REPORT RESUMES

ED 012 087 CG 000 250

GUIDE TO ASSESSMENT AND EVALUATION PROCEDURES, THE NEW ENGLAND EDUCATIONAL PROJECT.

BY- ANNAS, PHILIP A. AND OTHERS

NEW ENGLAND EDUCATIONAL ASSESSMENT PROJECT

PUB DATE OCT 66

EDRS PRICE MF-$0.09 HC-$1.48 37P.

DESCRIPTORS- *PROGRAM EVALUATION, *PILOT PROJECTS, *EVALUATION METHODS, GLOSSARIES, BIBLIOGRAPHIES, WORKBOOKS, NEW ENGLAND

THE RECENT PROLIFERATION OF FEDERAL SUPPORT PROGRAMS IN EDUCATION HAS BROUGHT AN INCREASED DEMAND FOR CAREFULLY PLANNED, FORMAL EVALUATION AT BOTH THE STATE AND LOCAL LEVELS. IN ORDER TO AID LOCAL SCHOOL SYSTEMS IN THE COMPLEX WORK OF EVALUATION OF TITLE I AND OTHER SPECIAL EDUCATION PROJECTS, THIS GUIDE HAS BEEN PREPARED. IT PRESENTS IN WORKBOOK FORM A STEP-BY-STEP PROCESS OF EVALUATING A PROJECT. A GLOSSARY OF TERMS AND BIBLIOGRAPHY ARE ALSO INCLUDED. (NS)

GUIDE TO ASSESSMENT AND EVALUATION PROCEDURES

THE NEW ENGLAND EDUCATIONAL ASSESSMENT PROJECT

Special Educational Programs Committee

Russell Capen, Connecticut
J. Wilfrid Morin, Maine
Everett Thistle, Massachusetts
Thomas Burns, New Hampshire
Edward T. Costa, Rhode Island
Walter D. Gallagher, Vermont

Consultant: Dr. Alexander Smith
Southern Connecticut State College

Co-chairmen: Philip A. Annas, Maine
Richard A. Dowd, Rhode Island

October, 1966

NEW ENGLAND COMMISSIONERS OF EDUCATION

William J. Sanders, Connecticut
William T. Logan, Jr., Maine
Owen B. Kiernan, Massachusetts
Paul E. Farnum, New Hampshire
William P. Robinson, Jr., Rhode Island
Richard A. Gibboney, Vermont

THE NEW ENGLAND EDUCATIONAL ASSESSMENT PROJECT

State Representatives

James M. Burke, Connecticut
Philip A. Annas, Maine
John A. Torosian, Massachusetts
Frank W. Brown, New Hampshire
Edward F. Wilcox, Rhode Island
Karlene V. Russell, Vermont
Acting Project Director, 1966, Dr. John A. Finger, Jr.
Project Director, Dr. Ermo H. Scott

PREFACE

Although most school programs are the constant subjects of informal evaluation by students, parents, and school officials, the recent proliferation of federal support programs in education has brought an increased demand for carefully planned formal evaluation at both the state and local levels.

The ability of the states to meet their obligations in this respect is dependent on successful evaluation at the local level. In order to aid local school systems in the complex work of evaluation of Title I and other special education projects, the New England Educational Assessment Project has prepared the *Guide to Assessment and Evaluation Procedures.*

The Assessment Project, funded under Title V of the Elementary and Secondary Act of 1965, is a cooperative effort on the part of the six New England state departments of education to develop criteria and procedures for assessing education programs and to strengthen regional education leadership through assessment. At the outset, it should be noted that the *Guide* is intended to be of help only in the general design of an evaluation program and in the use of some elementary instruments used in measurement. In some cases, the *Guide* will prove sufficient for the design of the complete assessment program. Where complex test design or statistical analysis may be necessary, the services of specialists should be recruited.

TABLE OF CONTENTS

(Blank assessment planning charts are included in the pocket on the back cover)

INTRODUCTION

Evaluation is the process of determining relative worth. This is usually done by comparing an established standard with something of unknown value. Once this has been done, the adequacy of the hypothesis under study can be stated in terms of its relationship to the established standard; i.e., the performance of the students under study is greater than, equal to, or less than the established standard, or better than, equal to, or worse than the norm which has been used for comparison. In any case, however, evaluation cannot begin until appropriate standards have been determined.

Examples:

ESTABLISHED STANDARD	COMPARISON
Last Year's Gain	**This Year's Gain**
achievement test shows .8 grade placement or 10 raw score points	achievement test shows 1.2 grade placement or 16 raw score points
350 books drawn from library	423 books drawn from library
—10 pupil hours in detention room	—18 pupil hours in detention room
—4 pupils dropping out of school	—6 pupils dropping out of school
6.4% pupils responding "yes" to question, "Do you ever visit the museum?"	8.7% pupils responding "yes" to question, "Do you ever visit the museum?"

Generally speaking, standards are of two kinds: those that have been established by comparison with other standards, or those that have been arbitrarily derived for local needs. Since the central problem of evaluation is to arrive at the most accurate and worthwhile judgments of value, the importance of choosing proper standards cannot be overstressed. The best evaluations are possible when they are based on quantitative data obtained from objective sources rather than from descriptive judgments obtained directly. The more objective the standard, the more valid can be the evaluation.

The process of evaluation is vitally important to the success of any special educational project, and although some highly complex statistical procedures have been developed in the quest for more accurate educational measurement, evaluation need not be thought of as a mysterious and unduly complex procedure reserved solely for highly trained specialists. The formulas suggested in the *Guide* should not discourage the project director who is not familiar with formal measurement procedures. They are included for the benefit of those who have some background in statistical procedures; but even when a statistical analysis is not attempted, if those who are involved in the implementation of special education projects follow the general guidelines of this publication, take great care that the instruments used for measurement are pertinent to the project activity and to the individual pupils involved, and seek expert assistance when necessary, a reasonable degree of success in evaluation can be expected.

HOW TO USE THE GUIDE

The *Guide* has been prepared in workbook style for easy use. The directions and information are presented in the actual order of use. Those who use the *Guide* are advised to follow the suggested process, step by step. In the graphic illustration below, the format of the *Guide* is illustrated in outline form.

THESE STEPS MUST BE TAKEN BEFORE PUPIL PARTICIPATION IN THE PROJECT BEGINS.

Fill out the sample assessment planning chart on pages 4 and 5.

Consult page 6 as an aid to completing the assessment planning chart.

Read pages 8–15 on testing.

Complete the experimental design according to the instructions on pages 16 and 17.

THESE STEPS MUST BE TAKEN DURING PUPIL PARTICIPATION IN THE PROJECT.

Read page 19 and follow the directions regarding the construction of the evaluative procedures which must occur during the time that pupils are participating in the project.

THESE STEPS MUST BE TAKEN AFTER PUPIL PARTICIPATION IN THE PROJECT IS OVER.

Read and follow the directions given on pages 20–25 which explain the methods of summarizing and analyzing the project data.

A glossary of terms and a bibliography follow the instructional procedures.

Several blank assessment planning charts have been included with the *Guide*. Some may find it more convenient to work on a blackboard, but for most cases, the planning chart should be sufficient. In using the planning charts, it will not be necessary to fill in every box in all four columns. Fill in only those boxes which are pertinent to your project.

Now, to begin, remove one of the blank planning charts, open the *Guide* to the next page, follow the directions carefully, and begin the construction of your own assessment and evaluation procedures.

STEP I

THE ASSESSMENT PLANNING CHART

(Step I must be completed before the pupils begin their participation.)

DIRECTIONS

OBJECTIVE

On a separate piece of paper, list all the objectives of the project. Then using a separate planning chart for each objective, write the objective in the box in column A. In parentheses under each objective write the code number as listed in *Instructions for Title I 1967 Application tion Forms OE-37003,* page 13. Extra planning charts are included in this guide.

METHODS OF EVALUATION

Expand each objective listed to include any outcome you expect or hope for, regardless of how difficult its measurement may seem. Remember, that for each objective listed, some kind of evaluation should be presented at the end of the project. The expanded objectives should be listed in the boxes in column B.

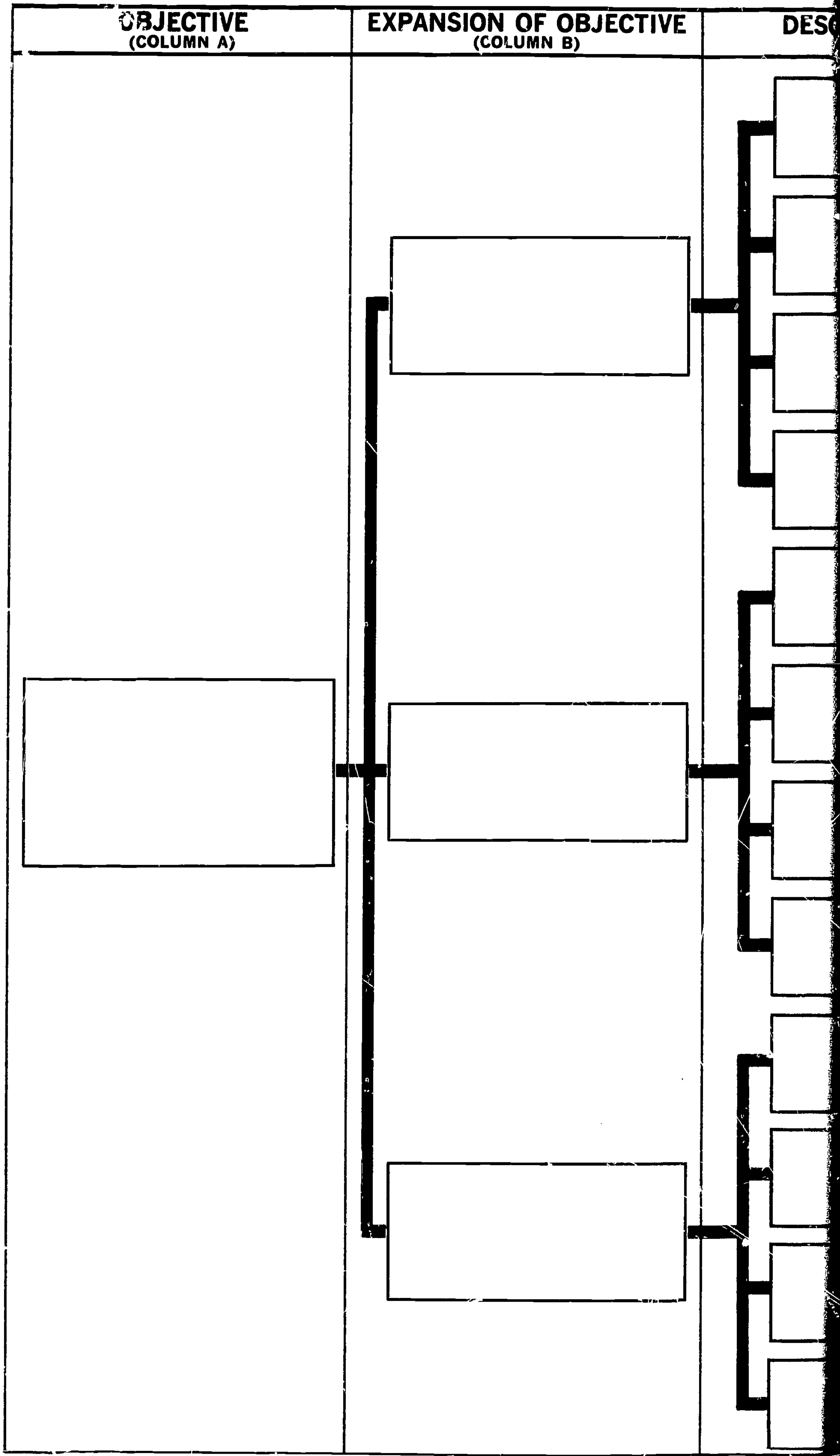
DESCRIPTION OF PUPIL BEHAVIORS

Translate each expanded objective listed in column B into brief descriptions of actual pupil behaviors. List each behavior separately and describe specifically what the pupil should do at the conclusion of the project. List the descriptions of pupil behaviors in column C.

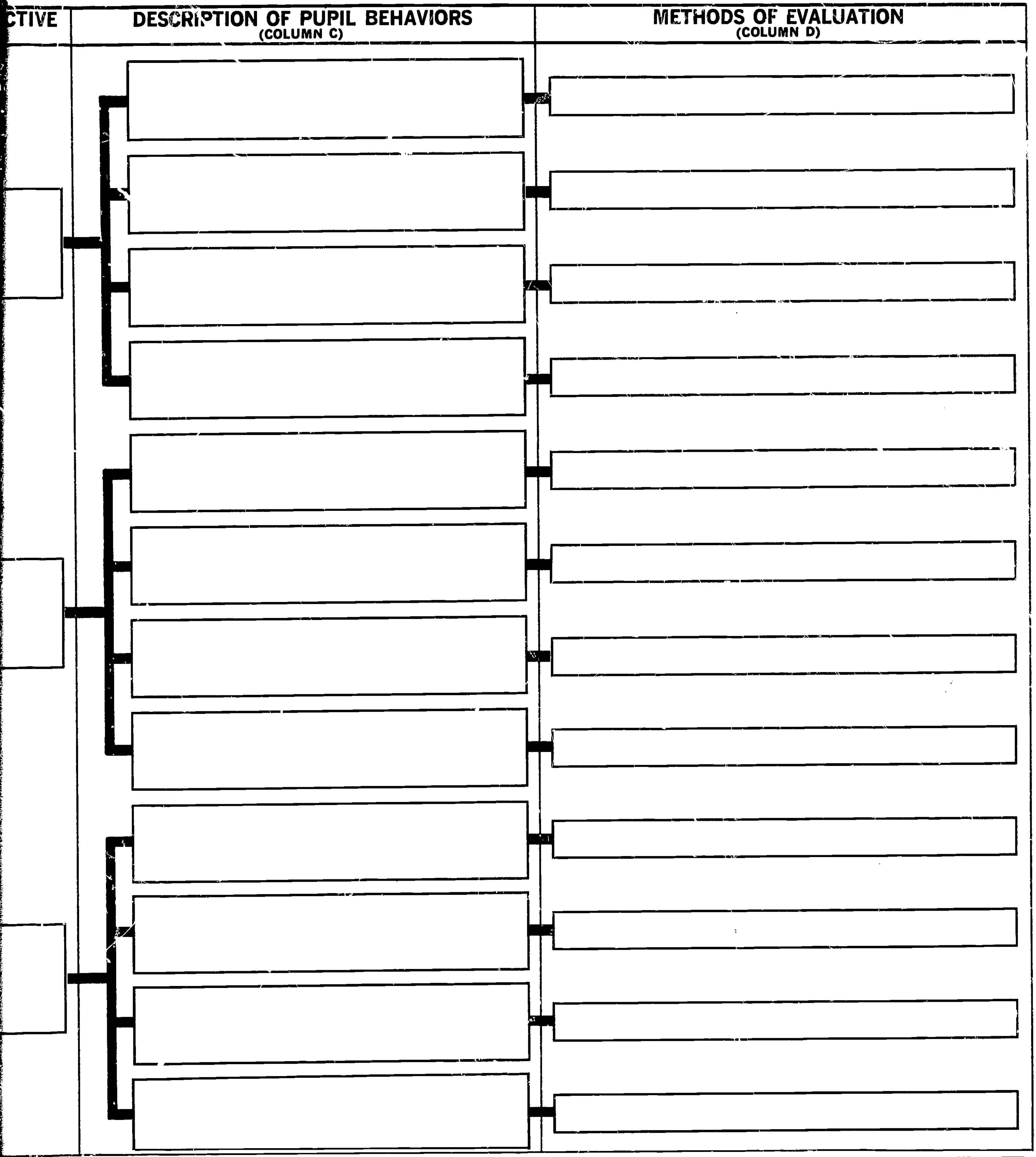
EXPANSION OF OBJECTIVE

Consult the table of representative learning outcomes and possible methods of evaluation that follows this sample planning chart. (See page 6.) From the table, choose the methods of evaluation which apply to the specific behaviors listed in column C. List the pertinent evaluation methods in column D.

ASSESSMENT PLAN



ENT PLANNING CHART (THIS MUST BE DONE BEFORE THE PUPILS ARRIVE)



CTIVE

DESCRIPTION OF PUPIL BEHAVIORS
(COLUMN C)

METHODS OF EVALUATION
(COLUMN D)

DIRECTIONS

OBJECTIVE

On a separate piece of paper, list all the objectives of the project. Then using a separate planning chart for each objective, write the objective in the box in column A. In parentheses under each objective write the code number as listed in *Instructions for Title I 1967 Application tion Forms OE-37003*, page 13. Extra planning charts are included in this guide.

METHODS OF EVALUATION

Expand each objective listed to include any outcome you expect or hope for, regardless of how difficult its measurement may seem. Remember, that for each objective listed, some kind of evaluation should be presented at the end of the project. The expanded objectives should be listed in the boxes in column B.

DESCRIPTION OF PUPIL BEHAVIORS

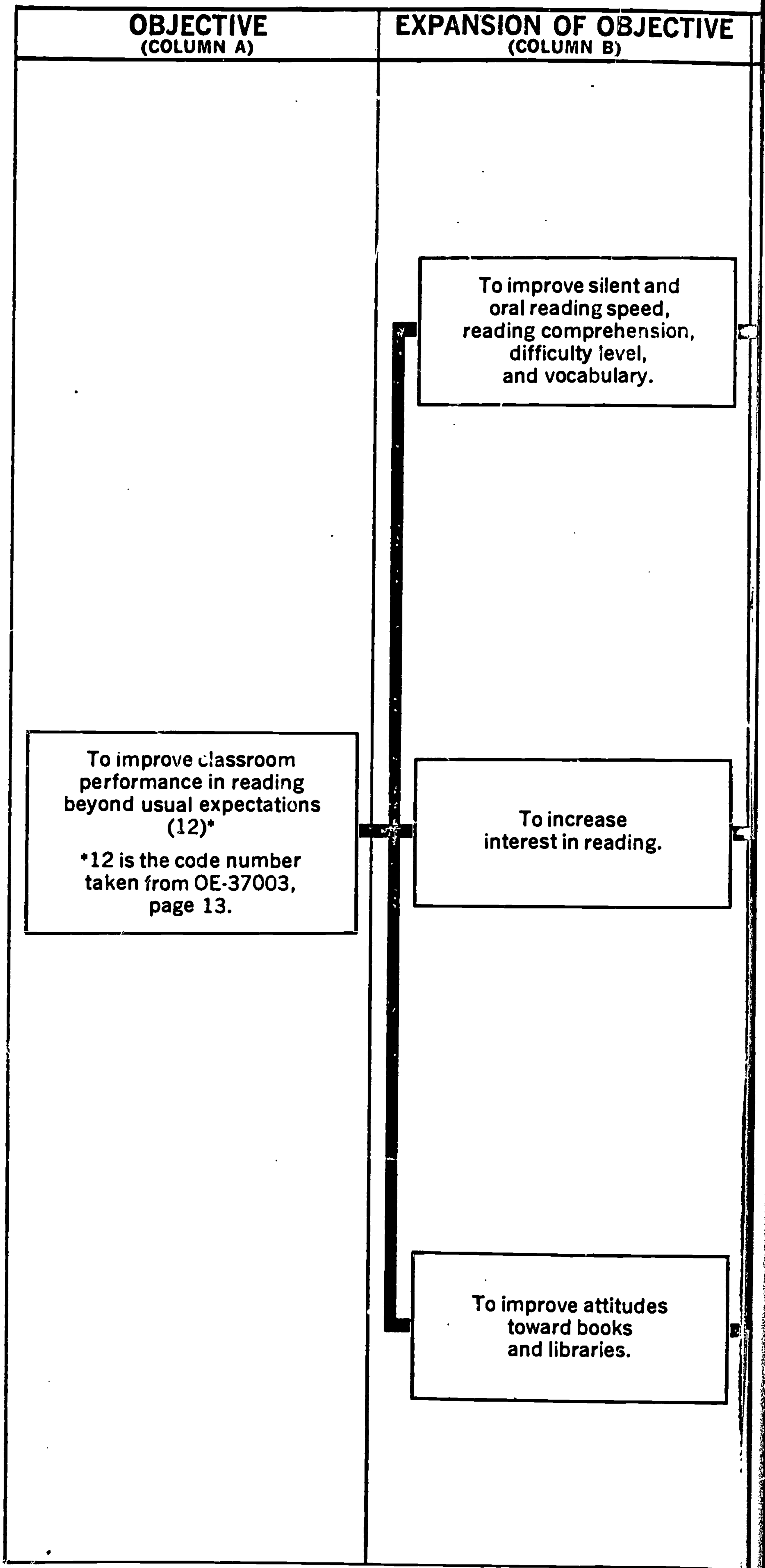
Translate each expanded objective listed in column B into brief descriptions of actual pupil behaviors. List each behavior separately and describe specifically what the pupil should do at the conclusion of the project. List the descriptions of pupil behaviors in column C.

EXPANSION OF OBJECTIVE

Consult the table of representative learning outcomes and possible methods of evaluation that follows this sample planning chart. (See page 6.) From the table, choose the methods of evaluation which apply to the specific behaviors listed in column C. List the pertinent evaluation methods in column D.

*NOTE: In the SAMPLE ASSESSMENT PLANNING CHART, column C, some boxes have been left undesignated. This has been done to indicate that although each expanded objective has many descriptions of pupil behaviors, some limitations should be arrived at in order to place proper emphases on the more important objectives and in order to assure that the evaluation process will be constructed in a practical length.

SAMPLE ASSESSMENT



OBJECTIVE (COLUMN A)

To improve classroom performance in reading beyond usual expectations (12)*

*12 is the code number taken from OE-37003, page 13.

EXPANSION OF OBJECTIVE (COLUMN B)

To improve silent and oral reading speed, reading comprehension, difficulty level, and vocabulary.

To increase interest in reading.

To improve attitudes toward books and libraries.

PLANNING CHART (THIS MUST BE DONE BEFORE THE PUPILS ARRIVE)

DESCRIPTION OF PUPIL BEHAVIORS (COLUMN C)	METHODS OF EVALUATION (COLUMN D)
Pupil reads at measurable greater rate than in pretest. (speed)	Objective test, standardized or locally made.
Pupil understands longer sentences with increased accuracy. (comprehension)	Objective test, standardized or locally made.
Pupil understands more difficult sentences, paragraphs, and passages. (difficulty)	Objective test.
Pupil reads orally with increased accuracy of pronunciation, word emphasis, inflection, sounding unknown words. (oral reading)	Checklist, objective test.
Pupil indicates increased preference for reading on self report checklist or rating scale. (interest)	Questionnaire, checklist, interest inventory.
Pupil is observed by teacher to show or express increased interest in reading on teacher rated checklist or anecdotal report. (interest)	Questionnaire, checklist, interest inventory.
*SEE NOTE IN DIRECTIONS COLUMN	
Pupil expresses or shows moderate or strong feelings about books or libraries. (attitude)	Rating scale, questionnaire, checklist, attitude loaded objective test.

TABLE OF REPRESENTATIVE LEARNING OUTCOMES AND POSSIBLE METHODS OF EVALUATION

(Numbers in parentheses refer to code designation taken from *Instructions for Title I 1967 Application Forms OE-37003*, page 13.)

TYPES OF BEHAVIORS (LEARNING OUTCOMES)	POSSIBLE METHODS OF EVALUATION
A Application (11-14) Concept Acquisition (11-14) Memorization of Facts (11-14) Problem Solving (11-14) Reading Comprehension (11-14) Skills (number, etc.) (11-14)	**A** Objective Test, Product Evaluation, Rating Scale, Checklist
B Performance (11)	**B** Rating Scale, Checklist, Product Evaluation
C Classroom Behavior (41-45)	**C** Rating Scale, Checklist, Attendance Record, etc.
D Interest (14)	**D** Questionnaire, Checklist, Interest inventory, Factual Vocabulary Test (with words from various interest fields)
E Attitude (31, 32)	**E** Rating Scale, Questionnaire, Checklist, or Objective Test (with factual material that has attitude-loaded responses.)
F Aspiration Level (33, 34)	**F** Rating Scale, Interview, Simple Objective Test, Word Association Test, Open Ended Sentences (psychologist needed.)
G Adjustment (53)	**G** Rating Scale, Anecdotal Report, Interview, Sociogram

After you have chosen the proper methods of evaluating the expected learning outcomes of your project, write your choices in column D on the assessment planning chart.

STEP II

SELECTION AND DEVELOPMENT OF INSTRUMENTS

(Step II must be completed before the pupils begin their participation.)

In this section:

STANDARDIZED TESTS

A Standardized Test is a test that has been given to a specified group of pupils (the norm group) and the results presented in organized fashion (tables of norms) so that a pupil who takes the test may be compared with this group. Types of interpretive scores given on norms tables are grade equivalents, intelligence quotients, mental ages, percentile ranks, and stanines. Sometimes raw scores, usually number of items correct, are converted to intermediate scores, such as standard scores, or converted weighted scores (to allow for equating raw scores from different forms of the test) before translation to these interpretative scores. Standardized tests are the most objective devices presently available for measuring factual recognition, certain skills, concepts, understandings, and problem solving, and sometimes interests, attitudes, and personality. They should be used only if found satisfactory for the project and for specific pupils involved on each of the following matters:

1. The test should be available in at least two equivalent forms (for pretest and post-test). If you wish to use the same test for selecting pupils for the project, it is best practice to use a different (third) form of the test for the selection.
2. Face validity analysis: Go over each item (question) of the test to be sure that:
   a. What the pupil actually does in getting the correct answer is the behavior that you want to test.
   b. The distractors (incorrect answers) are plausible to the pupils of your project, (or are they out of their range of familiarity?)
   c. All words and symbols or pictures are familiar to the culturally deprived (is there cultural bias?)—*unless* this is the sort of change you hope to measure.
3. Norm group analysis: Compare background and other descriptive information given by the publisher for the norm group with similar information for your project pupils. Probably this information is sparse, and you need be particular about this only if your experimental design is type 3 (Comparing gains with local, state, or national norms. See page17)
4. Appropriate score scale must be available for comparing pre-test with post-test. To see if such a score scale is presented with a test, or to select the most applicable for your project, the following suggestions are given:
   a. If there are no conversion (standard or converted) scores to which the raw scores are transposed from each form of the test, or if the publisher states that raw scores from all forms may be used interchangeably, then use raw scores.
   b. If raw scores are equated or transposed to standard, converted or grade equivalent scores by separate tables for each form of the test, then use the standard, converted, or grade equivalent scores. In addition, notice how much of a change in the standard, converted, or grade equivalent score is caused by first a one-point change, and then by a change of one standard error of measurement (as given in the manual) in raw score. This will make you more cautious in the interpretation of both individual differences in a class, and grade-to-grade differences for individuals and classes.
   c. If percentile rank scores are used, you may state only the pre-test and post-test percentile ranks for each pupil; or (as stated earlier) for a class, the highest, median and lowest percentile ranks; or, for each quarter or fifth of the class as ranked on the pre-test, the highest, median, and lowest percentile rank on the post-test. (Statistically, percentile ranks are not equal-interval scales so that the usual mean and standard deviation types of statistics are not applicable, and gains in percentile rank are not comparable.)
   d. Check with your State Department of Education concerning the type of score scale and the form to use in reporting data from standardized tests. It is required to tabulate each pupil's score on a given test statewide in reporting to the U.S. Office of Education. Therefore, the same score scale should be used throughout the state.

LOCALLY-MADE TESTS

Locally-made objective or essay type tests are necessary when standardized tests are inadequate for reasons of content, difficulty, scope, or cultural bias. Common procedures involve listing objectives and expressing them in terms of pupil behavior changes (as seen in the Assessment Planning Chart, pages 4 and 5, making a two-way blueprint of learning outcome vs. content coverage, writing items, trial administration, and assignment of scoring weights to items and parts.

The most versatile form with respect to objectivity, type of learning outcome, and discrimination possible is the multiple-choice. Points to watch:

1. An item (question) should test one idea only.
2. Language should be simple, unless complexity of language is an objective.
3. Format should be clear. Responses should be at the end of the sentence, and should be brief. Grammar should be correct for all responses.
4. Possible responses should be homogeneous and should be equally plausible to the pupil.
5. There should be three, four, or five possible responses (choices) according to the grade level.
6. The correct answer should be evenly spread among the choices (a, b, c, d, e) for the entire test.

In essay type tests, most care is needed in the phrasing of the questions and in the objectivity (consistency) of the scoring. Points to watch:

1. Wording should be simple, so that there is greater likelihood that the pupil will answer *what* is desired as well as in the desired scope and context (answer will not be too brief or too vague).
2. Scoring should consist of adding weighted parts for each question: possible points are assigned beforehand for each fact, concept, procedural step, or part-answer, with the numbers of possible points weighted according to importance. Total points for each question and for the entire test can then be treated as numerical scores to be organized and interpreted.

Since test construction is a specialized undertaking, seek assistance from the best authority available.

RATING SCALES

Rating scales may be devised to attempt to measure performance, attitude, interest, character, or personality. A rating scale allows classification along a continuum of either frequency of occurrence (always, usually, occasionally, never) or intensity (strongly agree, mildly agree, undecided, mildly disagree, strongly disagree) of reactions or behaviors. There is also a "person-to-person" (ranking) rating scale method. *In rating scales, the person doing the rating is the measuring instrument: the scale merely systematizes this human measuring.* Some brief pointers on the construction and use of rating scales are:

1. A rating scale should preferably have 5 to 9 equidistant rating points, identified by numbers usually arranged with the highest end or most desirable point the highest number. Ordinarily several scales are used in a cluster, and if clusters are to be summed or averaged this numbering must be consistent for proper weighting of the sum or average. If one scale is deemed less important than others in a cluster, its highest point can be set lower than the others; or, for consistency by the rater, all scales can be the same pointwise, but each would be multiplied by a different weight before the adding or averaging.

2. A scale must rate the same characteristic at all points.
3. Points on a scale should correspond to actual observable differences between pupils to be rated.
4. A scale may be either *descriptive*, with purely descriptive words at the points: (Value judgments are not included.)

Example:

Class participation

1	2	3	4	5
Almost never participates	Participates only with urging	Participates occasionally to about half the time	Participates most of the time	Always participates

or a scale may be *evaluative*, with value judgments implied by the words assigned to each point:

Example:

Class participation

1	2	3	4	5
Unsatisfactory, almost lacking	Poor, needs encouragement	Average	Good, above average	Enthusiastic, completely satisfactory

It is better to separate these two functions (descriptive and evaluative), as in anecdotal reports, so that the user of the results may make his own value judgments from the behaviors described.

5. Two practical arrangements of combining separate descriptive and evaluative scales are:
   a. Beside each descriptive scale insert a simple evaluative scale:

Example:

Class participation

Rate amount of participation below					Rate quality of participation below				
1	2	3	4	5	1	2	3	4	5
Almost never participates	Participates only with urging	Participates occasionally to about half the time	Participates most of the time	Always participates	Unsatisfactory	Barely satisfactory	Average	Above average	Very satisfactory

   b. After a cluster of descriptive scales insert one or two evaluative scales, in summary fashion:

Example:

Classroom behavior

1	2	3	4	5
Hardly ever interested or cooperative	Sometimes interested and cooperative	Average interest and cooperation	Usually interested and cooperative	Always interested and cooperative

Playground behavior

1	2	3	4	5
Either withdrawn or belligerent	Tends toward either withdrawal or anti-social	Average adaptation and participation	Usually dependable cooperative and/or resourceful	Resourceful and/or gets along in or leads wholesome activities

Overall behavior evaluation

1	2	3	4	5
Unsatisfactory, needs attention urgently	Requires improvement	Average	Above average	Very satisfactory

6. With each rating scale include a "familiarity indicator", where the rater can indicate how well qualified he feels he is, considering his own personal bias, degree of acquaintance with the pupil, or opportunity he has had to get an adequate and fair sampling of behavior, to allow satisfactory rating of that particular pupil.

Example:

(Beside each scale if possible; for each pupil at least)

AA Good opportunity for unbiased observation

AB Good opportunity but observation may be biased unduly

BA Fair opportunity for unbiased observation

BB Fair opportunity but observation may be biased unduly

CA Some good unbiased observation, but not enough for unconfirmed conclusions

CB Some good observation, not enough for unconfirmed conclusions, and probably biased

D Acquaintance or opportunity for observation not sufficient to make rating.

Reasons for bias are:__________

____

____

7. Agreement of raters. All (if possible) or a sample of raters should go over each scale together to attempt common interpretation of each point on each scale. This is particularly important if results from several raters are to be averaged in later summaries, or changes in ratings (pretest-posttest) are to be reported. Some raters avoid extremes; others may make hasty decisions or be unduly influenced (biased) by prior information or opinions ("halo effect"). If there is considerable diversity of interpretation of descriptions on a scale, the description may have to be reworded.

CHECKLIST

A checklist is a list of subjects or statements to which only two responses are possible.

1. Possible responses:
   (a) check or don't check
   (b) like or dislike
   (c) agree or disagree
2. Note that intensity of feeling is omitted. If intensity is desirable, expand the checklist into a rating scale.

Examples:

Checklist items:(agree or disagree)

______1. I think coming to school is fun.

______2. I seldom lose interest in this class.

______3. I have learned a lot this year.

______4. This class has made me a better reader.

Rating scale items:

1	2	3	4	5
Strongly Dislike	Mildly Dislike	Doesn't Matter	Mildly Like	Strongly Like

______1. Having to go to school in the summer.

______2. Going to school after regular hours.

______3. Reading a book at home.

______4. The way my teacher teaches me.

3. The descriptive checklist: a special type of checklist. The checking is usually done by the teacher, although it may be a pupil report device. Each description that applies is checked. (Examples below.)

Pupil Characteristics

______Cannot unlock words

______Speech interferes with reading

______Needs glasses

______Has weak phonics background

Program Characteristics

______Needs more visual aids

______Was planned with unclear goals

______Needs specialized personnel

______Lacks effective evaluation

4. Variations in the checklist. Sometimes more than two responses are possible in the nature of frequency-of-occurrence categories.

Example: never seldom often always (responses)

______ Look forward to coming to school

______ Feel the teacher is trying to help me

______See the use of what I am being taught

QUESTIONNAIRE

The questionnaire is a series of questions, usually of the "yes"–"no" variety, that is filled out by the pupil. Occasionally, numerical answers are requested, such as "How many books did you read last year?" Questionnaires must be carefully worked out, checked, tried out, and revised in order to avoid:

1. Vocabulary that is too difficult or otherwise unfamiliar.
2. Emotional overtones, or emotionally-toned experiences (unless this is what you are trying to measure indirectly). These overtones may distort the factual information you are seeking.

Example:

(Good) Has your teacher been able to help you when you did not understand what you were to do?

(Poor) Would you want to have a teacher who did not divide her attention equally among the whole class?

3. Social stereotypes, or descriptions that the pupil will either choose or avoid because they meet with or conflict with the general approval of society. Avoid topics of sex, religion, and masculinity-femininity on this account. A pupil will rarely report himself as "bad".
4. Other generally "visible" items including those wherein the pupil will likely choose the response he thinks the teacher "wants".

Example:

I try to do my homework on time every day.

5. Negative statements. Word all statements positively, so that the pupil does not have to review the logic of the statement each time to decide whether his agreement should be recorded "yes" or "no".

ANECDOTAL REPORT

Anecdotal report is the systematic writing down of observations of pupil behavior that cannot be measured or classified by more formal tests, rating scales or other devices. Many samples of the behavior of a pupil are required for adequate assessment of social interaction, character, attitudes, interests, motivation, and aspiration level by this method. Record each observation of a pupil and your best interpretations of that observation in separate parallel columns. Avoid over-interpretation. The separate columns are both a check on your interpretation (separating "facts" from projections,) and a help to other interpreters in making their own conclusions.

Example:

Date	**Setting**	**Actual Behavior Observed (what he did)**	**Implications of the Behavior (evaluation)**
10/66	Classroom	Johnny was willing to read before the class for the first time.	Johnny feels more sure of himself
10/66	Lunchroom	Mary cleaned up the entire table at which she had been eating, even though some of the debris was not hers.	Mary is beginning to develop some awareness of cooperation
10/66	Schoolyard (recess)	Fred admitted he had been tagged in this morning's game of tag football, even though the score was close and he was tagged so lightly that he easily could have disputed the play.	Fred is showing evidence of fair play.

PRODUCT ANALYSIS

Product analysis is the inspecting and rating of products which pupils have made in order to measure the behavior that went into making the product. A careful list should be made of the actual behavior presumed to have contributed to the product, so that the assumptions made are justified. Products for analysis may be writing samples (for penmanship), arithmetic samples or essays (parts of classroom tests), drawings, reports, projects or objects produced in industrial arts classes.

Possible general methods to use in product analysis are:

1. Comparison of the product with a standard:

   Examples: Handwriting scale
   Teacher's sample "perfect" essay paper, with weighted parts, as described under *locally made tests* page 9. (Sometimes called "factor counting with weights".)

2. Checklist—this is sometimes called "factor counting":

   Example: Reading a passage orally in class.

   Checklist of evident qualities

   ______reads at an even pace
   ______speaks clearly and loudly
   ______is able to deal with "new" words
   ______pauses in proper places

3. Factor rating. This method might involve a mark (A,B,C,D,F) on each of several factors sought:

   Example: Drawing (blueprint type) of a tool box for industrial arts

   Factor rating (mark each line)

Dimensions	B
Lines	A
Lettering	C
Neatness	C
etc.	

4. Overall rating. This would apply to an art class drawing or an English class composition. The teacher has in mind the characteristics of a product worthy of each mark, and matches them mentally. Applied to an essay-type test question, this is considered inferior (less accurate) than the factor counting with weights described above.

5. Overall ranking. This method might apply to the same types of product as overall rating (4). The papers from the entire class are arranged in order of merit from best to worst, keeping in mind the qualities sought. Since this "keeping in mind" is not organized in orderly fashion, the method is not highly regarded as a measurement procedure. The resulting "score" is a rough rank order number.

STEP III

EXPERIMENTAL DESIGN

(Step III must be completed before the pupils begin their participation.)

In this section: **Page**

EXPERIMENTAL DESIGN

1. Select the pupils for the project. Use the following means:
   a. Recommendation of teacher. Base selection on carefully worked-out criteria or rating scales. The criteria are best developed as a cooperative effort by the teachers who will select the students. If this cannot be done, at least have all teachers review the criteria together so that baselines and interpretations will be mutually understood.
   b. Rating scales. Select the students on the basis of some criteria such as: class marks, socio-economic status, cultural rating scales, or other data obtained from rating scales.
   c. Standardized test results. The use of standardized tests is usually best understood by specialists in educational research. If a standardized test is used for selection of pupils for the project, do not use the same form of the test as a pre-test when the pupils become actually involved in the project. A second test must be used. If further information regarding the use of standardized tests is needed, consult a research specialist.
   d. Locally constructed tests. These should be used if the local situation is such that standardized tests are not available.

2. Keep a record of as many pupil characteristics as may be appropriate; i.e., sex, ability level, reading level, socio-economic classifications, etc. You may wish to show scores on rating scales or achievement tests by these classifications; that is, you may want to compare the performance of the most and least capable students, or the boys and the girls.

3. Establish how the behavior of the pupils at the beginning of a project will be compared with their behavior at the end of the project. This will assure that necessary comparative behavior will be available at the end of the project.

4. Establish the type of experimental design. The U.S. Office of Education recommends the following five types:
   a. Experimental group vs. control group. (Test the project group; test the control group; compare the results.)
   b. Pre-test—post-test gain vs. expected gain. (Predict the expected gain of the project group; test the project group at the beginning of the project; test the project group at the end of the project; compare the difference between the two tests with the original predication based on publisher's norms.)
   c. Pre-test—post-test gain vs. local, state, or national norms. (Test the project group at the beginning of the project; test the project group at the end of the project; compare the difference in the two scores with the expected difference according to already established local, state, or national norms.)
   d. Pre-test—post-test gain vs. post-test projected from last year's class. (Test the project group at the beginning of the project; test the project group at the end of the project; compare the difference in scores with the expected difference as projected from last year's class.)
   e. Post-test or gain with no other basis for comparison.

STEP IV

PUPILS ARRIVE

In this section:

STEPS TO BE TAKEN AFTER PUPIL PARTICIPATION HAS BEGUN

1. At the beginning of the project, pre-test. Use as many of the instruments as possible that will be used in the post-test.
2. During the project, use whatever in-progress evaluation devices such as daily checklists, anecdotal reports, etc., as you have decided upon.
3. At the end of the project, post-test.

STEP V

ANALYSIS OF PROJECT DATA

(Step V is performed after pupil participation is over.)

In this section:

ANALYSIS OF PROJECT DATA

After pupil participation is over, the evaluator is faced with the problem of analyzing the data he has collected. This is done by comparing the data collected at the beginning of the project (pre-tests) with the data collected at the end of the project (post-tests). Four distinct procedures are required in order to complete this work, the last three of which require moderate acquaintance with statistical procedures. These last three procedures in data analysis have been reserved for the next section in the *Guide,* "Only for Use by Those with Some Knowledge of Statistics". Step 1, however, since it can be performed by most experienced teachers, is given here, separate from the statistical section.

Summarizing the data:

1. Test scores (raw, standard, grade equivalents) may be averaged and changes (gains or losses) may be averaged, but percentile ranks, by their nature, may not be averaged. You may find it convenient to give the highest, middle, and lowest percentile ranks of the students. A very meaningful way of describing group performance is to rank order the scores from highest to lowest. Then, divide the group into quartiles (four equal parts, more or less). Report the percentiles of the middle student in each of the four parts.
2. Rating scale results may be averaged, provided the scale was set up "equal interval" for each separate scale. In order to combine or average the results from more than one scale for a pupil or class, additional care should be exercised to see that extreme or intermediate ratings are uniformly desirable or undesirable for all scales.
3. Questionnaire results should be reported in number and percent of respondents.
4. Anecdotal reports and teachers' comments may be summarized into coarse categories, such as "little improvement", "slight improvement", "moderate improvement", "great improvement", or "very great improvement", giving the number and percent of the group in each category.
5. Pupil count should report number of pupils started, number finished, number improved, number no-change, and number regressed.
6. Over-all ratings, for inclusion in State Department of Education reports to the U.S. Office of Education, require for the Primary Objective, and Objective 2, etc., summarization or classification of a project as: "Little or No Progress Achieved", "Some Progress Achieved", or "Substantial Progress Achieved".

STEP VI

FOR USE BY THOSE WITH SOME KNOWLEDGE OF STATISTICS

(Step VI is performed after pupil participation is over.)

In this section:

1. **How to determine whether the gains made by pupils on tests and rating scales are statistically significant.**

To be considered statistically significant, any improvement or "regression" from pretest to posttest in the score of a pupil or the average score of a class must be shown to be greater than that expected purely by chance due to measurement error (standard error of measurement) of the test. The levels of chance commonly acceptable are 15, 5, or 1 time out of 100. For example, you would want to know whether a gain of 4.6 points would have occurred by chance as often as 15, 5, or 1 percent of the time.

For each level used (such as Grade 4, arithmetic reasoning) of a given test, the amounts of change may be estimated for each of these three levels of significance, and the change of each pupil or class compared with them and proclaimed as statistically significant at that level if it exceeds them, or as not statistically significant if it does not exceed them.

To estimate these minimum amounts of change required to be significant, it is necessary to compute the standard error of measurement of a difference between scores of the same pupil or class. The formulas are:

Standard error of measurement of a difference in one pupil's pretest-posttest scores on two forms of the same test $=\sqrt{2s^2_{meas.}}$

and

Standard error of measurement of a difference between mean pretest-posttest averages for a class of pupils on two forms of the same test $=\sqrt{\frac{2s^2_{meas.}}{N}}$

where

N is the number of pupils in the class

$s_{meas.}$ is the standard error of measurement (standard error of a score) of the test. (How to find it is given on page 24)

The standard error of measurement of a difference is multiplied by the appropriate multiplier (ordinate on a normal curve): 1.44 for the 15% level, 1.96 for the 5% level, and 2.58 for the 1% level, and the gains obtained in the project compared with it.

Example: Given N 20 and $s_{meas.} = 2.5$ Gain of Pupil C 6 points
Gain of Class 2.15 points

For pupil, standard error equals

$$\sqrt{(2)(2.5)(2.5)}$$

$$= 3.54$$

For class, standard error equals

$$\sqrt{\frac{(2)(2.5)(2.5)}{20}}$$

$$= .791$$

15% level:	(3.54) (1.44) = 5.10	(.791) (1.44) = 1.14
5% level:	(3.54) (1.96) = 6.94	(.791) (1.96) = 1.64
1% level:	(3.54) (2.58) = 9.13	(.791) (2.58) = 2.04

The conclusions are that for any pupil in the class gaining or regressing 6 points the change is statistically significant at the 15% level; for any pupil gaining or regressing 7 to 9 points, the change is statistically significant at the 5% level; and for any pupil gaining or regressing 10 or more points, the change is statistically significant at least at the 1% level. Similarly, for any class with an average change of 1.14 to 1.63 points, the change is statistically significant at the 15% level; with 1.64 to 2.03 points, at the 5% level; and with 2.04 or more, at the 1% level. Pupil C "makes it" at the 15% level, and the class at the 1% level.

To find the standard error of measurement of a test ($s_{meas.}$)

a. For standardized tests: The standard error of measurement is given, usually, for each grade level, for each test or subtest in the manual accompanying the test in the section or table labeled "reliability". Sometimes this section, in recent tests, is in a separate "technical report". Be sure the scale unit is the same as the one you are using: raw score, grade equivalent, standard score, etc. If not, you must convert it to match.

b. For locally made tests: The standard error of measurement can be estimated from the following formula.

$$s_{meas.} = SD\sqrt{1 - r_{rel.}}$$

where SD is the standard deviation of actual scores by the class for one testing session
$r_{rel.}$ is the coefficient of reliability of the test

To find the coefficient of reliability ($r_{rel.}$) of a locally made test:

A test's reliability may be conveniently estimated by correlating the two sets of scores for a group from a test-retest administration of the same form of test, using the Pearson Product-Moment Coefficient of Correlation, with no more than a day or two between test and retest.

Alternatively, if speed is not an important factor (each pupil actually attempts each question), either the split-half, corrected for length, or a Kuder-Richardson formula may be used. In the split-half, each pupil's paper from one administration of the test is scored in two parts: the odd-numbered items, and the even-numbered items. These two sets of "half-scores" are correlated by means of the Pearson Product-Moment Coefficient; and the result corrected from half-length by means of the Spearman-Brown Formula:

$$r_{rel.} = \frac{2r}{14r}$$ where r is the correlation between test halves.

Probably the most convenient Kuder-Richardson formula is number 21:

$$r_{rel.} = \frac{K}{K-1}\left(1 - \frac{M(K-M)}{K(SD)^2}\right)$$

where K is the number of items in the test
M is the arithmetic mean (average score)
SD is the standard deviation of these scores from one administration of the test to a class.

To find the reliability coefficient of a rating scale, probably the simplest method is the test-retest, as above, for each separate scale used. Possibly (with statistical consultation) the Kuder-Richardson formula given above could be used, with K = the largest number of points possible on that scale.

2. **Are the changes in questionnaire responses statistically significant?**

For example: if 15 of a class of 31 pupils responded "yes" or checked a statement in the pretest, and 21 on the posttest, is the change statistically significant? To compare this change with the three magnitiudes necessary to be significant at the 15%, 5% and 1% levels (as outlined above) a "two-by-two" table is set up:

	Posttest		
Pretest	not checked	checked	total
checked	a	b	a + b
not checked	c	d	c + d
Total	a + c	b + d	a + b + c + d

Example:

5	10	15
5	11	16
10	21	31

The number to find (called the "Critical Ratio" in this case) is found from the formula:

$$\text{critical ratio} = \frac{d - a}{\sqrt{a + d}} \qquad \text{Example: } \frac{11 - 5}{\sqrt{5 + 11}} = \frac{6}{4} = +1.5$$

Using the same critical numbers as before:

1.44 for the 15% level, 1.96 for the 5% level, and 2.58 for the 1% level, compare your result with these critical numbers to see at which, if any, level your results are significant beyond chance. Example: Since 1.5 is not larger than 1.96, but is larger than 1.44, the change is significant at the 15% level for this class.

3. **If there is no significant change between pretest and posttest results, of what value has the project been?**

As scientific research is not judged fruitless when results are zero or "negative", neither should unsuccessful projects be considered useless. Favorable results are not necessities. It is worthwhile to know what brings no or negative results in a certain educational community. This information, when combined with results from other communities, leads to improved education on a nationwide basis.

GLOSSARY

Anecdotal Report—An organized written record of an observed incident of pupil behavior, whether in classroom, on the playground, or elsewhere.

Aspiration Level—Personal or scholastic goals of the pupil

Behavioral Change—What the pupil does after the learning situation compared with what he did before. This should be analyzed in minute detail, so that as much of the change as possible will be noted in the evaluation.

Checklist—A list of words, phrases, or statements to be checked or not checked by the observer (sometimes by the pupil) according to directions.

Converted Score—A type of standard score

Culture-free test—An ideal test (not yet compl tely achieved) wherein the symbols are recognized and understood equally well by pupils of any culture.

Equal-Interval Scale—A test score scale where the units are of the same width, as in a thermometer. Usual test scores are treated as if they were on this type of scale, so that scores and gains may be averaged and otherwise compared.

Evaluation—Determination of the worthwhileness of instruction or other factors on the pupil, either by direct observation or with the aid of rating scales, tests or other measurement instruments.

Homogeneous Responses—Multiple-choice test choices that are roughly equal in the eyes of the pupil with regard to grammar, logic, and general plausibility. A distractor (incorrect choice) that is too different will not be chosen and should be replaced with a better one.

Instrument—An all-inclusive term for test, rating scale, checklist, questionnaire, anecdotal report, sociogram, or other device used in measurement and evaluation.

Item—A "question" on an objective test, or a word, phrase, or statement on a checklist. Used instead of "question" because it may not be stated in question form.

Learning Outcome—A classification of the phenomena or products of learning that is of concern in composing tests and other instruments used in evaluating scholastic achievement. Learning outcomes are generally sorted and identified according to complexity and type of behavioral response by the pupil. *Factual memorizations* (often called knowledge) are thought to be less complex than *concept acquisitions,* which are less complex than deeper *meanings* and *understandings.* There is not complete agreement on the classification of learning outcomes, but in *Taxonomies of Educational Objectives* (Bloom, and others) a comprehensive start has been made in both "cognitive" and "affective" areas (affective domain includes interests and attitudes).

Locally-constructed Test—A test made by local teachers or specialists rather than by a commercial publisher. It may have local norms.

Mean—Short for the Arithmetic Mean. A numerical average that may be applied to most test scores (scores that are equal-interval rather than rank orders or percentile ranks). The scores are added and then divided by the number of scores.

Measurement—Applying a number scale to the object to be evaluated, as is done with tests, rating scales, and checklists, as an aid in evaluation.

Median—Point on a score scale below which are half the scores of the group and above which are the other half. Identical with the 50th percentile.

Non-verbal—The content of a test includes no words, word symbols, or pictures to which usual word names have been associated. The directions may be written or spoken, however.

Norms—Tables of test scores achieved by a specified group of pupils (the norm group, or norm sample). These scores are translated into grade equivalents, percentile ranks, IQ, etc. in the process of standardization of the test. It is important to know as much as possible about the other characteristics (such as ability level and socio-economic environment) of the norm group pupils, when making interpretations of your pupils' scores.

Objectives—Aim or purpose of a test or project. Preferably stated in terms of the behavioral changes expected in the pupils.

Objectivity—Characteristics of a test permitting the same score for a pupil regardless of who does the scoring. Opposite of *subjectivity.*

Open-ended Sentences—A projective method of evaluating personality and aspiration level. For example: "The trouble with the army is________________." Should be devised and interpreted by a psychologist.

Percentile Rank—Often called "centile rank". An interpretive (derived) test score indicating the percent of the group scoring *below* the pupil receiving the score. (In case several pupils receive the same score, half are counted below when computing the percent, and all receive the same percentile rank). Since they are not equal-interval scale, percentile ranks cannot be averaged by the arithmetic mean.

Quartile—First quartile is the point on a score scale below which are one-quarter of the scores of the group. Third-quartile is the point below which are three-quarters of the scores.

Questionnaire—A list of questions either requiring "yes"–"no" response or a few words or numerical information. It is analyzed by computing the percentage of pupil responding "yes", or by tabulating the other types of response.

Rank-order scale—Papers or scores arranged in order, with the best or highest labeled "1", the next "2", etc. In case of tie for third place, for example, rank both "3½" and the next "5". As with percentile ranks, ranks are not equal-interval and cannot be averaged using the arithmetic mean.

Reliability—The accuracy or consistency of a test. If retested, how close would the score be? However accurate a test measures *whatever it does measure,* it may be measuring the wrong thing, and therefore still not be valid for your purposes for the pupils you are testing. The *reliability coefficient* is the Pearson Product-Moment Correlation between alternate forms or retestings, or halves of the test (corrected for length).

Scale (score)—A numerical or category system for classifying the results of measurement. Test score scales use raw scores (number right, sometimes corrected for guessing), standard scores, converted scores, grade equivalent scores, percentile ranks, IQs, stanines, etc.

Scale (rating)—A continuum (line) with usually equally spaced points which are numbered and accompained by word descriptions or evaluations, and used in rating pupil behavior such as personality characteristics.

Significance Level—The percent of the time a change in a pupil's score would occur by chance due to the test's measurement error. Its use in selecting experimental samples from populations (inferential statistics) is of no concern here. Example: A gain or loss as great as 6 points by a pupil on a test would occur by chance error only 4 times out of 100. His change is said to be significant at the 4% level of confidence. Instead of estimating the level for each pupil, we compute the change necessary for only three levels: the 15%, 5%, and 1%. The smaller the percent, the greater the significance.

Skill—A learning outcome characterized by a combination of mental and physical operations, such as typing or map reading, that shows improvement with practice. A skill is ranked in mental difficulty along with memorization of facts rather than with understanding or problem solving.

Sociogram—A diagram for interpreting social relationships in a class. Pupils write first and second choices of friends they would like to sit next to, and the diagram consists of small circles or triangles (one for each pupil) located on a target-like map, with the most popular pupils near the center, and connecting arrows representing choices.

Standard Deviation—A statistical measure of the variability, or spread of the scores of a class. In these projects, where the entire class is described, and projections are not made for a larger population, the simpler formula is used.

$$SD = \sqrt{\frac{\Sigma X^2}{N} - M^2}$$

where ΣX^2 is the sum of each score squared
M is the arithmetic mean of the class
N is the number of pupils in the class

ERRATA

Page 23 -

Standard error of measurement of a difference between mean pretest-posttest averages for a class of pupils on two forms of the same test

$$= \sqrt{\frac{2s^2 \text{ meas.}}{N}}$$

Page 24 - The Spearman-Brown Formula should read:

$r \text{ rel.} = \frac{2\,r}{1 + r}$ where r is the correlation between test halves.