DOCUMENT RESUME

ED 078 021 TM 002 836

AUTHOR Farr, S. David; Subkoviak, Michael J.
TITLE Program Evaluators Handbook: Measurement.
INSTITUTION New York State Education Dept., Albany. Bureau of Urban and Community Programs Evaluation.
NOTE 34p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Achievement Tests; Classroom Observation Techniques; *Guides; Item Sampling; *Measurement; *Professional Training; *Program Evaluation; Sampling; Standardized Tests; Test Construction; Test Selection

ABSTRACT

This handbook for training program evaluators provides background information and practice activities in the following areas: (1) measurement: purposes, ideals, possibilities; (2) defining measurement domains; (3) person and item sampling; (4) test and item selection--with a selected list of standardized tests with pertinent information on each, suggestions for writing objective test items, and formulae for item analysis; and (5) objective observation. It is recommended that each group in training choose one of the four sample situations described and use it throughout the sessions. The situations are: (1) prekindergarten program for disadvantaged children; (2) introduction of teacher aides (elementary); (3) individualizing instruction through computer-based resource units (high school); and (4) improving interracial attitudes and knowledge. (KM)

TM

PROGRAM EVALUATORS HANDBOOK

Measurement

One of a Review Series
in the
Bureau of Urban and Community
Programs Evaluation

The University of the State of New York
THE STATE EDUCATION DEPARTMENT
Albany, New York 12224

PROGRAM EVALUATORS HANDBOOK

Measurement

One of a Review Series
in the
Bureau of Urban and Community
Programs Evaluation

The University of the State of New York
THE STATE EDUCATION DEPARTMENT
Albany, New York 12224

THE UNIVERSITY OF THE STATE OF NEW YORK

Regents of University (with years when terms expire)

1984	Joseph W. McGovern, A.B., J.D., L.H.D., LL.D., D.C.L., Chancellor	New York
1985	Everett J. Penny, B.C.S., D.C.S., Vice Chancellor	White Plains
1978	Alexander J. Allan, Jr., LL.D., Litt.D.	Troy
1973	Charles W. Millard, Jr., A.B., LL.D., L.H.D.	Buffalo
1987	Carl H. Pforzheimer, Jr., A.B., M.B.A., D.C.S., H.H.D.	Purchase
1975	Edward M. M. Warburg, B.S., L.H.D.	New York
1977	Joseph T. King, LL.B.	Queens
1974	Joseph C. Indelicato, M.D.	Brooklyn
1976	Mrs. Helen B. Power, A.B., Litt.D., L.H.D., LL.D.	Rochester
1979	Francis W. McGinley, B.S., J.D., LL.D.	Glens Falls
1980	Max J. Rubin, LL.B., L.H.D.	New York
1986	Kenneth B. Clark, A.B., M.S., Ph.d., LL.D., L.H.D., D.Sc.	Hastings on Hudson
1982	Stephen K. Bailey, A.B., B.A., M.A., Ph.D, LL.D.	Syracuse
1983	Harold E. Newcomb, B.A.	Owego
1981	Theodore M. Black, A.B., Litt.D.	Sands Point

President of the University and Commissioner of Education
Ewald B. Nyquist

Executive Deputy Commissioner of Education
Gordon M. Ambach

Deputy Commissioner for Elementary, Secondary, and Continuing Education
Thomas D. Sheldon

Associate Commissioner for Elementary, Secondary, and Continuing Education
William L. Bitner

Associate Commissioner for Educational Finance and Management Services
Stanley L. Raub

Director, Center for Planning and Innovation
Norman D. Kurland

Associate Director, Center for Planning and Innovation
Mark B. Scurrah

Associate Commissioner for Research and Evaluation
Lorne H. Woollatt

Director, Division of Evaluation
Alan G. Robertson

Chief, Bureau of Urban and Community Programs Evaluation
Leo D. Doherty

FOREWORD

The increased competition for the tax dollar has caused and will continue to cause more rigorous evaluations in all fields of education, particularly at the Federal level. Increasingly, legislators and their constituent taxpayers are demanding hard data which will indicate whether a costly program is achieving that which it has purported to achieve. Under these conditions evaluation at all levels must satisfy the criteria elements of significance, credibility, and timeliness. Within this framework evaluative techniques must be strengthened.

Appropriate departmental personnel believed that strengthening the evaluative effort of the State might start with categorically aid projects at the elementary and secondary education level.

Appropriate people from within the State were asked to prepare and conduct formal lessons accompanied by simulated experiences and related materials. Thus this document is one in a series of review manuals to be used by appropriate local education. The contents of the series are appropriate for use with large program evaluative problems such as those encountered in ESEA, Urban Education, or the like.

This document on Measurment was prepared by S. David Farr and Michael J. Subkoviak, State University of New York at Buffalo.

CONTENTS

	PAGE
FOREWORD	iii
ORGANIZATION	1
UNITS	
1. Measurement: Purposes, Ideals, Possibilities	2
2. Defining Measurement Domains	5
3. Person and Item Sampling	8
4. Test and Item Selection	12
5. Objective Observation	24
SAMPLE SITUATIONS	
A. Prekindergarten program for disadvantaged children	27
B. Introduction of teacher aides (Elementary)	28
C. Individualizing instruction through computer based resource units (High School)	29
D. Improving interracial attitudes and knowledge	30

TITLE III EVALUATORS TRAINING

MEASUREMENT

Organization

Objectives: Learn, practice, share insights and examples

Number of Units: 5

Time per Unit: 90 minutes

Time Within Units (Approximate):

30 minutes - Lecture
30 minutes - Student work
30 minutes - Reporting and discussion

Formation of Groups for Student Work:

Initial assignments made by instructor - changes permitted

Choice of Sample Situations: Recommended that each group choose a situation and use it throughout sessions

Reporting Student Work:

1) Rotate recorder within group (arbitrary assignments by instructors may be varied)

2) Recorder will keep notes in black ink (pen provided) for reproduction

3) Recorder will also serve as reporter

MEASUREMENT: PURPOSES, IDEALS, POSSIBILITIES

Although many people think of measurement only in terms of the student outcomes which an experimental program attempts to change, a much broader view is desirable. Outcomes are only one of three classes of observations, and student outcomes are a subset of that class. The setting of any study should be carefully described so that others may interpret results in terms of their own situation. This may include measures on the community, the teachers and other school personnel, the physical facilities, the students' initial abilities, and the society in general. In addition, a clear description of the nature of the experimental program is essential. Usually, only through observation or other measurement procedures, can the program be described as it really happened. Proof that specified "treatments" were really administered to the students and a record of resulting changes in classroom behavior are the only adequate description of the program. Viewed broadly, then, the measurement plan for an innovative program should include measures on the setting, the process of the program, and its outcomes. Adequate attention to these three classes of measurements requires a serious committment to the measurement effort.

Two general ideals apply no matter what is being measured or how the measurements are made. These are meaningfulness and precision. To make meaningful interpretations of measurements, the tasks assigned, the method of observing, and the way scores are formed must follow a logical system. Simplicity and directness often are helpful in producing meaningful scores. Precision deals with whether a measurement, when replicated, will produce the same result. Precision of individual measurements is less important when an aggregate, for example a student body, is the object to be measured.

TITLE III EVALUATORS TRAINING

MEASUREMENT UNIT 1

Measurement: Purposes, Ideals, Possibilities

I. Purposes: Accurate description of
   A. Setting
   B. Treatment (Process)
   C. Outcomes

II. Ideals
   A. Meaningfulness
   B. Accuracy, or precision

III. Possibilities
   A. Setting
      1. measurement on community
      2. teachers
      3. other personnel - administrators, aides
      4. physical facilities
      5. children
      6. historical events
   B. Treatment
      1. measurement of specified treatment details
      2. nonspecified general description
      3. other routine observations
      4. use of facilities
   C. Outcomes
      1. student behavior
      2. teacher behavior
      3. auxiliary personnel
      4. parent or community
      5. delayed effects, persistence of observed effects

IV. Summary
   A. Multiple purposes and variables
   B. Meaningfulness and accuracy

References:

R. W. Tyler, R. M. Gagne and M. Scriven Perspectives on Curriculum Education. No. 1, AERA Monograph Series on Curriculum Evaluation. Chicago: Rand McNally, 1967.

UNIT 1 ACTIVITIES

Using the assigned or selected sample situation, plan a comprehensive measurement (information gathering) program for that project. Since descriptions of the sample situation are necessarily sketchy you may assume reasonable additions to and specification of the objectives and procedures.

The primary task is to specify what you wish to observe or assess and when. Do not be concerned about exactly how traits or abilities will be assessed or behaviors observed.

Assemble your decisions in the form of a rough chronology of observation or data collection.

Do not hesitate to set up a more extensive plan than practical considerations will allow. Such a plan can always be pared down later.

Time: 30 minutes

The reporter will present a 5-10 minute summary to the group.

DEFINING MEASUREMENT DOMAINS

Once a decision to measure a certain variable has been reached, the problem of how the measurement is to be made must be faced. It is easy to talk about achievement, anxiety, valuation, or attitudes, but it is another thing to measure them. A promising approach is to define a domain of tasks, observations, and conditions relevant to the specified variable. Since these domains are usually very large, measurements are made by sampling some of the elements. By knowing the extent and boundaries of the domain, however, the meaningfulness of measurements based on such samples may be intelligently assessed.

This approach to measurement is associated with a theory of generalizability, proposed by Cronbach and others (1963). The task is stated as defining a domain of "conditions" where the word conditions has a very general meaning. It includes, for example, different elements of content. The problem 2 + 2 = __, and the problem 3 + 5 = __ are two different conditions for observing arithmetic skill. Similarly, a free response and a multiple choice item based on the same content would represent different conditions. In addition to such formal variations, conditions may vary temporally. For example, a domain may include delayed measures or only those taken at the close of the program. Variation in the situation in which observations are made is a more familiar use of "conditions" but is only one of many meanings assigned to the word in this conception.

There are no established procedures for defining measurement domains. Both listing of included and excluded elements, and stating rules for inclusion or exclusion would seem useful. In practice, some combination of these two techniques is often most effective.

TITLE III EVALUATORS TRAINING

MEASUREMENT UNIT 2

Defining Measurement Domains

I. Problem: Specifying what we wish to measure
   A. Rational constructs - achievement, anxiety, valuation
   B. Range of indicators must be defined

II. Domain Definition: Specification of "conditions" included in domain
   A. Definition of conditions: Any aspect of the observation or its setting which may vary
   B. Domain score: Mean score over all observations included in domain (percent correct if 1-0 scoring)
   C. Factors of domain definition
      1. entity to be measured - persons vs. classes or other aggregates
      2. content
      3. formal
      4. temporal
      5. observer
      6. situational

III. Techniques of Domain Definition
   A. Listing included conditions
   B. Rules for inclusion and exclusion
   C. Spiral use of listings and rules

IV. Summary
   A. Concept of domain of conditions
   B. Diverse ways conditions may vary
   C. Need for precise and complete specification

References:

B. S. Bloom, et. al. Taxonomy of Educational Objectives: Cognitive Domain. New York: Longmans, Green & Co., 1956.

L. J. Cronbach, et. al. Theory of Generalizability; Brit. J. Statistical Psychol. XVI Part II, 137-163, 1963.

D. Krathwohl, et. al. Taxonomy of Educational Objectives: Affective Domain. New York: David McKay, 1964.

Michael Scriven. The Methodology of Evaluation, in Perspectives of Curriculum Evaluation, AERA Monograph Series on Curriculum Evaluation, No. 1. Chicago: Rand McNally, 1967.

UNIT 2 ACTIVITIES

Using the selected sample situation, quickly select two of the traits, performances, behaviors, etc. that you wish to assess. Select:

1) one very clearly defined maximum performance domain, e.g. a skill or achievement
2) one typical behavior trait or domain, e.g. a habit, value, or attitude

Develop a definition of each domain. Be as complete as possible in the time allowed.

Maximum performance domain - 10 minutes

Typical behavior domain - 20 minutes

Do not limit yourself to paper-and-pencil self-report behavior for the typical behavior domain.

The reporter should be prepared to give a brief definition of each domain to the group and to note problems encountered in the definition process.

PERSON AND ITEM SAMPLING

Once a domain of conditions has been defined and a population of persons specified, two types of questions often appear. The first is whether the population of persons can perform adequately on a particular element of the measurement domain. This question is usually answered by estimating the proportion of the person population which can perform at or above some specified level. The second question is to what extent a particular person has mastered the entire measurement domain. This may often be approached by estimating the proportion of the conditions within the domain which the person could perform satisfactorily. Both questions can therefore be reduced to the estimation of a proportion, the proportion of persons passing a specified item, or the proportion of items passed by a specified person.

Estimation theory points out that in neither case is it necessary to make all possible observations to reach an adequate estimate. Therefore, the usual practice of selecting a few conditions (items) and administering them to all students in an innovative program is very often wasteful. When the primary interest is in the first type of question, it might be better to draw two sets of items and administer each set to half the students. Each set might be equally effective in estimating domain proportions for individuals, but the item performance data would be available for twice as many items. Once the habits of traditional procedures are broken, the range of possibilities for sampling items and persons expands.

Procedures can be developed from random sampling theory for describing how accurate the estimates provided by any particular sampling plan will be. Conversely, it is possible to specify the desired level of accuracy and find the number of items and persons which must be sampled. These procedures make possible the preparation of an efficient measurement plan for an innovative program.

TITLE III EVALUATORS TRAINING

MEASUREMENT UNIT 3

Person and Item Sampling

I. Problem: Describe typical level of performance of
   A. Universe of persons or
   B. Domain of tasks or items

II. Subproblems: Interest may be in
   A. Estimating proportion of persons correctly answering single item
   B. Estimating proportion of items correctly answered by a single person
   C. Estimating mean and dispersion of the proportion of items correctly answered
   D. Estimating covariation among items

III. Need for Sampling and Procedures
   A. Ideal
   B. Traditional research approach
   C. Norming approach
   D. Joint sampling

IV. Example of Various Sampling Techniques
   A. Item domain: 100 one digit multiple factor, paper and pencil free response, specified time
   B. Person domain: 500 students in study
   C. 50,000 responses: too great a number
   D. 10,000 response plans

V. Evaluation of Plans - Done in terms of questions
   A. Item "difficulty" (success of program on specific criteria)
      1. estimate proportion passing single items

$$T = Z\sqrt{\frac{\hat{p}\hat{q}}{n}\left(\frac{N-n}{N-1}\right)}$$

$$n = \frac{Z^2\hat{p}\hat{q}N}{Z^2\hat{p}\hat{q}+(N-1)T^2}$$

B. Individuals' scores (success of program with individuals)
   1. estimate proportion of items in the domain passed by this individual
C. Estimate mean and variance of population distribution of universe scores (what is typical performance)
D. Covariance among responses (is there a pattern of success and failure on items)

VI. Summary
A. Concern is for estimating several parameters
   1. typical performance (mean score)
   2. item probability
   3. person scores
   4. item covariances
B. Sampling designs for group measurement depend on
   1. desired accuracy of estimation
   2. relative importance of various objectives
   3. practical limitations
   4. use of random or stratified random sampling of available items and persons
C. Generalization beyond populations sampled is logical problem

References:

Thomas R. Knapp. An application of balanced incomplete block designs to the estimation of test norms. Educ. and Psychol. Meas., 28, 265-272; (1968)

Frederick M. Lord. Estimating norms by item sampling. Educ. and Psychol. Meas., 22, 259-267. (1962)

Lynnette B. Plumlee. Estimating means and standard deviations from partial data, Educ. and Psychol. Meas., 24, 623-630. (1964)

H. M. Walker and J. Lev. Statistical Inference. New York: Holt, 1953. (esp. pp. 68-72)

UNIT 3 ACTIVITIES

Choose a clearly defined measurement domain (such as the maximum performance domain from Unit 2 activities) and specify the size of the population of persons available in the sample situation selected. Assume all items are scored 1 or 0.

Work either Activity A or Activity B.

A. Develop a plan for sampling items and persons, and find the accuracy it gives for:
   1. estimating the proportion of persons correctly answering a specified item.
   2. estimating the proportion of the item domain which could be answered correctly by a specified person.
   3. estimating the mean performance for persons on the domain of items.

B. Specify the accuracy of estimation desired for:
   1. estimating the proportion of persons correctly answering an item (use an item with .50 difficulty) and
   2. estimating the proportion of the item domain an individual would pass (use a 50 percent person).

   Calculate the number of persons per item and items per person required and construct a sampling plan.

What is the accuracy produced for estimating typical performance on the domain of items?

Reporter should report

1) Which activity was attempted
2) What plan or tolerances were specified
3) What tolerances or plan resulted
4) Implications of results

TEST AND ITEM SELECTION

In selecting a standardized test or selecting items for a homemade test which will represent some measurement domain, the primary concern is whether the items used may be considered a reasonable sampling of the domain. A clear definition of the domain's dimensions and boundaries will provide the information necessary for a logical analysis of the question.

Empirical operations can also be helpful in analyzing a set of items by highlighting peculiar response patterns which suggest that an item is not dependent on the ability intended, and therefore may not be properly included as sampled from the specified domain. The item may then be discarded or revised. A classic example is the item response which has accidentally been incorrectly keyed. The tendency of students who otherwise perform well to miss this item and to choose the option which is actually correct calls the error to the examiner's attention.

The inconsistency between item performance and some more general performance which revealed the miskeyed item illustrates the general nature of item analysis. Items in a domain are expected to be homogeneous enough so that there is positive covariation between almost any pair of items and certainly between any item and the domain scope, a property usually called internal consistency. Most item analysis procedures are designed to show a lack of internal consistency.

Analyzing the nature of the inconsistency by studying the distribution of responses over the options of multiple choice items may assist the evaluator to find the source of the irregularity. These techniques provide empirical checks on domain definition and sampling which are helpful to the test constructor.

TITLE III EVALUATORS TRAINING

MEASUREMENT UNIT 4

Test and Item Sampling

I. Two types of Achievement Test:
   A. Subjective test - The grader is allowed to extensively exercise personal judgement in scoring the test.
   B. Objective test - The grader is permitted little, if any, freedom of personal judgement in scoring the test.

   The present discussion will be restricted to type B.

II. Standardized and Self-Made Achievement Tests
   A. Standardized test - A test for which items have been carefully selected and which has been administered to various normative groups.
      1. example of standardized tests - see handout
      2. advantages of standardized tests
      3. references for standardized tests - see handout
      4. considerations in choosing a standardized test
   B. Self-made test - A test constructed for a specific purpose and which has not been extensively used.
      1. item writing - see handout
      2. item analysis - see handout

References:

N. M. Downie. Fundamentals of Measurement: Techniques and Practices. New York: Oxford University Press, 1967.

J. R. Gerberich. Specimen Objective Test Items. New York: Longmans, Green & Co., 1956.

H. A. Greene, A. N. Jorgensen and J. R. Gerberich. Measurement and Evaluation in the Secondary School. New York: David McKay Co., 1964.

N. E. Gronlund. Measurement and Evaluation in Teaching. New York: The MacMillan Co., 1965.

A SELECTED LIST OF STANDARDIZED TESTS[1]

Test Name (Publisher's no.)*	Grade† Levels Covered	Testing* Time Minutes	Major Areas Measured**	MMY†† Review
ACHIEVEMENT BATTERIES				
California Achievement Tests (3)				
Elementary Levels	1-9	89-178	Reading, arithmetic, language	5-2
High School Level	9-14	178	Reading, mathematics, language	5-2
Essential High School Content Battery (6)	9-12	205	Mathematics, science, social studies, English	4-9
Iowa Tests of Basic Skills (7)	3-9	279	Reading, arithmetic, language, work-study skills	5-16
Iowa Tests of Educational Development (10)	9-12	459	Understanding and skills in English, mathematics, science, and social studies	5-17
Metropolitan Achievement Tests (6)				
Elementary Levels	1-9	105-255	Reading, arithmetic, language, science, social studies, study skills	4-18
High School Level	9-13	315	Reading, mathematics, language, science, social studies, study skills	New
SRA Achievement Series (10)	1-9	360-480	Reading, arithmetic, language, work-study skills	5-21
Sequential Tests of Educational Progress--STEP (5)				
Elementary Levels	4-9	455	Both levels-reading, writing, listening, essay, mathematics, science, social studies	5-24
High School Level	10-14	455		
Stanford Achievement Tests (6)				
Elementary Levels	1-9	127-255	Reading, arithmetic, language, science, social studies, work-study skills	New (5)
High School Level	9-12	282	Reading, mathematics, language, science, social studies, study skills	New

A SELECTED LIST OF STANDARDIZED TESTS
(Continued)

Test Name (Publisher's no.)*	Grade Levels† Covered	Testing* Time Minutes	Major Areas Measured**	MMY †† Review
READING TESTS				
Davis Reading Test (9)	8-13	40	Level and speed of comprehension	5-625
Diagnostic Reading Tests-Survey Section (10)	4-13	50-80	Rate, comprehension, vocabulary, word recognition	4-531
Durrell Analysis of Reading Difficulty (6)	1-6	30-45	Difficulties in silent and oral reading, listening comprehension, word analysis, phonetics, pronunciation, writing and spelling	5-660
Gates Advanced Primary Reading Tests (2)	2-3	40	Word recognition, paragraph reading	5-630
Gates Basic Reading Tests (2)	3-8	70	General significance, directions, details, vocabulary, comprehension	5-631
Gates Reading Survey (2)	3-10	45-60	Speed, accuracy, vocabulary, comprehension	5-633
Iowa Silent Reading Tests (6)	4-13	45-49	Rate, comprehension, word and sentence meaning, work-study skills	3-489
Kelley-Greene Reading Comprehension Test (6)	9-13	75	Rate, comprehension, directed reading, retention of details	5-636
Nelson-Denny Reading Test (7)	9-16, A	30	Vocabulary, comprehension, rate	New (4)
Nelson Reading Test (7)	3-9	30	Vocabulary, comprehension	New (4)
Reading Comprehension: Cooperative English Test (5)	9-14	40	Vocabulary, level and speed of comprehension	New (5)
SRA Reading Record (10)	6-12	45	Vocabulary, comprehension, rate	4-550
(Also see reading tests in achievement batteries listed above.)				

* The publishers' numbers (in parentheses) refer to the list of publishers which follows.
† Gives total span only--not the number of separate levels available. (K = kindergarten, A = adult).
* Ranges in time are mainly due to different forms used at different grade levels.
** Indicates the general areas measured but not the specific scores available.
†† Refers to Mental Measurements Yearbook and entry (e.g., 5-2 = second entry in fifth Yearbook). New (5) = new edition, old edition reviewed in Fifth Yearbook.

[1] Reproduced from: N.E. Gronlund, Measurement and Evaluation in Teaching. New York: The MacMillan Co., 1965.

Sources of Information About Standardized Tests

O. K. Buros. Tests in Print. Highland Park, N. J.: Gryphon Press, 1961.

O. K. Buros. Mental Measurements Yearbook. Gryphon Press. Published periodically.

Test Publishers:

1. American Guidance Service, Inc.
   720 Washington Avenue, S.E.
   Minneapolis, Minnesota 55414

2. Bureau of Publications
   Teachers College,
   Columbia University
   New York, New York 10027

3. California Test Bureau
   Del Monte Research Park
   Monterey, California 93940

4. Consulting Psychologists Press, Inc.
   577 College Street
   Palo Alto, California 94306

5. Cooperative Test Division
   Educational Testing Service
   Princeton, New Jersey 08541

6. Harcourt, Brace & World, Inc.
   757 Third Avenue
   New York, New York 10017

7. Houghton Mifflin Company
   2 Park Street
   Boston, Massachusetts 02107

8. Personnel Press, Inc.
   188 Nassau Street
   Princeton, New Jersey 08541

9. Psychological Corporation
   304 East 45th Street
   New York, New York 10017

10. Science Research Associates, Inc.
    259 East Erie Street
    Chicago, Illinois 60611

TITLE III EVALUATORS TRAINING

MEASUREMENT

Suggestions for Item Writing

General Suggestions

1. Express the item as clearly as possible.
2. Choose words that have precise meaning wherever possible.
3. Avoid complex or awkward word arrangements.
4. Include all qualifications needed to provide a reasonable basis for response selection.
5. Avoid the inclusion of cofunctional words in the item.

   Poor: When sailors put out to sea for long periods of time, vitamin C, in most instances, is added to diets to prevent
   A. beri-beri
   B. cretinism
   C. sterility
   + D. scurvy

   Better: Vitamin C is added to diets to prevent
   A. beri-beri
   B. cretinism
   C. sterility
   + D. scurvy
6. Avoid unessential specificity in the stem or the responses.

   Poor: If President Nixon and Vice President Agnew were to die, they would be succeeded by
   + A. Speaker of the House McCormack
   B. Chief Justice of the Supreme Court Warren
   C. Secretary of State Rogers
   D. Secretary of Defense Laird

   Better: If the President and Vice President of the United States were to die, they would be succeeded by
   + A. the Speaker of the House
   B. the Chief Justice of the Supreme Court
   C. the Secretary of State
   D. the Secretary of Defense
7. Avoid irrelevant inaccuracies in any part of the item.
8. Adapt the level of item difficulty to the group and purpose for which it is intended.
9. Avoid irrelevant clues to the correct response.

   Poor: A test is said to be valid when
   + A. it measures what it is supposed to measure
   B. including only multiple-choice items
   C. reliability is important too
   D. to score it one is objective

   Better: A test is said to be valid when
   + A. it measures what it is supposed to measure
   B. it includes only multiple-choice items
   C. it is reliable
   D. it is objective
10. In order to defeat the rote-learner, avoid stereotyped phraseology in the stem or the correct response.
11. Avoid irrelevant sources of difficulty.

Suggestions for Item Writing

(Continued)

Short Answer Form

1. Use the short-answer form only for questions that can be answered by a unique word, phrase, or number.
2. Do not borrow statements verbatim from context and attempt to use them as short-answer items.
3. Make the question, or the directions, explicit.
4. Allow sufficient space for pupil answers, and arrange the spaces for convenience in scoring.
5. In computational problems specify the degree of precision expected, or better still, arrange the problems to come out even unless the ability to handle fractions and decimals is being tested.
6. Avoid overabundance of completion exercises.

The True-False Form

1. Base true-false items only on statements which are true or false without qualifications.
   Poor: It is a short trip from Chicago to Detroit. (T or F)
   Better: In a superjet, it is a short trip from Chicago to Detroit. (T)
2. Avoid the use of long and involved statements with many qualifying phrases.
   Poor: If the President were to die and if the Vice President were to assume command and then also die, the Speaker of the House would become President. (T)
   Better: If the President and Vice President both die, the Speaker of the House becomes President. (T)
3. Avoid the use of sentences borrowed from texts or other sources as true-false items.

Multiple-Choice Form

1. Use either a direct question or an incomplete statement as the item stem.
   Poor: Charles Darwin
   A. was a renowned chemist
   + B. formulated a theory of evolution
   C. discovered the proton
   D. proved the Central Limit Theorem
   Better: Charles Darwin was a
   A. chemist
   + B. naturalist
   C. physicist
   D. statistician
2. In general, include in the stem any words that must otherwise be repeated in each response.
   Poor: One of the major functions of the adrenal gland is
   + A. to regulate the amount of sugars in the blood
   B. to regulate the amount of protein sent to body cells
   C. to regulate the secretion of wastes
   D. to regulate the amount of insulin.

Suggestions for Item Writing

(Continued)

Better: One of the major functions of the adrenal gland is to regulate the

+ A. amount of sugars in the blood
B. amount of protein sent to body cells
C. secretion of wastes
D. secretion of insulin

3. If possible, avoid a negatively stated item stem.
4. Provide a response that competent critics can agree is the best.
5. Make all the responses appropriate to the item stem.
6. Make all distracters plausible and attractive to examinees who lack the information or ability tested by the item.

Poor: The area of a circle with a diameter equal to 12 is approximately

A. 19 (using πr)
B. 38 (using πd)
+ C. 113 (using πr^2)
D. 453 (using πd^2)

7. Avoid highly technical distracters.
8. Avoid responses that overlap or include each other.
9. Use "none of these" as a response only in terms to which an absolutely correct answer can be given; use it as an obvious answer several times early in the test but use it sparingly thereafter; avoid using it as the answer to items in which it may cover a large number of incorrect responses.
10. Arrange the responses in logical order, if one exists, but avoid consistent preference for any particular response position.
11. If the item deals with the definition of a term, it is often preferable to include the term in the stem and present alternative definitions in the responses.
12. Do not present a collection of true-false statements as a multiple-choice item.

Matching Exercises

1. Group only homogeneous premises and homogeneous responses in a single matching item.

Poor:

____ 1.	$-\Sigma X/N$	A: standard deviation
____ 2.	statistician	B. mean
____ 3.	$\Sigma(X-\bar{X})^2/N$	C. Samuel Wilks
		D. variance

Better:

____ 1.	$\Sigma X/N$	A. standard deviation
____ 2.	$\frac{X-\bar{X}}{S}$	B. mean
		C. standard score
____ 3.	$\Sigma(X-\bar{X})^2/N$	D. variance

Suggestions for Item Writing

(Continued)

2. Use relatively short lists of responses.
3. Arrange premises and responses for maximum clarity and convenience to the examinee.
4. The directions should clearly explain the intended basis for matching.
5. Do not attempt to provide perfect one-to-one matching between premises and responses (more responses than premises).

Item Analysis Formulae

1. DIFFICULTY OF ITEM i $= \frac{C_i}{T_i} \times 100$

C_i = the number of persons answering item i correctly
T_i = the total number of persons who responded to item i

2. DISCRIMINATING POWER OF ITEM i $= \frac{C_{Ui} - C_{Li}}{T_i/2} = D_i$

C_{Ui} = the number of persons scoring in the upper half on the test and who answer item i correctly.
C_{Li} = the number of persons scoring in the lower half on the test and who answer item i correctly.
T_i = the total number of persons who responded to item i

3. TEST RELIABILITY = KR20 $= \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k} p_i q_i}{S^2}\right]$

k = total number of items on the test.
$p_i = \frac{C_i}{T} \equiv$ proportion of persons responding correctly to item i. ($C_i \equiv$ number of persons answering item i correctly; $T \equiv$ total number of persons taking the test)
$q_i = 1 - p_i$

$S^2 = \frac{\sum_{t=1}^{T} (X_t - \bar{X})^2}{T} \equiv$ variance of the total test scores

Exercise - A group of 100 persons took a four-item test; and the following outcomes were observed.

Item No.	No. Correct in Upper Half	No. Correct in Lower Half	Total No. Responding
1	50	0	100
2	15	5	100
3	50	50	100
4	10	40	100

The mean and variance of the test were determined to be 2.20 and 1.32 respectively. Determine:

(1) the difficulty of each item. Which items are too easy and which are too difficult?
(2) the discriminating power of each item. Which items are good discriminators and which are not?
(3) Is the test highly reliable or not?

Answers to Exercise

Item No.	No. Correct in Upper Half	No. Correct in Lower Half	Total No. Responding
1	50	0	100
2	15	5	100
3	50	50	100
4	10	40	100

$\overline{X} = 2.20$

$S^2 = 1.32$

(1)

i	C_i	T_i
1	50 + 0 = 50	100
2	15 + 5 = 20	100
3	50 + 50 = 100	100
4	10 + 40 = 50	100

Difficulty of 1 = $\frac{50}{100}$ x 100 = 50%

Difficulty of 2 = $\frac{20}{100}$ x 100 = 20% (too difficult)

Difficulty of 3 = $\frac{100}{100}$ x 100 = 100% (too easy)

Difficulty of 4 = $\frac{50}{100}$ x 100 = 50%

(2)

i	C_{Ui}	C_{Li}	$C_{Ui} - C_{Li}$	T_i	$T_i/2$
1	50	0	50	100	50
2	15	5	10	100	50
3	50	50	0	100	50
4	10	40	-30	100	50

$D_1 = 50/50 = 1.00$ (good discriminator)

$D_2 = 10/50 = .20$ (weak discriminator)

$D_3 = 0/50 = .00$ (does not discriminate)

$D_4 = -30/50 = -.60$ (negative discriminator)

(3)

i	$C_i P_i$	q_i	$p_i q_i$
1	(50 + 0)/100 = .50	.50	.2500
2	(15 + 5)/100 = .20	.80	.1600
3	(50 + 50)/100 = 1.00	.00	.0000
4	(10 + 40)/100 = .50	.50	.2500
			.6600 = $\sum_{i=1}^{4} p_i q_i$

$$KR\text{-}20 = \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k} p_i q_i}{s^2}\right]$$

$$= \frac{4}{3}\left[1 - \frac{.66}{1.32}\right]$$

$$= \frac{4}{3}\left[1 - \frac{1}{2}\right]$$

$$= \frac{4}{3} \times \frac{1}{2}$$

$$= \frac{2}{3} \simeq .67 \text{ (not highly reliable)}$$

OBJECTIVE OBSERVATION

The concern for meaningful measurements often suggests the use of techniques other than the typical pencil and paper test. A broad and useful class of data gathering procedures is objective observation. There are, however, many ways error may creep into such measurements, many possible sources of "slippage" between the raw input to the observer and the recording of a number or symbol representing that observation. This problem has recently been studied by Webb, et. al. (1966) who have emphasized two qualities of measurements which help achieve the general ideals of meaningfulness and precision.

The first is nonreactivity, a quality achieved when the measurement process does not affect the thing being measured. A common problem is that reactive measurements often become an important part of the treatment. On the other hand, a reactive measure may produce a temporary effect, making a meaningful measurement impossible. The effect of observers in small groups is an obvious example. Allowing adaptation periods and undetectable observation are two techniques for countering reactivity. The latter, of course, raises questions of ethics.

The second quality is consistency of calibration, a property existing when the same phenomenon observed twice will produce the same measurement. The tendency of participant observers to notice certain things when they first join a new culture, and other things after they have observed for some time, illustrates inconsistency of calibration. A more common illustration is provided by the decrease of alertness resulting from fatigue during a series of consecutive observations. Training, simplicity and clear definition of procedures, and attention to physical limitations help keep calibration consistent.

A final concern not emphasized by Webb is the need for reasonable sampling of the behavior domain. The risk of using a single behavior to represent a domain is an instance of generalizing from one case.

TITLE IV EVALUATORS TRAINING

MEASUREMENT UNIT 5

Objective Observation

I. Stages
   A. Collection
   B. Storage
   C. Reduction
   D. Storage
   E. Summarization
   F. Storage
   G. Reporting

II. Some Issues
   A. Should collection and reduction be combined?
   B. How may selectivity be controlled?

III. Ideals
   A. Consistency of calibration
   B. Nonreactiveness
   C. Unbiased sampling of domain of conditions

IV. Reduction of Information:
   A. Accuracy (objectivity)
   B. Meaningfulness

V. Storage: Problems and process
   A. Files: liquor cabinet vs. cemetery
   B. Coded data
   C. Housekeeping vs. housecleaning
   D. Written reports

VI. Summary
   A. Major processes - observation, reduction, storage
   B. Objectives - accuracy and meaningfulness
   C. Techniques

Reference:

J. W. Webb, D. T. Campbell, R. D. Schwartz and L. Sechrest. Unobtrusive Measures. Chicago: Rand McNally. 1966.

UNIT 5 ACTIVITY

Choose one of the characteristics, behaviors, etc. suggested in the Unit 1 Activities (perhaps the typical performance domain analyzed in Unit 2). Suggest as many ways in which the behaviors of the domain could be observed as you can in 20 minutes. In the final 10 minutes, analyze each observation procedure for (1) nonreactiveness and (2) consistency of calibration.

Be free with suggestions during the first phase.

The reporter should select a few of the procedures for presentation on the basis of creativity, quality, or interesting problems presented.

SAMPLE SITUATION A

Prekindergarten Program for Disadvantaged Children

The primary purpose of this program is to increase verbal communication skills and broaden the children's range of experience. Ninety 4-year-old children will be selected from a depressed area of the city. They will spend ½ day, 5 days a week at the center, for 1 school year. The curriculum will be planned by three teachers and a developmental psychologist available 1 day a week. The teachers will be assisted by the equivalent of six full-time persons recruited from the childrens' parents. It is believed that participation by a parent may have a substantial effect on the home environment. The program will be conducted in the basement of the Methodist church. Available are two office-sized rooms, two slightly larger rooms, and a large open area. Desired equipment will be provided by project or community funds.

SAMPLE SITUATION B

Introduction of Teacher Aides

The primary objective of this program is to provide the teacher more teaching time by assigning nonteaching tasks to teacher aides. It is assumed that achievement of the objective will result in improved student achievement and teacher morale. Twenty 4th grade classrooms will participate. An aide will be available to each teacher during all school hours. After orientation by the central unit, each aide will be assigned to a teacher, to do whatever the teacher asks. The central unit will provide short training sessions as requested by the teachers or aides. The aides must have some college education. The setting is suburban.

SAMPLE SITUATION C

Individualizing Instruction Through Computer Based Resource Units

The primary objective of this program is to increase individualization of instruction by providing lists of materials, activities, and projects appropriate to the teacher's objectives and the child's individual characteristics. Twenty 11th grade social studies classes will participate. For each class the teacher will choose from a list of objectives and record each child's individual characteristics on a check sheet. Abilities, interests, and background factors are included. From the computer-stored unit, the teacher will receive lists of group activities and individual lists of resources and activities for each child. Each teacher will use three such units during a single semester. No special provision of materials will be made. It is expected that successful individualization will result in improved student interest and achievement and a feeling of productivity in the teachers.

SAMPLE SITUATION D

Improving Interracial Attitudes and Knowledge

This is a two-pronged study to be conducted in the 8th grades of four school districts. The objective is to insure knowledge of the Negro's contribution to past and present sccieties, and to produce favorable attitudes toward other groups. Each of the schools is about 50 percent white. Preparations will be made during the fall semester and activities conducted during the spring.

A panel of teachers, augmented by a curriculum specialist, a Negro leader, and a full-time clerical worker will collect materials and activities relevant to the units normally taught in 8th grade, stressing the contributions of Negroes and the Negro community. The widest possible range of subject matters will be covered. The panel will also suggest ways the special materials can be integrated into the usual unit presentation. All teachers will use at least some of the materials.

The second prong consists of an interested university group training teachers in techniques for changing attitudes. Procedures relevant to each major subject will be provided. Procedures for altering both whites' attitudes toward blacks and blacks' attitudes toward whites will be supplied. Each major subject teacher agrees to use two of the provided attitude change routines during the second semester.