z-statistic computed and interpreted.

Derivation of the Computational Formula for the z-Statistic*

The general definitional formula for the z-statistic is:

$$z = \frac{Y2-Y1}{SE(Y2-Y1)}$$

* The variables employed in this section were previously defined (see p. 174).

where:

$Y2-Y1$ = magnitude of the difference between aggregate Test and Retest scores.

$SE(Y2-Y1)$ = Standard error of difference between Test and Retest scores.

In order to derive the computational formula, the standard error (SE) of Y2-Y1 (i.e., for items guessed correctly) must be defined in raw score form (the "v" refers to the "variance of").

$$v(Y2-Y1) = v(Y2) + v(Y1)$$

but, $v(Y2) = v(X2 + \text{No. items guessed correctly on Post-Tes}$

$$= \frac{1}{N}\left(1-\frac{1}{N}\right)(T-X2)$$

Similarly, $v(Y1) = \frac{1}{N}\left(1-\frac{1}{N}\right)(T-X1)$

Therefore, $v(Y2-Y1) = 2\cdot\frac{1}{N}\left(1-\frac{1}{N}\right)\left(T - \frac{X1+X2}{2}\right)$

Since the true values of X1 and X2 cannot be known, each will be expressed in terms of Y1, Y2 and T as follows:

$$\left(Y1+Y2 - (X1+X2)\right) = \frac{2T - (X1+X2)}{N}$$

or, $$\frac{X1+X2}{2} = \frac{\frac{N}{2}(Y1+Y2) - T}{N-1}$$

therefore, $$v(Y2-Y1) = 2\cdot\frac{1}{N}\left(1-\frac{1}{N}\right)\left(T - \frac{\frac{N}{2}(Y1+Y2) - T}{N-1}\right)$$

$$= \frac{2}{N}\left(T - \frac{Y1+Y2}{2}\right)$$

the $SE_{(Y2-Y1)} = \sqrt{v} = \sqrt{\frac{2}{N}\left(T - \frac{Y1+Y2}{2}\right)}$

$\therefore$ the derived computational formula is:

$$z = \frac{Y2-Y1}{\sqrt{\frac{2}{N}\left(T - \frac{Y1+Y2}{2}\right)}} = \frac{\sqrt{N}\,(Y2-Y1)}{\sqrt{2\left(T - \frac{Y1+Y2}{2}\right)}}$$

(The sampling distribution of the z-statistic is normal with a mean of 0 and a standard derivation of 1.)

(see pp. 90-91)

Fb

TESTING SIGNIFICANCE OF DIFFERENCE BY APPLICATION OF THE CHI SQUARE TEST OF INDEPENDENCE

In order to illustrate the computational formula for deriving the Chi Square (χ^2) Statistic, the observed frequencies in the 2 X 2 contingency table (see text, page 91) can be symbolized as follows:

	ITEMS		
Testing	Correct	Incorrect	
Pretest	A (60)	B (53)	A + B
Posttest	C (78)	D (35)	C + D
	A + C	B + D	N (=A+B+C+D)

Using the above scheme, the computation of the χ^2 Statistic is carried out according to the formula

$$\frac{N\,(AD-BC)^2}{(A+B)\;(C+D)\;(A+C)\;(B+D)}$$

Sample Computation

Employing the Composite Score Data for examinee #1 (as shown in Figure 5, p. 52, and reproduced below) the chi square statistic is computed as follows:

	C	I	
Pre	60	53	113
Post	78	35	113
	138	88	226

$$\chi^2 = \frac{226((60 \times 35) - (53 \times 78))^2}{(113)\ (113)\ (138)\ (88)}$$

$$= \frac{226\ (2100 - 4134)^2}{(12769)\quad (12144)}$$

$$= \frac{226\ (4137156)}{155066736} = \frac{934997}{155067}$$

$$= 6.03$$

Interpreting the Computed Chi Square Value

In the sample 2 x 2 table, the observed cell frequencies are classified two ways: by "correct vs incorrect" items, and by time of testing (Pre-Test vs Post-Test).

In terms of testing for significant differences, the essential question is whether or not the two ways of categorizing the observed cell frequencies are independent of each other.

If the two ways of classifying are independent, then the distribution of correct and incorrect item responses does not depend upon the time of testing. This is the same as stating that the Pre-Test scores (i.e., the distribution of correct item responses) do not differ significantly from the Post-Test scores.

If the categorizations are not independent (i.e., if they are correlated) then there is evidence for the fact that the Post-Test differs significantly from the Pre-Test in the distribution of correct and incorrect item responses.

As stated earlier, the 5 and 1% levels of significance are used for significance testing in the analysis section. The values of χ^2 required for significance (for any 2 X 2 table) are:

3.84 for significance at (or beyond) the 5% level
6.64 for significance at (or beyond) the 1% level

When the Null Hypothesis (that the Pre-Test and Post-Test distributions of correct and incorrect responses are independent of each other) is rejected at (or beyond) the 5 or 1% level of significance, the alternative hypothesis that the Pre-Test and Post-Test distributions are correlated, and therefore significantly different, is supported. Furthermore if the Null Hypothesis is rejected, and at the same time, the Post-Test score is greater than its Pre-Test counterpart, then it can be concluded that the Post-Test score gain is statistically significant. If the assumption that appropriate controls (discussed in the text) were employed during test construction and administration is accepted, then there is evidence for inferring that the significant score gains reflect increases in subject competence brought about as the result of training instruction.

(For the sample data, the computed χ^2 value of 6.03 was significant beyond the 5% level. Furthermore, the fact that the Post-Test score was higher than the Pre-Test gave support to the inference cited above for the positive impact of instruction on increasing the levels of general or composite subject matter competence for trainee #1.)

When, on the other hand, the computed χ^2 value falls short of significance at the 5% or 1% level (whichever had been preselected), there is no unequivocal evidence for significant statistical differences between Test and Retest score distributions. Inferences of positive effects of instruction on subject matter competence for the specific examinee under assessment would not be supported.

(see pp. 91-98)

Fc

QUANTITATIVE PROCEDURES FOR CONSTRUCTING ITEM PATTERN ANALYSIS TABLES

To illustrate the construction, a set of data from the Rennes field testing will be used, consisting of the responses of the 31 trainees to the 51 items of Set 1.

Figure F3 illustrates a worksheet used for recording scores from individual answer sheets, so that all the information is on one form (where the number of trainees or items is too large for inclusion on a single sheet, trainees or items can be broken into subgroups and several sheets used, with totals added-up on a cover sheet).

The worksheet is set up so that the number of correct and incorrect responses on the Test and Retest may be totalled both for each trainee and for each item. In addition, the direction of movement of each item from Test to Retest is shown.

Across the top of the sheet, the number of each item is entered. Down the left hand column, the ID number of each trainee. To fill in the worksheet, the response data from each trainee's answer sheet is transferred to the appropriate column. That is, taking the Pre-Test answer sheet for trainee #01 and moving across the worksheet, enter a C (for each correct response) or an I (for an incorrect response) under the appropriate item number. Then, enter the total number C and the total number I in the appropriate boxes (to the right of the heavy black line). The same procedure is then followed with the Post-Test answer sheet for trainee #01. After the responses of both testings have been entered, the third horizontal column is used to indicate the response pattern for each item from Test to Retest, numbered as follows:

Test (T)	C	C	I	I
Retest (R)	C	I	C	I
→	1	2	3	4

FIGU

SAMPLE WORKSHEET FOR

FP/MCH Program - 11/73-12/73

TRAINEE ID		ITEM NUMBER	1	2	3	4	5	6
01	T		C	C	I	I	I	I
	R		C	I	I	I	C	I
	→		1	2	3	3	4	3
02	T		C	I	I	C	C	I
	R		C	C	C	C	C	I
	→		1	4	4	1	1	3
03	T		C	C	I	I	I	C
	R		C	C	I	C	C	I
	→		1	1	3	4	4	2
28	T		C	I	I	I	C	I
	R		I	C	C	C	C	I
	→		1	1	3	4	4	2
29	T		C	C	C	I	I	C
	R		C	C	C	C	I	C
	→		1	1	1	4	3	1
30	T		I	I	I	C	I	I
	R		C	C	I	C	I	I
	→		4	4	3	1	3	3
31	T		C	C	I	I	I	C
	R		C	I	I	C	C	C
	→		1	2	3	4	4	1
TOTAL CORRECT:	T		8	10	7	11	22	14
	R		11	11	16	27	27	26
(1)	C→C		3	6	6	10	21	13
(2)	C→I		5	4	1	1	1	1
(3)	I→I		15	16	14	3	3	4
(4)	I→C		8	5	10	17	6	13

F3

ITEM PATTERN ANALYSIS

48	49	50	51	C	I	(1) C→C	(2) C→I	(3) I→I	(4) I→C
C	C	C	C	28	23				
C	I	C	C	34	17	24	4	13	10
1	2	1	1						
C	I	I	C	30	21				
C	C	C	C	42	9	26	4	5	16
1	4	4	1						
I	C	C	I	13	38				
I	C	C	I	19	32	6	7	25	13
3	1	1	3						
I	I	C	C	29	22				
I	I	I	C	39	12	25	4	8	14
3	3	2	1						
I	C	C	C	18	33				
C	C	C	C	38	13	16	2	11	22
4	1	1	1						
I	C	I	C	22	29				
C	C	C	C	35	16	16	6	10	19
4	1	4	1						
I	C	C	C	26	25				
C	C	C	C	31	20	23	3	17	8
4	1	1	1						
11	4	5	13	737	844	615	122	372	472
12	14	0	19	1087	494				
4	4	0	10			615			
7	0	5	3				122		
12	17	26	9					372	
8	10	0	9						472

The last four vertical columns are filled in according to the frequency of occurrence of each of the four Pre-Test to Post-Test item patterns for each trainee.

All of this information is entered for each trainee, and the vertical columns are then totalled. Across the bottom of the page are the response patterns for each item: the number of times they were answered correctly on the Test and Retest, the number of times they were correct on both, incorrect on both, and the number of times they went from correct to incorrect or from incorrect to correct.

Down the right-hand columns are the same patterns as they apply to each trainee: how many correct responses, how many times they were correct both times, incorrect both times, how many times their responses were incorrect on the Pre-Test and correct on the Post-Test and how many times they were correct on the Pre-Test and incorrect on the Post-Test.

Finally, both horizontal and vertical columns are totalled in the lower right-hand corner (below and to the right of the heavy lines). This serves as a check against errors in calculations or recording--the totals should be the same for both the horizontal and vertical columns.

Note: Using the worksheet, a table can be constructed displaying the totals only. Figure F4 is an example of a table displaying the summary scores and pattern frequencies for each item (on the worksheet, the table values correspond to the totals for each horizontal column). Figure F5 is an example of a table summarizing the scores and response pattern frequencies for each trainee (the table values represent the totals for each vertical column on the worksheet). The table in Figure F5 is similar to the type of table used in the discussion of the item response pattern analysis on pp. 91-98. When the table has been completed and the appropriate statistics (i.e., percentages) calculated, the analysis of the data will be conducted as described on pp. 94-98.

FIGURE F4

ITEM RESPONSE PATTERNS
BY INDIVIDUAL ITEM

ITEM	(1) TOTAL CORRECT PRE	(1) TOTAL CORRECT POST	PRE---- POST C-- C	PRE---- POST C-- I	(2) TOTAL INCORRECT PRE	(2) TOTAL INCORRECT POST	PRE--- POST I-- I	PRE--- POST I-- C	TOTAL RESPONSES (1+2)
1	8	11	3	5	23	20	15	8	31
2	10	11	6	4	21	20	16	5	31
3	7	16	6	1	24	15	14	10	31
4	11	27	10	1	20	4	3	17	31
5	22	27	21	1	9	4	3	6	31
6	14	26	13	1	17	5	4	13	31
7	26	30	25	1	5	1	0	5	31
8	23	25	18	5	8	6	1	7	31
9	14	9	4	10	17	22	12	5	31
10	21	22	14	7	10	9	2	8	31
11	21	27	19	2	10	4	2	8	31
12	15	26	13	2	16	5	3	13	31
13	24	29	23	1	7	2	1	6	31
14	19	27	18	1	12	4	3	9	31
15	22	21	18	4	9	10	6	3	31
16	20	29	18	2	11	2	0	11	31
17	26	22	20	6	5	9	3	2	31
18	25	30	24	1	6	1	0	6	31
19	4	3	1	3	27	28	25	2	31
20	3	13	0	3	28	18	15	13	31
21	7	1	1	6	24	30	24	0	31
22	6	21	6	0	25	10	10	15	31
23	20	21	16	4	11	10	6	5	31
24	27	29	27	0	4	2	2	2	31
25	17	24	15	2	14	7	5	9	31
26	15	27	12	3	16	4	1	15	31
27	7	27	6	1	24	4	3	21	31
28	23	28	23	0	8	3	3	5	31
29	25	28	23	2	6	3	1	5	31
30	5	18	4	1	26	13	12	14	31
31	17	24	13	4	14	7	3	11	31
32	1	17	1	0	30	14	14	16	31
33	17	31	17	0	14	0	0	14	31
34	14	21	12	2	17	10	8	9	31
35	23	29	23	0	8	2	2	6	31
36	10	21	8	2	21	10	8	13	31
37	24	30	23	1	7	1	0	7	31
38	5	16	3	2	26	16	13	13	31
39	10	9	4	6	21	22	16	6	31
40	10	23	8	2	21	8	6	15	31
41	20	30	19	1	11	1	0	11	31
42	13	25	12	1	18	6	5	13	31
43	11	10	6	6	20	21	16	4	31
44	10	29	10	0	21	2	2	19	31
45	8	25	8	0	23	6	6	17	31
46	6	21	6	0	26	10	10	15	31
47	18	26	17	1	13	5	4	9	31
48	11	12	4	7	20	19	12	8	31
49	4	14	4	0	27	17	17	10	31
50	5	0	0	5	26	31	26	0	31
51	13	19	10	3	18	12	9	9	31
TOTAL	737	1087	615	122	844	494	372	472	1581

46.6% - - - - - - - - 53.4% - - - - - - - - - - - - →100%

68.8% - - - - - - - - - - - - 31.2% - - - - - - - - →100%

100% ← - - - - 83.4% - - 16.6%   100% ← - - - - - 44.1% - - - 55.9%

FIGURE P5

ITEM RESPONSE PATTERNS
BY INDIVIDUAL TRAINEE

ID	(1) TOTAL CORRECT PRE	POST	PRE-----→ POST C-->C	C-->I	(2) TOTAL INCORRECT PRE	POST	PRE---→ POST I-->I	I-->C	TOTAL ITEMS (1+2)
1	28	34	24	4	23	17	13	10	51
2	30	42	26	4	21	9	5	16	51
3	13	19	6	7	38	32	25	13	51
4	26	43	25	1	25	8	7	18	51
5	23	36	19	4	28	15	11	17	51
6	29	37	28	1	22	14	13	9	51
7	33	33	28	5	18	18	13	5	51
8	23	31	16	7	28	20	13	15	51
9	28	32	22	6	23	19	13	10	51
10	29	34	25	4	22	17	13	9	51
11	25	34	21	4	26	17	13	13	51
12	13	29	12	1	38	22	21	17	51
13	18	35	15	3	33	16	13	20	51
14	22	35	17	5	29	16	11	18	51
15	24	36	20	4	27	15	11	16	51
16	16	36	11	5	35	15	10	25	51
17	31	41	30	1	20	10	9	11	51
18	20	36	17	3	31	15	12	19	51
19	19	36	16	3	32	15	12	20	51
20	21	41	20	1	30	10	9	21	51
21	29	37	25	4	22	14	10	12	51
22	8	31	5	3	43	20	17	26	51
23	27	35	22	5	24	16	11	13	51
24	29	36	23	6	22	15	9	13	51
25	28	38	26	2	23	13	11	12	51
26	23	38	21	2	28	13	11	17	51
27	27	29	15	12	24	22	10	14	51
28	29	39	25	4	22	12	8	14	51
29	18	38	16	2	33	13	11	22	51
30	22	35	16	6	29	16	10	19	51
31	26	31	23	3	25	20	17	8	51
TOTAL:	737	1087	615	122	844	494	372	472	1581
	46.6%				53.4%				→100%
	68.8%				31.2%				→100%
	100% ←		83.4%	16.6%	100% ←		44.1%	55.9%	

APPENDIX G

MATHEMATICAL EQUATION AND PARAMETER VALUES FOR GENERATING THE WEIGHTED ACHIEVEMENT/COMPETENCE SCORE CURVES

The general equation that describes mathematically the family of curves, one series of which is presented in Figure 15, is

$$\frac{(R - T)\ N^2}{N^2 - T^2} = I$$

where: T = Test Score
R = Retest Score
N = Total Number of Items in Test Instrument
I = Weighted Achievement/Competence Score

The values of the T, R and I parameters* used to generate the specific curves presented in the figure are as follows:

I=0		I=10		I=20		I=30		I=40	
T	R	T	R	T	R	T	R	T	R
0	0	0	10	0	20	0	30	0	40
10	10	10	19.9	10	29.8	10	39.7	10	49.6
20	20	20	29.6	20	39.2	20	48.8	20	58.4
30	30	30	39.1	30	48.2	30	57.3	30	66.4
40	40	40	48.4	40	56.8	40	65.2	40	73.6
50	50	50	57.5	50	65.0	50	72.5	50	80.0
60	60	60	66.4	60	72.8	60	79.2	60	85.6
70	70	70	75.1	70	80.2	70	85.3	70	90.4
80	80	80	83.6	80	87.2	80	90.8	80	94.4
90	90	90	91.9	90	93.8	90	95.7	90	97.6

I=50		I=60		I=70		I=80		I=90	
T	R	T	R	T	R	T	R	T	R
0	50	0	60	0	70	0	80	0	90
10	59.5	10	69.4	10	79.3	10	89.2	10	99.1
20	68.0	20	77.6	20	87.2	20	96.8		
30	75.5	30	84.6	30	93.7				
40	82.0	40	90.4	40	98.8				
50	87.5	50	95.0						
60	92.0	60	98.4						
70	95.5	70	99.6						
80	98.0								
90	99.5								

*N=100

REFERENCES

Notes to the Text

1 Sherman N. Tinkelman, Planning the objective test. In Robert L. Thorndike (Ed.), Educational measurement. (2nd ed.) Washington, D. C.: American Council on Education, 1971. p. 51.

2 Robert L. Ebel, Measuring educational achievement. Englewood Cliffs, N. J.: Prentice-Hall, 1965. Pp. 68-72.

3 Norman E. Gronlund, Measurement and evaluation in teaching. (2nd ed.) New York: Macmillan, 1971. p. 130.

4 Gronlund, Measurement and evaluation in teaching, p. 224.

5 Ebel, Measuring educational achievement, p. 60.

6 Lee J. Cronbach, Educational psychology. (2nd ed.) New York: Harcourt, Brace & World, 1963. p. 554.

7 Tinkelman, in Thorndike, Educational measurement, p. 50.

8 Guidelines have been adapted from discussions of item construction in: Robert L. Thorndike, & Elizabeth Hagen, Measurement and evaluation in psychology and education. (2nd ed.) New York: John Wiley, 1961. Pp. 61-85; Richard H. Lindeman, Educational measurement. Glenview, Ill.: Scott Foresman, 1967. Pp. 84-85; and Alexander G. Wesman, Writing the test item. In Robert L. Thorndike (Ed.), Educational measurement. (2nd ed.) Washington, D. C.: American Council on Education, 1971. Pp. 81-129.

9 Ebel, Measuring educational achievement, p. 468.

10 Leslie Kish, Some statistical problems in research design. American Sociological Review, 1959, 24, 328-338.

11 Hanan C. Selvin, A critique of significance in survey research. American Sociological Review, 1957, 22, 519-527.

12 Ebel, Measuring educational achievement, p. 81.

13 Paul B. Diederich, Pitfalls in the measurement of gains in achievement. In William C. Morse & G. Max Wingo (Eds.), Readings in educational psychology. Fairlawn, N. J.: Scott Foresman, 1962. p. 363.

14 Lindeman, Educational measurement, p. 89.

Notes to the Appendices

1 Benjamin Bloom (Ed.), Taxonomy of educational objectives: the classification of educational goals. Handbook 1. The cognitive domain. New York: David McKay, 1956. Pp. 28-29.

2 Lee J. Cronbach, Essentials of psychological testing. (3rd ed.) New York: Harper & Row, 1970. p. 24.

3 Donald T. Campbell, & Julian C. Stanley, Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1963. p. 7.

4 Campbell, & Stanley, Experimental and quasi-experimental designs for research, p. 7.

5 Campbell, & Stanley, Experimental and quasi-experimental designs for research, Pp. 7-12.

6 M. J. Apter, D. Boorer, & S. Murgatroyd, A comparison of the effects of multiple-choice and constructed response pretests in programmed instruction. Programmed Learning, 1971, 8(4), 251-256.

7 Campbell, & Stanley, Experimental and quasi-experimental designs for research, p. 48.

8 Alexander W. Astin, & Robert J. Panos, The evaluation of educational programs. In Robert L. Thorndike (Ed.), Educational measurement. (2nd ed.) Washington, D. C.: American Council on Education, 1971. p. 744.

9

Guidelines adapted from: Gronlund, Measurement and evaluation in teaching, Pp. 183-192 & 196-216; Wesman, in Thorndike, Educational measurement, Pp. 81-128; Thorndike, & Hagen, Measurement and evaluation in psychology and education (2nd ed.), Pp. 60-95; and Robert L. Thorndike, & Elizabeth Hagen, Measurement and evaluation in psychology and education. (3rd ed.) New York: John Wiley, 1969. Pp. 104-116.

BIBLIOGRAPHY

Cited Sources

Apter, M. J., Boorer, D., & Murgatroyd, S. A comparison of the effects of multiple-choice and constructed response pre-tests in programmed instruction. Programmed Learning, 1971, 8(4), 251-256.

Astin, Alexander W., & Panos, Robert J. The evaluation of educational programs. In Robert L. Thorndike (Ed.), Educational measurement (2nd edition). Washington, D. C.: American Council on Education, 1971. Pp. 733-751.

Bloom, Benjamin (Ed.) Taxonomy of educational objectives: the classification of educational goals. Handbook 1. Cognitive domain. New York: David McKay, 1956.

Campbell, Donald T., & Stanley, Julian C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1963.

Cronbach, Lee J. Educational psychology (2nd edition). New York: Harcourt, Brace & World, 1963.

Cronbach, Lee J. Essentials of psychological testing (3rd edition). New York: Harper & Row, 1970.

Diederich, Paul B. Pitfalls in the measurement of gains in achievement. In William C. Morse & G. Max Wingo (Eds.), Readings in educational psychology. Fairlawn, N. J.: Scott Foresman, 1962. Pp. 359-364.

Ebel, Robert L. Measuring educational achievement. Englewood Cliffs, N. J.: Prentice-Hall, 1965.

Gronlund, Norman E. Measurement and evaluation in teaching (2nd edition). New York: Macmillan, 1971.

Kish, Leslie. Some statistical problems in research design. American Sociological Review, 1959, 24, 328-338.

Lindeman, Richard H. Educational measurement. Glenview, Ill.: Scott Foresman, 1967.

Roscoe, John T. Fundamental research statistics for the behavioral sciences. New York: Holt, Rinehart & Winston, 1969.

Selvin, Hanan C. A critique of significance in survey research. American Sociological Review, 1957, 22, 519-527.

Thorndike, Robert L., & Hagen, Elizabeth. Measurement and evaluation in psychology and education (2nd edition). New York: John Wiley, 1961.

Thorndike, Robert L., & Hagen, Elizabeth. Measurement and evaluation in psychology and education (3rd edition). New York: John Wiley, 1969.

Tinkelman, Sherman N. Planning the objective test. In Robert L. Thorndike (Ed.), Educational measurement (2nd edition). Washington, D. C.: American Council on Education, 1971. Pp. 46-80.

Wesman, Alexander G. Writing the test item. In Robert L. Thorndike (Ed.), Educational measurement (2nd edition). Washington, D. C.: American Council on Education, 1971. Pp. 81-129.

Supplementary Readings

Anastasi, Anne. Psychological testing (3rd edition). New York: Macmillan, 1968.

Dyer, Henry S. On the assessment of academic achievement. In William G. Morse & G. Max Wingo (Eds.), Readings in educational psychology. Fairlawn, N. J.: Scott Foresman, 1962. Pp. 353-359.

Gerberich, J. R. Specimen objective test items: a guide to achievement test construction. New York: Longmans & Green, 1956.

Guilford, J. P. Fundamental statistics in psychology and education (4th edition). New York: McGraw-Hill, 1965.

Karmel, Louis J. Measurement and evaluation in the schools. New York: Macmillan, 1970.

Pierotti, Daniel J.-A., Lecorps, Philippe, & Revson, Joanne E. Une analyse par l'equipe de formation. Rennes, France: Ecole Nationale de la Sante Publique, Section de Sante et Protection de la Famille, February 1974.

Manuals for Evaluation of Family Planning & Population Programs:

#1 A Framework for the Selection of Family Planning Program Evaluation Topics, by Jack Reynolds*

#2 A Framework for the Design of Family Planning Program Evaluation Systems, by Jack Reynolds*

#3 A Method for Estimating Future Caseload of Family Planning Programs, by Jack Reynolds and Rukmani Ramaprasad*

#4 Operational Evaluation of Family Planning Programs Through Process Analysis, by Jack Reynolds*

#5 The Fertility Pattern Method: Estimation of Fertility Change by Retrospective Quasi-Cohort Analysis of Group-Specific Fertility Patterns, by Samuel M. Wishik and Donald W. Helbig

#6 A Checklist for Evaluative Overviews of Family Planning Program Activities, by Jack Reynolds*

#7 Couple-Years of Protection A Measure of Family Planning Program Output, by Samuel M. Wishik and Kwan-Hwa Chen

#8 Evaluating Training Effectiveness and Trainee Achievement. Methodology for Measurement of Changes in Levels of Cognitive Competence, by Bernard G. Pasquariella and Samuel M. Wishik

*Also available in Spanish.