DOCUMENT RESUME

ED 086 727 TM 003 378

TITLE Evaluation Guidelines.
INSTITUTION Nevada State Dept. of Education, Carson City.
PUB DATE Jul 73
NOTE 75p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Educational Accountability; Educational Assessment; *Evaluation Techniques; *Guidelines; Manuals; *Models; Performance Specifications; Program Design; *Program Evaluation; Student Evaluation

ABSTRACT

This practical reference manual is intended for use in the evaluation of learner performance. The guidelines are presented in such a manner so as to make the evaluation process accessible and understandable to everybody involved in educational instruction and administration. A glossary of terminology is included to facilitate this process. The text follows the progression in an evaluation model. The model is designed to depict the necessary steps in evaluation in their proper time sequence. The major steps of the evaluation model include Needs Assessment, Design Program Evaluation, Implementation of Program Evaluation, Measurement of Objective Attainment, and Reporting and Recycling. (NE)

NEVADA

EVALUATION GUIDELINES

July, 1973

DEPARTMENT OF EDUCATION · STATE OF NEVADA

KENNETH H. HANSEN
Superintendent of Public Instruction

STATE BOARD OF EDUCATION

MEMBERS

DIVISION OF PLANNING AND EVALUATION

James Kiley, Associate Superintendent

Dr. Kay W. Palmer, EMIS Director

Dr. R. H. Mathers, Consultant, Assessment & Evaluation

Dr. Austin Haddock, Consultant, Planning

NEVADA STATE DEPARTMENT OF EDUCATION

Carson City, Nevada

Table of Contents

Title	Page
Foreword	i
Evaluation Model	ii
1. Needs Assessment	1
description of variables	
performance objectives	
item pools	
measurement of learner status	
report of learner status	
2. Program Evaluation Design	6
program evaluation questions	
performance objective review	
comparison of variables	
sampling procedures	
instrument selection	
measurement controls	
statistical techniques	
report format	
monitoring system	
calendar of events	
3. Program Evaluation Implementation	44
monitoring instructional, institutional variables	
interim measurements	
revision of evaluation design	
revision of calendar of events	
4. Measurement of Objective Attainment	47
collecting postprogram data	
analyzing data	
comparisons with preprogram data	
comparing attainment with performance objectives	
writing recommendations for use of measurement data	

Page

5. Reporting and Recycling 50

reporting postmeasurement
reporting pre-post comparative data
reporting analysis of performance objective attainment
revising variables and sample procedures
revising objectives
revising measurement controls and instruments
recommending program changes
writing recommendations for learners not attaining objectives

6. Glossary of Terms 55

7. Appendix 61

FOREWORD

In this time when "accountability" is the watchword, it is appropriate that we consider evaluation and its role in accountability. If accountability means an accounting by the educational system to its product, the learner, then evaluation assumes a significant role in the process, since any educational accountability requires two components:

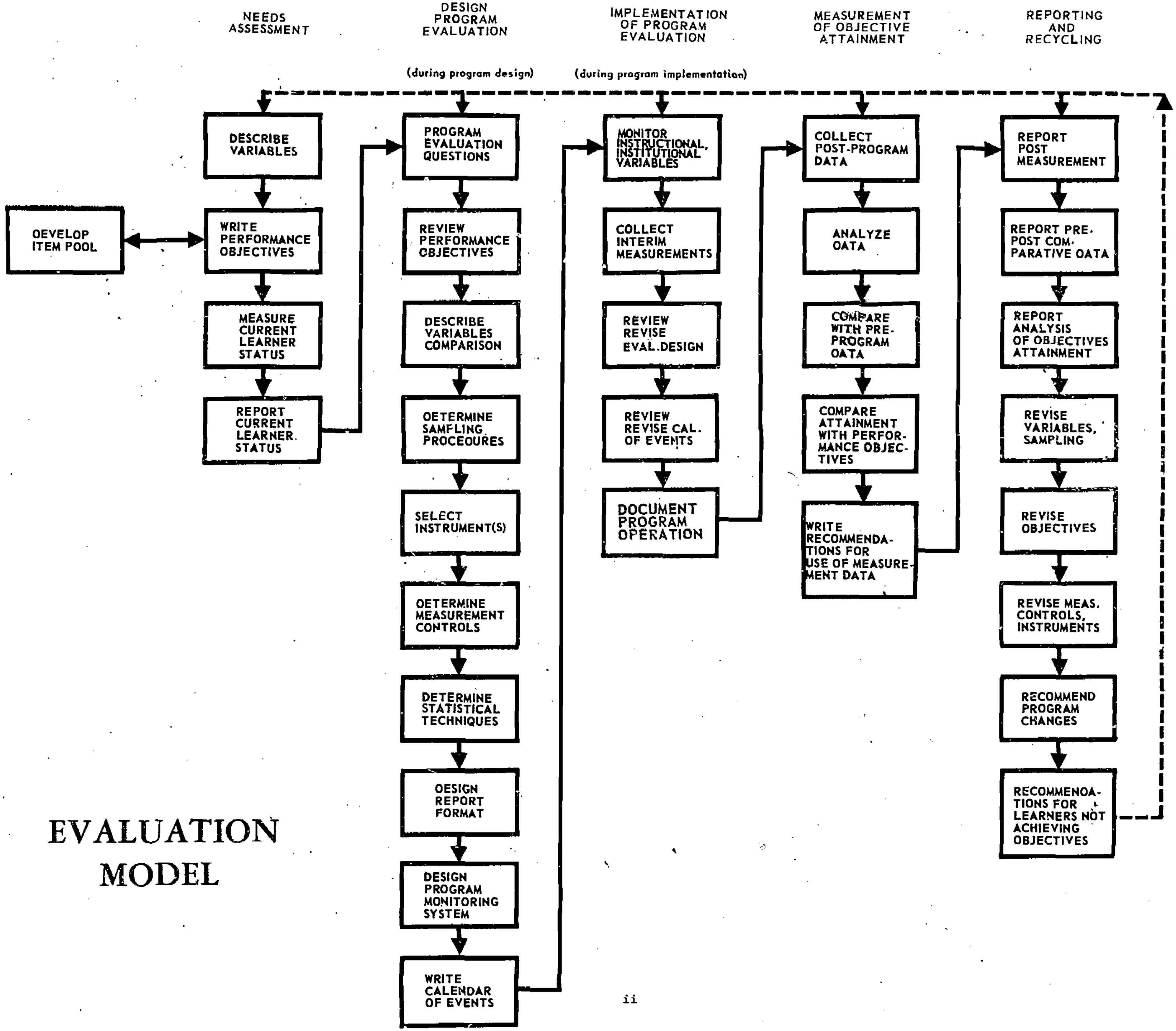
1. a precise definition of the objectives of the educational undertaking, and
2. a method of measuring these objectives in order that judgments may be made which will alter the educational process.

Without either of these components, accountability becomes an empty and futile concept.

Evaluation is a process in which measurement and judgment are combined to make possible decisions which will change and improve education. The Guidelines presented here are intended to make the evaluation process accessible to everybody involved in educational instruction and administration. They are not intended as a scholarly tract, but rather as a practical reference manual to which the educator may turn for assistance in the evaluation of learner performance.

The text of the Guidelines follows the steps in the evaluation model which may be seen on the next page. The model was designed to depict the necessary steps in evaluation in the proper time sequence.

R. H. Mathers
Division of Planning and Evaluation



EVALUATION MODEL

I. NEEDS ASSESSMENT

In the following discussion of needs assessment the term is intended to mean only those aspects of assessment that relate directly to expectations of student performance and to the measurement of that performance. Other aspects of needs assessment, such as determining educational need as perceived by persons involved in the educational process, are not considered here. For a more thorough discussion of needs assessment, please see the work entitled Needs Assessment Guidelines, also published by the Nevada State Department of Education.

Description of Variables

The variables involved in any educational undertaking may be described under three broad headings: institutional, behavioral, and instructional. Institutional variables are the persons involved in the undertaking, such as students, teachers, or other members of the community. Behavioral variables may be thought of as cognitive, affective, and psychomotor. Instructional variables include such things as content, method, and cost. Definitions of these variables are included in the glossary of terms in the back of this work. For our purposes here, however, let us take a more complete look at the behavioral variables.

The cognitive variable has six levels:

1. Knowledge
2. Comprehension
3. Application
4. Analysis
5. Synthesis
6. Evaluation

Note that these levels are listed in order of complexity, comprehension being thought to be more complex that knowledge.

The affective variable has five levels:

1. Reception
2. Response
3. Valuation
4. Organization
5. Characterization

Again, these levels are listed in order of complexity.

The psychomotor variable has five levels:

1. Imitation
2. Manipulation
3. Precision
4. Articulation
5. Naturalization

These levels are also listed in order of complexity.

Detailed definitions of each of the levels listed above are stated in the Appendix. Please read them before continuing on to the next section on performance objectives.

Performance Objectives

The reason we needed to examine behavioral levels is because of their use in writing performance objectives. A performance objective is one which talks about desired changes in behavior by the learner. Such an objective contains six components:

1. the time required to attain the stated performance
2. the institutional variable (who is involved?)
3. the behavioral variable (knowledge, comprehension, etc.)
4. the instructional variable (subject area, content, etc.)
5. the level of proficiency to be attained in the performance
6. the method of measuring that attainment

For example, here is a cognitive performance objective in sentence form:

1. By May 15, 1973 (time)
2. third-grade pupils (institutional variable)
3. will increase their knowledge (behavioral variable)
4. of reading vocabulary (instructional variable)
5. by ten months (proficiency level)
6. as measured by the reading vocabulary section of the Comprehensive Tests of Basic Skills (the method of measuring that attainment).

An example of an affective performance objective might be:

1. By the end of the first semester
2. fourth-grade pupils
3. will respond positively
4. in their attitude toward school
5. as evidenced by a 20% increase in their total score
6. on the Self-Concept Index.

The only way effective evaluation can ever take place is when we know where the learner is (in regard to a given behavioral variable) and where he should be (the performance objective). This is in fact our definition of educational need--the difference between the learner's status and his expected performance.

Item Pools

Since standardized tests rarely, if ever, contain all the items necessary to measure classroom objectives, it is highly desirable to begin the formation of item pools. These are simply collections of items which test the attainment of a specific learner performance objective. For example, a teacher might have a performance objective for one or more learners such as:

1. By the end of October,
2. pupils in my class
3. will apply knowledge of multiplication,
4. by multiplying two-digit numbers larger than 11 by a one-digit number larger than 4,
5. with at least 90% accuracy
6. as measured by problems from the item pool.

The item pool to test that objective would contain items like: 12x5, 26x7, 38x9, etc. Preferably the pool would contain a dozen or more such items from which random selections could be made to test the attainment of this specific performance objective. A sizable pool gives assurance that the learner had not merely memorized one particular answer, but understood the principles involved in the operation. Such items are called criterion-referenced test items, since the test item refers to a specific criterion, namely the ability to perform the stated operation. In many content areas, banks of objectives are available, such as the IOX collections in reading K-3 and mathematics K-3, among many others.[1]

---

[1] Instructional Objectives Exchange, P.O. Box 24095, Los Angeles, California 90024

Measurement of Learner Status

One of the most important aspects of needs assessment is the measurement of learner status. Ultimate pupil performance has no meaning unless measured against a starting point. If we measure pupil performance in reading at the end of the third grade and find pupils are reading at a fifth-grade level, we still have not learned anything about the quality of the instructional program, or about the increase or decrease in individual performance. For this reason among others, we recommend measurement of learner status as closely as possible to the beginning of an instructional program and as close to the end of the program as possible. Once-a-year measurement has two major defects. It may or may not reflect learner status at the beginning of a program. If it does, then it cannot reflect learner status at the end of a program. The second defect is that once-a-year measurement does not permit as many comparisons (and hence more complete information) because of high student mobility in some geographic areas.

To be sure, once-a-year measurement costs less, in money and personnel, but that is scarcely a good reason for its existence. Again we see that some better solution may ultimately lie in criterion-referenced measurement, especially of individual classroom-level objectives, since this would enable us to check pupil performance at many points during a program.

Reporting Learner Status

It is important that learner status be measured by, or reported to, the teacher as ar the beginning of the program as possible. The teacher can then compare status with performance objectives and begin to make necessary program alterations. By the same token, school and district offices should have access to such measurement data, since it will enable them to make comparisons with school and district performance objectives respectively. School and district offices may thus often be able to spot potential problem areas in the instructional program before they occur.

II. DESIGNING PROGRAM EVALUATION

1. Program evaluation questions

In order to evaluate a program, a design for such evaluation should be developed before the program starts. The best way to do this, in our opinion, is to consider what questions you want the evaluation to answer. There is no point in evaluating anything unless it sheds light on questions not previously answered and provides new judgments.

There are of course a large number of evaluation questions which may be posed, depending on the program, but let us examine a few common ones. One thing we usually want to know is whether a particular program or treatment is more effective than what we have been doing. 1. Did this program result in higher performance (by the group or the individual) than would have been the case in the regular program? Others might be-- 2. Can this new program (or treatment) be generalized to other grade levels or subjects? 3. Were the results achieved more costly than in the regular program? 4. If so, was it worth it?

In order ultimately to answer such evaluation questions, a way to answer them must be conceived in the preprogram planning. For example, to be able to answer the first question (above) we might use an evaluation design which included a control group using the regular program. To answer the second question, the treatment would have to be applied simultaneously to other grade and/or subject levels, and so on. The point here is that evaluation designs must be conceived before the program, not after, in order to be of maximum usefulness.

2. Review performance objectives

This step in the evaluation design is necessary for two reasons. First, you may need to revise the performance objectives in the light of information obtained from the measurement of learner status. For example, you may find that the performance objectives you have written for the class as a whole are unrealistic (too high or too low a proficiency level, insufficient time allotted for a particular objective, etc.).

Second, it may happen that there is a teacher change in the program, or somebody becomes involved who is unfamiliar with the objectives. Any change of personnel involved should trigger a review of the performance objectives, in order that everyone involved may be aware of the objectives. This applies to learners, as well as to faculty and administration. Changes in instructional materials, equipment, class times, etc. should also occasion a review of performance objectives.

3. Describe variables comparison

A very clear pre-program concept should be developed regarding the types of evaluative data to be derived from the program. In addition to classroom test or survey data, the program may call for school and district data. Consideration should be given to the collection of ethnic and socioeconomic data, and any other data which might not otherwise be available. Kinds of analyses and reports of the data required should be made in this time period in order to determine the cost and availability of such analyses and reports. For example, do you want scores reported in the form of raw scores, percentile ranks, stanines, grade equivalents or some other form? Do you want class mean scores? If so, in what form?

At this time you should also consider the kinds of comparisons you wish to make of evaluative data from the program. Do you wish to compare class data with school, district, state or national data? If so, is such information available? Where, and at what cost? There are many other kinds of comparisons you might wish to make, both for performance and for diagnostic reasons. All of them should be designated well in advance of the program and their availability and cost determined.

Another factor in determining what kinds of comparisons of data should be made is that of ease of interpretation. Neither raw nor treated data should ever be presented without adequate explanation of the significance of such data. Conversely, whoever uses evaluative data should familiarize himself with the precise meanings of the data presented. For example, if a mean grade equivalent score of 4.0 is reported for a new third-grade class, what does that mean, and what is the significance of such a score for such a group? It is very important that one knows precisely what a "standard deviation" is, or what a "correlation coefficient" is when one talks or reads about them. What is "standard" about a "standard score", for example? Definitions given in the glossary of items in the back of the Guidelines will help to refresh your knowledge of many measurement and evaluation terms.

4. Determine sampling procedures

In selecting samples (of learners or whatever) in a program, it is important to remember that samples are almost never perfectly representative of the population from which they are drawn. Where feasible, whole populations should always be included in any evaluative study, but often this is not possible. A school district might wish to collect data about the proficiency level of its fifth-grade pupils in arithmetic computation. If there are 50,000 fifth-grade pupils in the district this might not be feasible because of cost. Therefore a suitable sample of, say, 10,000 might be selected to give the district a reasonably accurate picture of the proficiency level of these pupils. The picture will not be exact, but if the sample is suitably drawn, the results will almost always be close enough to be of value.

The key word in sampling is that the sample should be representative of the population from which it is drawn. We might accomplish this reasonably well by systematic sampling. That is, we might select every fifth student in the third-grade classes of a district in order to obtain a sample representative of third-graders in that district.

Random sampling is sampling in which every person or thing to be selected has an equal chance of being selected. As you can see, selecting every fifth person, as in the example above, is not random sampling, because not every student had a chance to be selected. Probably the best way to accomplish random sampling is through the use of a table of random numbers. Through the use of such a table the educator eliminates any systematic, built-in bias in the sample selection. A table of random numbers is included in the Appendix, together with an example of its use.

Stratified sampling means that the sample is chosen from subgroups within the total population. For example, a testing program might wish to obtain data by sex or different ethnic groups. Sample selection should then make sure that proportionate numbers of the sex or ethnic subgroups be included in the sample. While the total sample would not then be random, selection within the subgroups could be conducted on a randomized basis.

One other technique in sampling will be discussed here, since its use is increasing in the field of evaluation design. This is the technique known as matrix sampling. Matrix sampling is simply a way of estimating test scores or other data for groups of people. In addition to selecting the persons for inclusion in the sample, matrix sampling also selects items randomly. For example, suppose that a district wishes to determine the proficiency level of fifth-graders in reading vocabulary. The district could administer a reading vocabulary test of 40 items, say the CTBS, to the total group of 3000 fifth-graders, but of course this would be expensive and perhaps not feasible. With the matrix sampling techniques the district could select say 300 students and 10 of the 40 items in order to get estimates of the mean proficiency level. There would thus be 300 x 10 = 3000 examinee-by-item responses, whereas with the total group there would have been 3000 x 40 = 120,000 examinee-by-item responses. The technique obviously represents a great saving in time and money. Matrix sampling should probably be reserved for those situations where total population measurement would be infeasible. It should never be considered an across-the-board substitute for individual evaluation. In any kind of sampling, however, keep in mind that there will be a sampling error. The size of this error will depend on the particular sample selected. Where random samples are involved, the degree to which the mean of the sample is representative of the total population mean can be estimated by the standard error of the mean formula:

$$SE_m = \frac{\sigma}{\sqrt{N}}$$

SE_m = the standard error of the mean

N = the size of the sample

and σ = the standard deviation computed by the formula:

$$\sigma = \sqrt{\frac{\Sigma d^2}{(N - 1)}}$$

σ = standard deviation

Σ = the sum of

d^2 = the squared deviations from the mean

N-1 = the size of the sample minus 1

Standard deviation may also be computed from the formula:

$$\sqrt{\frac{\Sigma d^2}{N}}$$

but particularly for small samples (30 or less) the formula containing (N-1) should be used because it is more accurate.

Now, what do we do with this statistic (standard error of the mean) when we get it? What does it tell us? Let us look at an example.

Suppose we find, in testing a sample of 36 students in arithmetic operations, that the mean score is 70 and the standard deviation is 18.

Then $SE_m = \frac{18}{\sqrt{36}} = \frac{18}{6} = 3$

Now, since approximately 2/3 of the scores, in a normal distribution, lie within one standard devi tion of the mean, we can say that the chances are two out of three that our sample mean is within ± 3 of the total population mean (67-73). Furthermore, the chances are about 19 in 20 that the sample mean is within ± 6 of the total population mean (64-76). Thus the standard error of the mean gives us a fairly precise method of estimating how accurate our sample mean is when compared to the total population.

5. Select instruments

Quite often measurement instruments are selected at a district level, which may or may not permit the individual evaluator any latitude in the selection of such instruments. The evaluator should make his opinions known, however, concerning the value of such instruments in determining the attainment (or lack of it) of performance objectives.

There are several problems involved in the selection of appropriate measurement instruments. First of all, in the cognitive domain, we have usually had to resort to standardized tests for the major measurement events in a program. Standardized tests have their advantages and their disadvantages. They are relatively inexpensive, easy to administer, and, in the case of the better ones, have been standardized on carefully selected national norm groups. Usually, scoring services are also available at additional cost. But, on the minus side, standardized tests have a serious drawback in that they seldom, if ever, contain items which will test all of the classroom teacher's performance objectives. The reason for this is simple. The test publishers are forced to select a relatively small number of performance areas among the many hundreds existing in a given subject.

The best way to select a standardized test of cognitive achievement is to examine the test, item by item, comparing each item with your list of performance objectives. Select the test that affords an opportunity for testing the largest number of your objectives. There are, of course, other considerations in standardized test selection. For example, you should consult the publisher's examiner manual and technical manual to answer questions you should have about the standardization of the instrument. Was the norm group diverse in nature or was it a regional (and hence probably biased) group? Does the instrument have parallel forms for the same grade level? What is the test-retest reliability of the instrument? And so on. But most important is the comparison of test items and performance objectives. If the instrument does not test at least a majority of your performance objectives, it is of no value to you. It is thus not hard to see why the development of criterion-referenced tests is of critical importance to the teacher or evaluator.

Many collections of cognitive performance objectives and related test items are now available. It would be worth your time to look at these and see if they might be useful in your program. A list of the collections is included in the Appendix.

The situation in the affective and psychomotor domains is perhaps even more distressing. Affective surveys should usually be administered by persons not otherwise connected with the learners involved, because of the emotional components of such surveys. In other words, the learner is more likely to answer accurately a question such as "Does your teacher yell at you?" if that question is put to him by someone other than the teacher. If possible, assurances should be given the pupil that teachers and administrators will not see individual item responses. Another negative in affective surveys is the fact that few reliability studies have been made for such instruments. On the plus side, however, is the fact that there are now available many collections of both affective objectives and corresponding survey items. Many of these collections have been refined for use at various grade levels and can be very useful in assessing learner attitudes.

In the psychomotor domain useful instruments are scarce, and for some grade levels unavailable. Furthermore, the relationships between psychomotor skills and cognitive skills have not yet been well researched. The same is true of the psychomotor and affective relationships. Only in the cognitive-affective relationships is there sizable research and it is spotty. In many areas we are still not able to answer such questions as "Does a positive attitude toward the subject correlate positively with cognitive achievement?" We feel, however, that such a lack of interrelated research in the three domains is not due to a lack of interest by researchers, but is due simply to the massiveness of the problems involved. This is another reason why local efforts in the writing of performance objectives in the three domains, and in developing the corresponding criterion-referenced instruments, is of such major importance.

While pretests and posttests have characteristically made use of standardized instruments, interim measurement usually has not. By "interim" we mean the measurement of performance objectives which takes place during the instructional program. Most interim measurement takes place because of the teacher's desire to measure instructional effectiveness while there is still time to make changes. This is an area where criterion-referenced tests, designed by the teacher or evaluator, may be of greatest significance. Among other things, it gives the teacher a chance to devise test items which measure his specific performance objectives, rather than those constructed by somebody else. Thus in terms of teacher evaluation, the teacher himself can have some first-hand input into the evaluation system.

6. Determine measurement controls

Prior to any measurement situation, controls should be developed which will insure as little contamination of the results as possible. There are at least six control areas which should be taken into consideration:

1. history of the class
2. testing times and dates
3. practicing for the test
4. changes in measurement
5. differences among experimental and control groups
6. statistical regression

History of the class refers to events which took place in the class which might affect test scores. For example, a third-grade class might not have had sufficient exposure to arithmetic applications, a fact which might tend to lower scores in that part of the subject area.

Testing times and dates are important because they often affect scores. For example, the "summer lapse" in cognitive achievement has been noted by many observers. Testing late in the day, when fatigue becomes a factor, might lower scores. Time of day for pretests and posttests should be the same. Psychomotor and affective assessments may change considerably due to the time of administration.

Practicing for the test, in the sense used here, means simply the effect upon a test score caused by having taken the test before. If the administrations are far enough apart (six months or more) the practice effect is usually negligible. Furthermore, if the teacher uses actual test questions for review purposes, scores often are increased.

Changes in measurement instruments or observers often cause score changes. In addition to differences of content between two instruments, it is very difficult to obtain reliable comparative scores on measures which have been normed on different groups. In measurement involving observation, such as in oral reading tests, scores vary because of the perceptions of different observers. Where possible the same well-trained observer should be used on different administrations of the same measure. With teacher-made instruments, there is sometimes an inclination to change grading standards during the course of instruction. This is one more reason for writing precise performance objectives prior to instruction.

Differences among experimental and control groups sometimes produce misleading results. For example, an experimental class might consist of low ability students who did not achieve during the program as well as students in the control group. The reason might be the low ability factor rather than the design of the experimental program. If one intends to measure certain factors in the program treatment, the two groups should be made to match as closely as possible.

Statistical regression refers to the tendency of scores at the extremes of a distribution to move toward the mean upon retesting. This could lead the evaluator to misleading inferences about such scores.

To summarize--in the evaluation design everything possible should be done to insure uniform measurement conditions. Moreover, any extraneous factors which affect scores, such as statistical regression, should be taken into account in evaluation procedures (see Appendix).

7. Determine statistical techniques

This section of the guidelines is intended both as a basic review of certain statistical concepts and as a guide to the selection of appropriate statistical techniques in evaluation. The intent of the review is to provide an easily accessible place to find basic statistical definitions and formulas.

Normal Curve

Basic to an understanding of statistical techniques is the normal curve of probability or distribution. It is also called the bell-shaped or Gaussian curve. Most distributions of chance events in any area of life will exhibit a more or less bell-shaped curve if we plot the frequency of occurrence of each event. The following chart shows a distribution

Chart I

Chart 1

Frequency distribution of number of heads when eight coins are thrown.

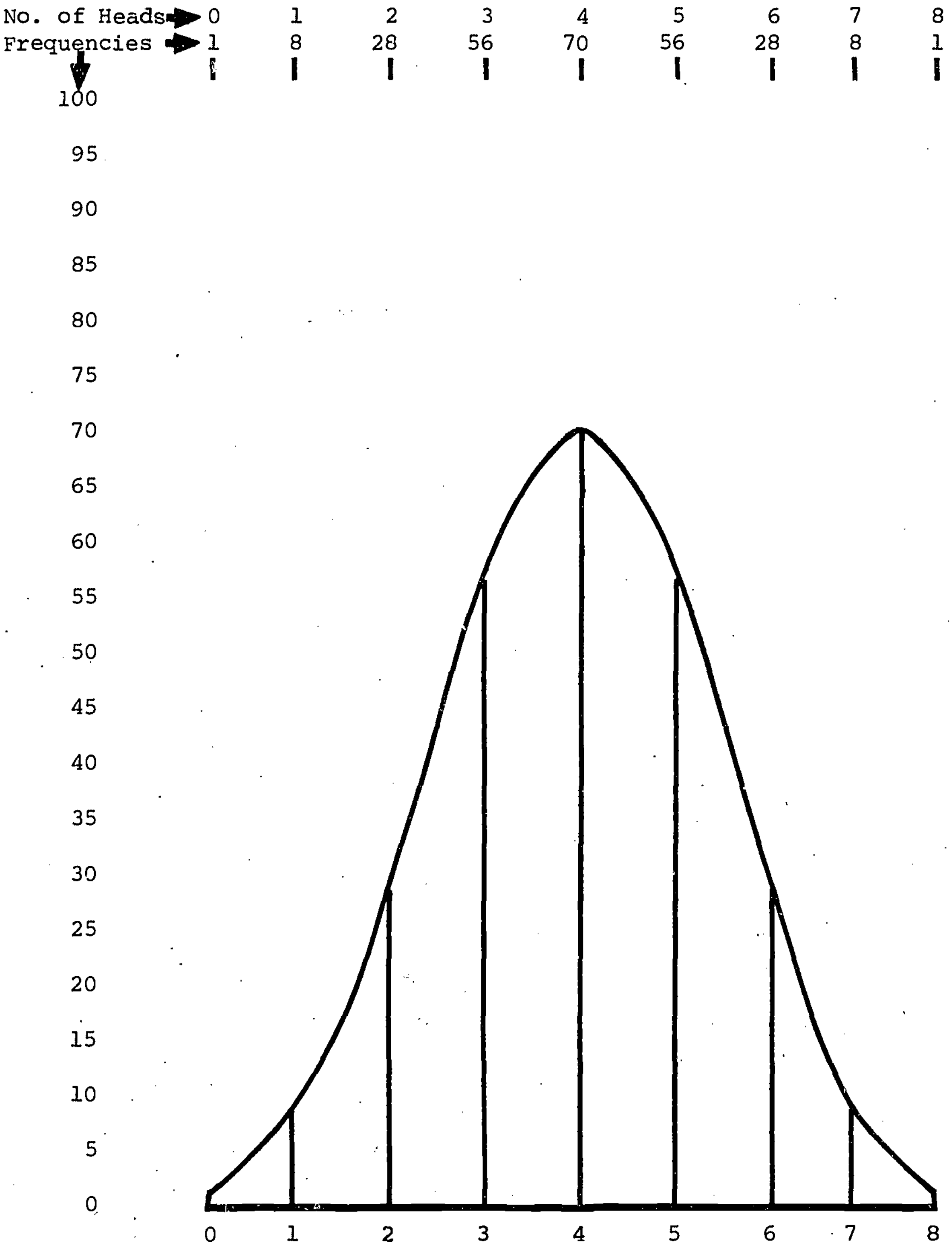
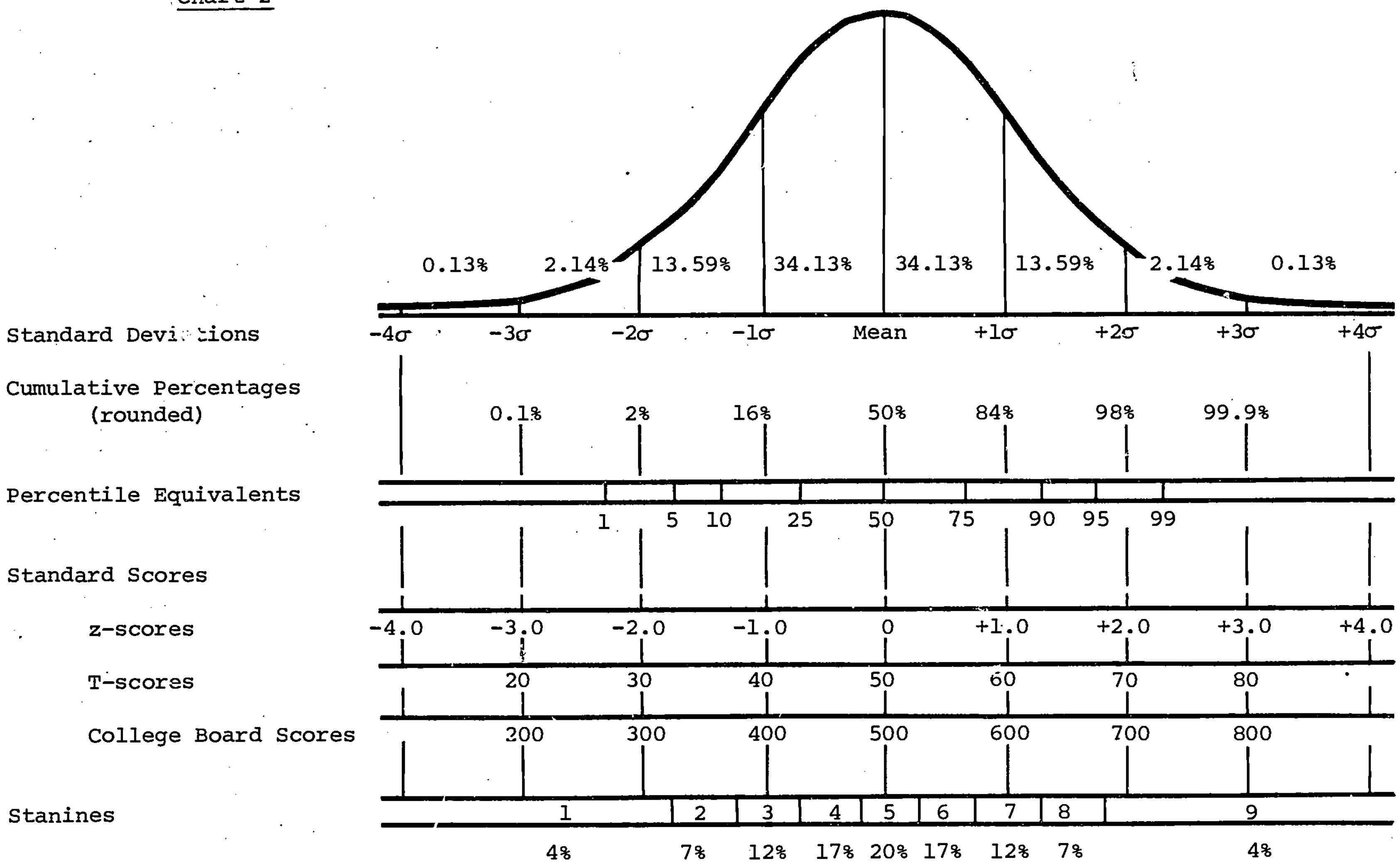


Chart 2


0.13%
2.14%
13.59%
34.13%
34.13%
13.59%
2.14%
0.13%
Standard Devi tions
-4σ -3σ -2σ -1σ Mean +1σ +2σ +3σ +4σ
Cumulative Percentages (rounded)
0.1% 2% 16% 50% 84% 98% 99.9%
Percentile Equivalents
1 5 10 25 50 75 90 95 99
Standard Scores
z-scores
-4.0 -3.0 -2.0 -1.0 0 +1.0 +2.0 +3.0 +4.0
T-scores
20 30 40 50 60 70 80
College Board Scores
200 300 400 500 600 700 800
Stanines
1 2 3 4 5 6 7 8 9
4% 7% 12% 17% 20% 17% 12% 7% 4%

of the frequencies of heads which can occur when eight coins are tossed. Out of a total of 256 such tosses, in a normal distribution there would be one time when no heads showed, one when all heads showed, eight times when one head showed, and so on. Note that when the frequencies are plotted on the chart, and the resulting points are connected, the chart assumes a bell-shape.

Chart 2

The normal curve is useful in many ways. For example, along the baseline (bottom horizontal line) we can measure the scale. Note the z scores, the T-scores, and the College Board scores, and their relationship to the mean and to the standard deviations. The midpoint of the baseline is the mean score and the percentage of cases under study is measured by the area between the curve and the baseline. Notice that the curve never reaches the baseline, although it gets closer and closer. At a distance of three standard deviations or sigmas on each side of the mean, the remaining area is only 1/10 of one percent on each side.

Standard Deviation

In the preceding paragraph we used the term "standard deviation." Let us review the meaning and use of that term. If you look at any distribution, say of test scores, and compute the mean of those scores, you can then talk of the variability of each score from the mean. For example, if the mean of a distribution of scores is 50, and we wish to discuss a certain score in reference to the mean, we say the score has a deviation of -10 from the mean. But if we ask "What is the total deviation of all scores from the mean?" then we come up with the answer zero, since of necessity for each set of scores above the mean there must be a set equidistant below the mean. For this reason a procedure is needed which will describe not only the deviation, but also the total amount of deviation is some <u>standard</u> and acceptable way. For this reason the <u>standard</u> deviation was devised.

If we have a set of scores, 1, 2, 3, 4, 4, 5, 6, 7, the mean score is 4. The standard deviation of this set of scores is computed by the formula:

$$\sigma = \sqrt{\frac{\Sigma d^2}{N-1}}$$

σ	=	standard deviation
$\sqrt{\quad}$	=	square root of
Σ	=	the sum of
d^2	=	the deviations from the mean squared
N	=	the number of cases

The score 5 deviates by 1 from the mean (4), so we write down $1^2 = 1$ (see below). In similar fashion we square each of the deviations from the mean:

(3)	$-1^2 = -1 \times -1 = 1$	(5)	$1^2 = 1$
(2)	$-2^2 = -2 \times -2 = 4$	(6)	$2^2 = 4$
(1)	$-3^2 = -3 \times -3 = 9$	(7)	$3^2 = 9$

Adding (Σ) up the squared deviations (d^2) we get a total of 28. We then divide by 7 (N-1) for a result of $\sqrt{4}$. Since the square root of 4 is 2, the standard deviation of this set of scores is 2.

Thus we can describe any score within two of the mean as being within one standard deviation of the mean. In our example this would include all the scores from 2 through 6.

By reference to Chart 2, we see that, in a normal distribution, scores within one standard deviation of each side of the mean include 68.23% of the cases. This enables us to determine what kind of group we have, whether or not it resembles a normal distribution or is "skewed" to the right or left.

Similarly, 95.38% of the cases in a normal distribution fall within two standard deviations on each side of the mean. In scores such as those in the Scholastic Aptitude Test of the College Entrance Examination Board (commonly called College Board scores) the scores are "transformed" so that the mean is 500 on each test and the standard deviation is 100. Thus a score of 400 to 600 is within one standard deviation of the mean, and hence the range of scores is 200 to 800 (three standard deviations on each side of the mean).

Frequency

Frequency (f) in statistics refers simply to the number of times an event occurs. For example, we might have a distribution of scores which looks like this:

Scores	(f)
21-30	2
31-40	5
41-50	7
51-60	8
61-70	4
71-80	1

The column designated by (f) refers to the frequency of occurrence of a given set of scores. Thus, there were seven occurrences of scores in the 41-50 group. The concept of frequency is useful because it gives a quick view not only of the range of the thing measured, but also of the points at which results occurred in large numbers.

Central Tendency

One way to describe a group of measurements (scores or whatever) is in terms of some central tendency exhibited by the group. This usually takes the form of a number representing some kind of <u>average</u>. The most common of these averages are the mean (arithmetic average), the median (middle point of a series), and the mode (most frequent occurrence in a series). Shown below are examples of each:

Scores	(f)
5	
5	3
5	

Scores	(f)
6	1
8	1
11	2
11	
14	1
17	1
Total 82	

Mean = 82/9 = 9.1

Median is 8 (the middle number - four scores on each side)

Mode is 5 - the most frequent number

As you can see, the kind of average you select makes a difference in your description of a group or series. Sometimes one average may take preference over another. For example, if nine men each earn $10,000 a year, and a tenth man earns $1,000,000, the averages look like this:

Mean	=	$109,000
Median	=	10,000
Mode	=	10,000

Obviously if you describe the income of this group in terms of the mean, you convey little, if any, information. On the other hand, in the previous example of scores, the use of the mean or the median better describes the group of scores.

This raises an interesting point which ought to be carefully observed: When talking in measurement or statistical terms, always use terms which communicate best. Don't assume any statistical sophistication on the part of your reader or listener. Make your descriptions as simple as you can, without sacrificing accuracy.

Some Frequently Used Scores

Raw Scores

There are many ways of expressing measurement scores, each of which has certain advantages in given situations. The raw score is simply the number of right answers or occasionally the number right with a correction for guessing. Raw scores are often more useful in the immediate classroom context for judging individual performance than are other types of scores. If a fourth-grade learner spells correctly 100 words of an appropriate level of difficulty, this raw score gives us an immediate and direct measure of his ability in this skill in a classroom context. Similarly, if a teacher has an objective of teaching learners to multiply 5 sets of two-digit numbers lying between 45 and 50, a raw score is directly indicative of the ability or inability to perform the task.

Rank

Another basic kind of score is rank. Rank in class, or rank in a test, is often a good descriptor of a learner's position relative to others. If we say a learner is fifth in his class of 20 we compare him to others at the same level of instruction. Of course, such a rank makes no comment about the learner's level of mastery or ability or achievement, but only where he stands relative to others. Note that where there are <u>tied</u> ranks the next rank should be <u>two</u> or more below the tied ranks. For example:

Scores	Rank
38	1
35	2
32	3
32	3
30	5
24	6
24	6
24	6
21	9
20	10

If two runners tie for first place, the next runner is <u>not second</u>. He is third.

Derived Scores

Raw scores are often translated into other kinds of scores in order that they may be compared with scores of other tests and also to make them more meaningful. In addition to ranking scores (discussed above) we may also translate them into standard scores, percentiles, grade scores, and intervals.

Standard Scores

In order to translate raw scores into standard scores we first compute the mean and the standard deviation of the raw scores. Suppose that in a group of scores we find a mean of 20 and a standard deviation of 2. We can then express each score in the group in terms of standard deviations. For example, if a score is 19 we can say it is 1/2 standard deviation below the mean.

At this point we can construct a standard score scale to suit our convenience. If we arbitrarily chose one with a mean of 100, and a standard deviation of 10, then our score of 19 would now become 95 (1/2 standard deviation below the mean). A score of 22 would be 110 (1 standard deviation above the mean), and so on. Thus we see that a standard score scale can be arbitrarily set to suit the evaluator's purposes. Scores on that scale must, however, reflect the position of the original scores relative to the mean, in terms of standard deviations. Note on Chart 2, for example, that the College Board standard score scale has an arbitrary mean of 500 and a standard deviation of 100. Other examples of standard scores are T-scores and z-scores. See the glossary of terms for definitions of these as well as other statistical terms.

Percentiles

The relationship of a score to other scores may also be stated as a percentile. If a raw score of, say, 35 is at the 90th percentile of a group of scores, we use this percentile as another way to describe the score. A 90th percentile score means that the score is higher than 90% of all the scores in the group. One advantage of using percentiles to describe scores is obviously that it describes the score relative to the group and this is easy for non-statistically-minded people to understand. A standard score of 20 might not have much meaning to a parent, for example, but an equivalent percentile rank of 80 probably would.

Grade Scores

Test scores, particularly those of grades 1-8, are often expressed as grade scores or grade equivalent scores. Thus a third-grade learner might have a reading vocabulary score of 4.3. This could mean either fourth year, third month grade equivalent, or fourth year plus 3/10 of a year grade equivalent, depending on the test publisher.

Care should be taken in using and interpreting grade scores. On some tests a difference of one raw score point can change the grade score by several months. Since most test instruments will show some measurement error, a difference of one or two raw score points on two different administrations is to be expected. Hence one should always view grade scores in the light of measurement error. Computation of measurement error is discussed in a later section of these guidelines.

Interval Scores

Another kind of derived test score which has wide use is the interval. These may be in the form of quartiles or deciles or stanines. A decile is any of the nine points that divides a score scale into ten intervals. Each interval includes one-tenth of the total frequency. Similarly, a quartile is any of the three points on the score scale that divides it into four parts of equal frequency. Stanines are intervals which represent nine divisions of the baseline on the normal curve of distribution. Each division of the stanine is 0.5 σ long on the baseline, with the exception that the end divisions (1 and 9) includes the remainder of the area. Stanine 5 is in the center of the baseline and runs from -0.25σ to $+0.25\sigma$ on each side of the mean. Note on Chart 2 the percentages of the area of the normal curve which are in each stanine division. Stanines are also useful in describing scores to persons who are not too familiar with test score terminology. Since they are single-digit descriptions, their relative position is easy to understand. When using stanines for score descriptions, however, always be sure to include a statement about the percent of cases contained in each division. Otherwise your audience may get the impression that each division contains equal percentages of the cases.

Standard Error of Measurement

The standard error of measurement is a quantity which gives us some idea how far a given learner's score is from his true score. In other words, the standard error of measurement is an estimate of the standard deviation of a learner's score if he were to be measured several more times. Standard error of measurement (SE_{meas}) is computed from the formula:

$$SE_{meas} = \sigma \sqrt{1 - r}$$

in which r is the reliability coefficient of the instrument used and σ is the standard deviation of the scores on the test. If the reliability coefficient of a given test is .84 and the standard deviation is 10, then:

$$
\begin{aligned}
SE_{meas} &= 10 \sqrt{1 - .84} \\
&= 10 \sqrt{.16} \\
&= 10 \times .4 \\
&= 4
\end{aligned}
$$

This quantity, 4, is <u>one</u> standard error and tells us that approximately two-thirds of the time, if the test were repeated, individual or group scores would fall within one standard error ($\pm$4) of their "true" score. Similarly, <u>two</u> standard errors would be 8, and we could say that approximately 95 times out of 100, retest scores would fall within $\pm$8 of the true score.

Correlation

As the name implies, correlation is a method of describing how two or more things are related. In testing, correlation descriptions are precise mathematical ways of stating the relationship between test scores or between a score and some other presumably related occurrence, such as a grade in a class, for example. These mathematical descriptions are called correlation coefficients.

There are several methods of computing correlation coefficients. We shall discuss here one that is probably most useful in the evaluator's work. It is called the Pearson r. If we are attempting to compute a correlation coefficient (r) between two sets of scores, here is how we proceed:

1. List the two sets of scores
2. List the deviation of each score from the mean and square the number
3. List each score as a standard deviation
4. Multiply each standard deviation of a score in the first test by its corresponding standard deviation in the second test
5. Add the sum of the products obtained in (4)
6. Divide this sum by the number of persons tested
7. Result is the correlation coefficient(r)

	(1) Scores		(2) Deviations		(3) Standard Deviations		(4) $SD_x \times SD_y$
Learner	Test X	Test Y	DX^2	DY^2	Test X	Test Y	
1	13	24	9	16	1.50	1.00	1.50
2	12	26	4	36	1.00	1.50	1.50
3	12	24	4	16	1.00	1.00	1.00
4	11	22	1	4	0.50	0.50	0.25
5	11	18	1	4	0.50	0.50	0.25
6	10	20	0	0	0.00	0.00	0.00
7	9	22	1	4	0.50	0.50	0.25
8	9	18	1	4	0.50	0.50	0.25
9	9	14	1	36	0.50	1.50	0.75
10	7	18	9	4	1.50	0.50	0.75
11	7	14	9	36	1.50	1.50	2.25
TOTALS	110	220	40	160			8.75
Means	10	20					

Standard Deviations $= \sqrt{\frac{40}{10}} \quad \sqrt{\frac{160}{10}}$

$= \quad 2 \quad 4$

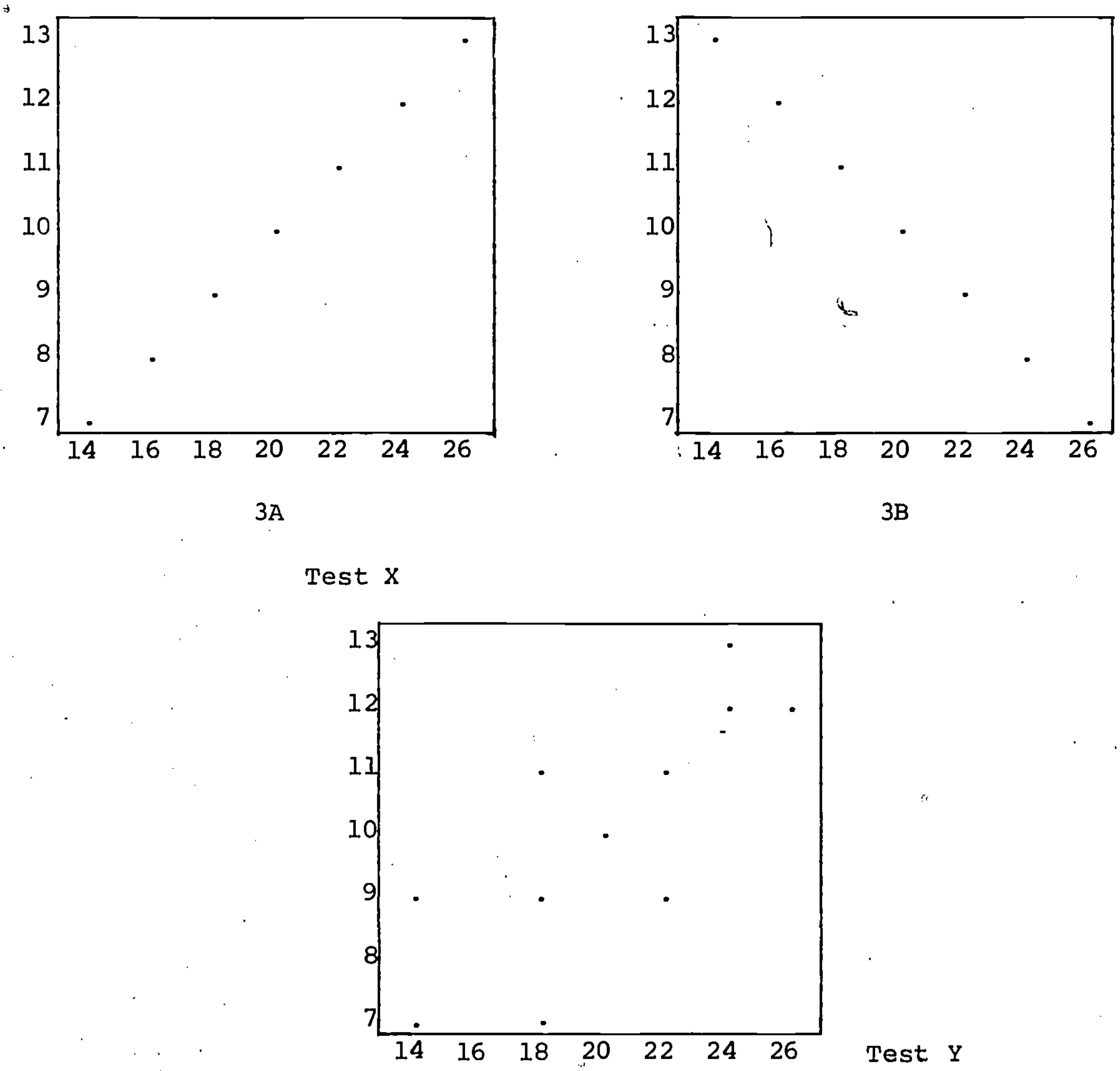
(5) 8.75

(6) $\frac{8.75}{11} = .795$

(7) $r = .795$

Computation of Correlation Coefficient (Pearson r)

A correlation of 0.0 means the scores are not related. A correlation of 1.00 indicates a perfect positive relationship and a correlation of -1.00 indicates a perfect negative relationship. The following scattergrams in Figure 3 show various correlations of two sets of test scores. 3A shows a perfect positive correlation. 3B shows a perfect negative correlation. 3C shows the correlation computed above.



3C

Chart 3

Another way to compute the correlation coefficient, without computing the standard deviation, is by using the formula:

$$rxy = \frac{\Sigma xy}{\sqrt{(\Sigma x^2)\ (\Sigma y^2)}}$$

rxy = correlation between x and y

x = deviation of any x score from the mean in test X

y = deviation of any y score from the mean in test Y

Σ = the sum of

How high does a correlation have to be to have much significance? This is a difficult question to answer, because it depends a great deal on what amount of relationship is useful in a particular situation. For example, if you use test scores as a basis for sectioning English classes, then any correlation (positive or negative) is better that drawing names out of a hat. On the other hand, if you are trying to improve successful placements where there is already a 70% success factor you would need to have a very high correlation for it to be of value. In general, however, we can say that correlations from 0 to $\pm$30 to $\pm$70 are increasingly useful, and those from $\pm$70 to $\pm$1.00 provide powerful indices of relationships. A handy index to consider in trying to determine the usefulness of a correlation coefficient is the quantity r^2. This is the percentage of variance in Test Y, for example, that is explained by Test X.

Reliability — By the reliability of a measurement device we mean the extent to which it is consistent in yielding the same score on different occasions.

Validity — By the validity of a measurement device we mean the extent to which it measures what it is supposed to measure. Both reliability and validity are special forms of correlation.

There are several methods for determining the reliability of a test. Three of the most widely used ones are the test-retest, the use of different forms of the same test, and the split-half method.

The correlation coefficient computed in the last section (see page 34) is actually a reliability coefficient, since the problem dealt with a set of test-retest scores. Let us look at another, more common, method of calculating reliability, namely the split-half method. In this method, the test is administered and scored, then <u>two</u> scores for each person are calculated by scoring alternate halves of the test. This could be done by scoring all the odd numbered questions and all the evens. The formula for computing a reliability coefficient from split-halves data is called the Spearman-Brown Prophecy Formula:

$$r_{ww} = \frac{2r_{\frac{1}{2}\frac{1}{2}}}{1 + r_{\frac{1}{2}\frac{1}{2}}}$$

r_{ww} = reliability of the whole test

$r_{\frac{1}{2}\frac{1}{2}}$ = the correlation of half of the test with the other half

If the half-tests have a correlation of .60, then:

$$r_{ww} = \frac{1.20}{1.60}$$

$$= .75$$

Correlations of measurement device scores with related criteria are called validity coefficients. An example would be the correlation between a score on an English test and a grade (criterion) in a class in English. This is another example of the correlation calculation explained on page 34.

Multiple Correlation and Prediction

But a criterion (grade in an English class, for example) is rarely if ever due to just one cause (ability in English, for example). The factors which affect the criterion are often many, and thus we speak of a multiple correlation. If the factors which result in a given criterion performance are indeed related to that criterion, then usually a higher (and hence more useful) correlation will exist. If we could isolate all the factors which produce a given criterion performance we would have a perfect multiple correlation.

Multiple correlations are often computed for the purpose of predicting future performance. For example, the multiple correlation of several junior high school course grades and test scores with high school grades can provide counselors with information to aid in class placement and course selection.

8. Design report format

In designing a format for reporting measurement and evaluation data, great care should be exercised to make the format simple and intelligible to the different publics who will read the report. If statistical terms or numbers are used, these should be explained in footnotes or in a glossary. Never assume that the reader is familiar with any words or terms other than those used in everyday conversational English.

The report format should include, in a prominent place, the six-item performance objectives as described on page 3. The reason for this is so that the person reading the report can compare the measurement and evaluative data directly with the stated objectives. Performance evaluation which is presented without reference to specific performance objectives is worthless.

Preprogram measurement data and any other pertinent baseline data should be listed early in the report, because this sets the stage for understanding learner need. Similarly, postprogram data should be presented in such a way that it can easily be compared with preprogram data.

In addition to listing mean scores, which could be in the form of standard scores, grade level equivalents, or stanines, indications should be given of mean growth and the number and percentage of individuals who attained each objective and the number and percentage of those individuals who did not achieve the objective. This provides a basis for evaluating the program in terms of objective attainment. Where available, subgroup mean scores should be included in the report, since often such information is not detectable within the set of larger group scores. For example, a total sample of third-grade reading vocabulary scores might show a mean grade level equivalent of 4.0, concealing the fact that some subgroup, say boys, had a mean score of only 3.0. In this way better program evaluation is possible and new approaches may be tried to resolve learner needs detected by this method.

All measurement data presented in the report should include standard error of measurement information, where possible, in order that decisions based on the data may be more accurate. In some cases the standard error of measurement is too large to permit valid conclusions to be made from the reported data.

9. Design program monitoring system

A monitoring system is a method for determining whether or not the planned program has been implemented.

There are two good reasons why monitoring systems should always be designed for instructional programs:

1. to provide information and documentation about the conduct

of the program (organization, facilities, cost, etc.) and

2. to provide feedback for change in the program

Only if the monitoring design provides timely, periodic, and accurate information about the conduct of the program, can the evaluation of outcomes be valid and realistic.

Following is an example of part of a program monitoring system:

Objective No. 2 Time Interval Sept. 15 - Oct. 30

INSTITUTIONAL VARIABLES	INSTRUCTIONAL VARIABLES				
	Organization	Content	Method	Facilities	Cost
Student	Class - 5 hr. per week Lab. - 2 hrs. per week	Algebra I Basic Linear Equations			
Teacher			Lecture Work-groups		
Adminis-trator	Organizes Training	In-Service Meeting			
Educ. Special-ist		In-Service Meeting	Tutoring	Teaching Machine	$50 per day 2 days per week
Family		Conference with Teacher			
Commun-ity			Service Club Presen-tations		

The system designed should then provide for monitoring each of the items listed in the chart.

10. Write calendar of events

Once all of the planning for events in Section I (Needs Assessment) and Section 2 (Design of Program Evaluation) has been completed, a calendar of events should be constructed to show the sequence and flow of work to be accomplished by specific dates. This is a very important step in evaluation design, because it helps evaluators and program planners to find the real constraints within which, or around which, they must work. Often such a calendar can point up resource deficiencies (people, time, money) and highlight problems which can be resolved only by timely planning. Ideally, when the calendar is constructed it should include time, cost and people allotments for all the events from the earliest steps of the needs assessment to the final recycling recommendations. Wherever possible in the calendar alternative dates should be established to help overcome unforeseen interventions.

On the following page is an example of part of a calendar of events.

Expected Date of Event	Alternative Date(s) of Event	Activities, Materials, Facilities, Costs	Persons Responsible	Actual Completion Date
Sept. 3	No later than Sept. 4	Deliver Math Pretests to Teacher	Curriculum Coordinator	Sept. 3
Sept. 7	No later than Sept. 11	Administer Math Pretests	Teacher	Sept. 10
Sept. 13	None	Return Math Pretest Results to Teacher	District Test Director	Sept. 17
Sept. 14	None	Begin Instruction in Linear Equations	Teacher	Sept. 18
Sept. 17, 18 and each Wed. and Thurs. thereafter until Oct. 25	Sept. 18 and 19	Individual Tutoring	District Math Specialist	Sept. 19
Nov. 1	Nov. 2	Posttest in Linear Equations Delivered to Teacher	Curriculum Coordinator	Nov. 1
Nov. 3	Nov. 5	Administer Posttest in Linear Equations	Teacher	Nov. 3

III. IMPLEMENTATION OF PROGRAM EVALUATION

1. Monitoring instructional and institutional variables

Instructional variables include program organization, content, method, facilities and cost. A systematic monitoring system will attempt to collect periodic information on each of these variables, for only in this way can an accurate evaluation be made of the factors which really caused learner change.

Program organization refers to the ways in which learners are organized for instruction - nongraded class, homogeneous ability grouping, etc. Content defines the particular body of knowledge to be included in the program - history, geometry, etc. By method is meant the various types of activities or systems by which teaching is effected - lecture, team teaching, student aides, multimedia approaches, etc. Facilities include not only classroom space, but supportive areas such as language labs. Equipment and expendable materials are also classified as facilities. The cost variable should include not only operational costs but calculations related to outcomes. Programs are sometimes established which produce desired results but at per pupil outcome costs which make them prohibitive.

Institutional variables include students, teachers, administrators, specialists, the family, and the community--in short, all those involved in a particular educational process.

The student variable may be described in terms of age, sex, ethnic origin, achievement level, etc.

The teacher variable might include grade level background, teaching majors or minors, special training or degrees held.

The administrator would be the person directly responsible for a specific educational program - usually the principal.

The specialist is a person who provides assistance in some specific aspects of the program, as for example, a tutor in linear equations or a laboratory reading specialist.

Family includes those persons in the student's immediate family group.

Community includes service groups, political groups, the P.T.A., and so on.

2. Collecting interim measurements

In order for evaluation to be meaningful and useful to decision-makers, it should be an ongoing process, rather than something which occurs only at the end of a program. Interim measurement can often provide clues for the improvement of instruction and for diagnosis of individual problems. One note of caution here--test-retest procedures have often led to false conclusions that learner change was taking place whereas actually the change was a function of measurement error.

Always check on measurement error, particularly in working with performance contracting programs.

3. Review and revise evaluation design and calendar of events

Careful monitoring of instructional variables and interim measurement will often bring about changes in the evaluation design and in the calendar of events. Such procedures might show, for example, that within-program objectives were too ambitious for a given time span, or, on the other hand, that the objectives were attained more rapidly than anticipated and that time is now available for additional objectives. There is nothing sacred about evaluation designs or calendars of events. Each must be flexible enough to accommodate changes indicated by unanticipated situations in the program.

4. Document program operation

The reason for this documentation is to indicate any changes from the original intent.

IV. MEASUREMENT OF OBJECTIVE ATTAINMENT

This is the payoff in educational evaluation. This is where we learn what the actual outcomes of the program are in terms of learner performance. It is essential that the measurement of objective attainment be conducted carefully in order that decision-making about future programs will not be contaminated by faulty conclusions.

1. Collecting postprogram data

Ideally postprogram data should be collected by persons not associated with the instructional program, using instruments specifically designed to measure performance related to objectives. Care should be exercised to preserve the security of measurement instruments in order that "teaching the test" and cheating procedures may be reduced to a minimum. Uniformity of scheduling of administration times should be maintained. There is always a problem as to precisely when postprogram data should be collected. On the one hand, if data collection is at or near the end of the program, the analysis of such data may not be completed in time for individual learner need diagnosis and counseling. On the other hand, if data collection is scheduled too early, the instructional program may not have been sufficiently completed to permit optimum attainment of performance objectives. Where programs are ongoing, for example a three-year Title I reading program, it is probably better to schedule annual data collection well before the end of that year's phase, so that evaluation conclusions can be built into continuation plans. In any event, postprogram data collection should take place when it can be of maximum effect in evaluating performance objective attainment.

2. Analyze data

By this term is meant simply the interpretation of what the data mean. Such interpretation should always be in terms the non-evaluator can understand, and should always be accompanied by estimates of the amount of credibility we can assign to the data. It is a fact of our educational life, that if we evaluate programs in units of semesters or years, we will usually get increasingly greater ranges of performance, hence greater measurement error, and hence less credibility in the results. Knowing that, however, we can take steps to guard against false conclusions.

3. Compare with preprogram data

The comparison of preprogram and postprogram data is essential to the evaluation of performance objective attainment and to a determination of remaining learner need. Here the evaluator will use the statistical techniques determined in the program evaluation design. Obviously, any comparisons should be made in similar terms, i.e. comparing grade equivalent scores on the pretest with grade equivalent scores on the posttest, etc.

4. <u>Compare attainment with performance objectives</u>

If the performance objectives are properly written they will contain proficiency levels against which to measure performance attainment. Be sure to point out in comparing attainment with performance objectives that while the group as a whole may have achieved the desired proficiency level, there may be many learners or subgroups who did not achieve this level. For example, a desired proficiency level for a beginning third-grade class in reading might be a 4.0 grade level equivalent by the end of the year. If the group indeed achieves a 4.0, obviously a considerable number of the learners will have scores below 4.0. In other words, don't let the forest hide the trees.

Performance objective attainment data should always include percentages of learners who attained the objective and those who did not. It is important to keep in mind that to the extent that one learner failed to attain the performance objective--to that extent the program failed.

5. <u>Write recommendations for use of measurement data</u>

We take the position that it is not enough for the evaluator merely to evaluate and let it go at that. He is the person who must impress upon his colleagues the significance and relevance of his findings, and then to make specific recommendations which are derived from his findings. There are, of course, many areas in which the evaluator can make valuable recommendations to the learner, to the school, and to the community. To mention only a few--learner placement and grouping, learner diagnosis, counseling and guidance, identification of exceptional children, interpretation of the school to the community, and for educational research.

In the next section on reporting and recycling we shall see some suggestions for getting your recommendations into the right hands.

V. REPORTING AND RECYCLING

Reporting and recycling are purposely presented together in the evaluation model, because it is only through the evaluation report recipients that program modification can be effected. Recycling, in the sense used here, means a complete return to the first section of the model, needs assessment, to determine how that section, and subsequent sections, may be modified in the light of evaluative information obtained during and after the conduct of the program.

1. Report postmeasurement

After postprogram data have been collected and analyzed, they should be disseminated in a manner designed to create program improvement. Everybody immediately involved in the program--learners, teachers, administrators--should be apprized of the program outcomes in terms of learner performance. Too often postprogram measurement and evaluation are routinely reported to some higher official and the significance of the evaluation is lost somewhere in a steel file. That is the fault of the evaluator. Part of his job is to see to it that his reports are discussed and his recommendations acted upon.

2. Report pre-post comparative data

Much of the same may be said of pre-post comparisons. Postmeasurement is always better understood in the light of original learner status. It also provides a basis for determining the role a given program may have played in learner achievement. Furthermore, pre-post comparisons give persons not acquainted with the program a bird's eye view of program outcomes. Such data are of great value in reports to the general public. Incidentally, another role of the evaluator is to follow up on evaluation releases to see that they represent an accurate portrayal of the situation and that the message has not been editorially obscured.

3. Report analysis of objectives attainment

The report of the analysis of objectives attainment should be made on an individual basis to learners, program instructional personnel, and parents. Mean scores of group and subgroups should be reported to administration and other officials when necessary (such as federal program officials).

4. Revise variables, sampling

In the light of performance objectives outcomes, revision of sampling methods and institutional, instructional, and behavioral variables may be advisable for subsequent program offerings. In small population areas some sampling methods may not be adequate for evaluating some subgroups. For such subgroups "oversampling" (including more than a proportional representation in the sample) may be advisable.

5. Revise objectives

On occasion a revision of performance objectives may be necessary, if it is found that substantial percentages of learners are not attaining objectives in what are thought to be good programs. Such a step should only be taken, however, if there is control group evidence, by a higher ability group, for example, that the objectives have too high proficiency levels. Before changing proficiency levels, however, it is advisable to look at the measurement devices being employed, to determine if they represent adequate measurement of objectives.

6. Revise measurement controls, instruments

No standardized instruments are likely to measure all of the objectives of a given instructional program. To the extent that they do not, additional measures should be contructed or selected from existing collections. Collections in many cognitive and affective areas are available through the National Assessment of Educational Progress, a project of the Educational Commission of the States, in Denver, Colorado, or through private groups such as the Instructional Objectives Exchange in Los Angeles, California. These collections are available at nominal cost and can be valuable supplements to standardized measurement instruments.

Measurement controls should also be revised on the basis of feedback derived from the monitoring system. Time of day, time of year, scheduling, etc. may require revision, as may administration procedures.

7. Recommend program changes

The evaluator is in a better position than anyone else to recommend program changes, based on his experience with performance outcomes. Instructional personnel are often too close to the scene to be able to see program deficiencies. Obviously, the evaluator may tread on a few toes in offering his recommendations, but that is one of the occupational hazards of being an evaluator.

8. <u>Recommendations for learners not achieving objectives</u>

Probably the most important postevaluation task the evaluator has is to make recommendations for learners who do not achieve the stated performance objectives. Obviously the program failed for these learners, whatever the reasons, and it is incumbent upon the evaluator not only to deduce as much as he can from the data available to him as to why these learners failed to meet the objectives, but beyond that he must follow through on his recommendations (to learners, parents, teachers, administrators) to see that additional assistance is provided. To do otherwise is to deny the whole purpose of educational evaluation.

VI. GLOSSARY OF TERMS USED IN THESE GUIDELINES

Affective	that variable of human behavior which relates to feelings or emotion
Baseline data	information used as a reference point for comparative purposes
Behavior change	an increase in any of the levels of behavior
Behavioral dimension variables	the variables of individual behavior; three variables are generally considered--cognitive, affective, psychomotor
Calendar of events	a calendar which indicates the projected dates of all events in a system
Central tendency	an average
Coefficient of correlation	a number (called r) which expresses the degree of relationship of two variables. The number may extend from +1.00 (perfect positive relationship) through zero (no relationship) to -1.00 (perfect inverse relationship)
Cognitive	that variable of human behavior which relates to knowledge and to the development of intellectual abilities
Control group	a group of people who serve as a reference point for another group under study
Correlation	the degree of relationship between two variables
Diagnosis	analysis of the nature of a problem
Evaluation	determination of the value of any object under study; measurement plus judgment
Expectation	mathematically, the chance that an event will occur (expressed as a fraction, e.g. 1/3) times the payoff
Experimental group	persons being studied in a program or "treatment"

Feedback — the return of system outputs to the input phase

Formulas

1. expectation

$$E = \frac{1}{n} \times P$$

E = expectation
n = number of possible chances
P = payoff

2. standard deviation - $\sigma = \sqrt{\frac{\Sigma d^2}{N}}$

σ = standard deviation
$\sqrt{\ }$ = the square root of
Σ = the sum of
d^2 = the square of the deviations from the mean
N = the number of cases

3. standard error of the mean

$$SE_m = \frac{\sigma}{\sqrt{n}}$$

SE_m = standard error of the mean
n = size of sample
σ = standard deviation

4. standard error of measurement -SE meas = $\sigma \sqrt{1-r}$

SE meas = standard error of measurement
σ = standard deviation of the test
$\sqrt{\ }$ = square root of
r = reliability coefficient of the test

5. z-score

$$z \text{ or } \sigma \text{ score} = \frac{d}{\sigma} = \frac{X - M}{\sigma}$$

z = z score
d = deviation
X = score
M = mean
σ = standard deviation of the test

Gain Score	a score which indicates an increase, such as an increase of one grade level
Institutional variables	(people variables) - the different persons involved in an educational program: students, teachers, administrators, specialists, families, communities
Instructional variables	those variables which affect the nature of instruction: organization, content, method, facilities, cost
Learner	a student at any level in any program
Matrix sampling	a sampling method which samples items as well as people
Mean	an arithmetic average
Measurement	the process of determining the current status of human behavior
Measurement control	a device to control any factors which might influence measurement outcomes
Measurement instrument	any written document whose purpose is to measure human behavior
Measurement of objective attainment	the determination of the degree to which a previously established objective has been accomplished
Median	the middle item of a distribution
Mode	the most frequent item in a distribution

Monitor	to keep track of, regulate, control
Multiple correlation	a relationship of two or more items to another item
Need	the difference between the present behavioral status of the learner and the proficiency level of the stated performance objective
Needs assessment	the processes of determining a need
Normal curve of probability	a mathematical model of the theoretical distribution of an infinite number of scores or measures
Percentile rank	the position of any score in a distribution indicating the percentage of scores below that position
Performance objective	a statement which predicts a future change in a behavioral level
Placement	the assignment of a person to a suitable place
Postprogram instrument	those instruments administered following, or near the end of, a program
Prediction	inferring future performance from a measurement score
Preprogram instruments	those instruments administered prior to the start of a program
Probability	the ratio of the outcomes that would produce a given event to the total number of possible outcomes
Proficiency level	a description of the status of the behavior being studied
Program evaluation	the process of measuring and judging the value of a program
Program evaluation implementation	putting into effect the various elements of an evaluation design
Psychomotor	that variable of human behavior which relates to muscular activity ensuing from prior mental activity

Range	the difference between the smallest and largest values of a variable
Recycling	utilizing evaluation data to improve planning processes
Reliability	a special form of correlation, the consistency of a measurement
Sampling	selecting a subset of a population
Scattergram	a device for illustrating the relationship between two variables
Standard deviation	a measure of variability which takes into account the actual variation of each item from the mean.
Standard error of measurement	a numerical statement of the probable difference between a measured score and a "true" score (See Formulas)
Standard error of the mean	a numerical statement of the error of estimation in any sampling situation; the standard deviation of the distribution of sample means
Standard score	a transformation of a z-score into a distribution with an arbitrary mean and standard deviation
Statistical regression	tendency of extreme scores to move toward the center of the distribution upon a second administration
Statistical techniques	methods of measuring by mathematical processes
Statistics	a branch of mathematics dealing with the collection, analysis, and interpretation of numerical data
T-score	a "normalized" score obtained by transforming the raw scores of a frequency distribution into equivalent scores in a normal distribution
Validity	the degree to which an instrument measures what it is supposed to measure

Variable (adjective)	capable of change
Variable (noun)	a quantity that may assume any one of a set of values
z-score	the deviation of a score from the mean, divided by the standard deviation of the test (See Formulas)

VII. APPENDIX

Page

1. Behavioral level definitions 62
2. Table of random numbers and example 64
3. Performance objective and test collections 66
4. Measurement control designs 67

1. SOME DEFINITIONS OF BEHAVIORAL LEVELS

COGNITIVE

Knowledge --	the recall of specifics and universals, the recall of methods and processes, the recall of a pattern, structure, or setting
Comprehension --	understanding in which the individual knows what is being communicated and can make use of the cognitive material without necessarily relating it to other material
Application --	the use of abstractions in particular and concrete situations
Analysis --	the breakdown of cognitive material into its constituent parts and detection of the relationships of the parts and of the way they are organized
Synthesis --	putting together of elements of cognitive material to form a cogent whole
Evaluation --	making judgments about the value, for some purpose, of cognitive materials

AFFECTIVE

Receiving --	awareness of, and willingness to receive, phenomena or stimuli
Responding --	sufficient involvement in a subject or activity to produce active commitment
Valuing --	acceptance of, and preference for, a value; commitment to a goal or objective
Organization --	conceptualization and organization of a value system
Characterization --	consistent action in accordance with the value system; the person can be "characterized" by his value system

PSYCHOMOTOR (tentative hypotheses by R. H. Dave)

Imitation --	imitation of an observable action

Manipulation -	development of skill in following direction; performance of selected actions
Precision -	proficiency of performance in reproducing a given act reaches a high level
Articulation -	coordination of a series of acts and establishing internal consistency among them
Naturalization -	automatic and spontaneous response in the performance of an act or series of acts; performance becomes "second nature"

2. Table of Random Numbers

	00000 01234	00000 56789	11111 01234	11111 56789	22222 01234	22222 56789	33333 01234	33333 56789
00	23157	54859	01837	25993	76249	70886	95230	36744
01	05545	55043	10537	43508	90611	83744	10962	21343
02	14871	60350	32404	36223	50051	00322	11543	80834
03	38976	74951	94051	75853	78805	90194	32428	71695
04	97312	61718	99755	30870	94251	25841	54882	10513
05	11742	69381	44339	30872	32797	33118	22647	06850
06	43361	28859	11016	45623	93009	00499	43640	74036
07	98806	20478	38268	04491	55751	18932	58475	52571
08	49540	13181	08429	84187	69538	29661	77738	09527
09	36768	72633	37948	21569	41959	68670	45274	83880
10	07092	52392	24627	12067	06558	45344	67338	45320
11	43310	01081	44863	80307	52555	16148	89742	94647
12	61570	06360	06173	63775	63148	95123	35017	46993
13	31352	83799	10779	18941	31579	76448	62584	86919
14	57048	86526	27795	93692	90529	56546	35065	32254
15	09243	44200	68721	07137	30729	75756	09298	27650
16	97957	35018	40894	88329	52230	82521	22532	61587
17	93732	59570	43781	98885	56671	66826	95996	44569
18	72621	11225	00922	68264	35666	59434	71687	58167
19	61020	74418	45371	20794	95917	37866	99536	19378
20	97839	85474	33055	91718	45473	54144	22034	23000
21	89160	97192	22232	90637	35055	45489	88438	16361
22	25956	88220	62871	79265	02823	52862	84919	54883
23	81443	31719	05049	54806	74690	07567	65017	16543
24	11322	54931	42362	34386	08624	97687	46245	23245

Suppose that from a group of 100 you wish to select a sample of 10 persons. Assign each person a number from 00 to 99. Then select any column of numbers from Table I and write down the last two digits of each of the first ten rows (or any other set of two digits you care to use). For example, if you select the last column, and use the last two digits in each of the first ten rows, your sample of 10 would be the persons with numbers 44, 43, 34, 95, 13, 50, 36, 71, 27 and 80.

Notice that if you had selected the second column you would have had numbers 59, 43, 50, 51, 18, 81, 59, 78, 81, 33. Number 59 occurred twice and number 81 occurred twice. You would then have to decide whether to include both 59's and both 81's in your sample or to discard one 59 and one 81 and select two additional numbers. In a truly random sample both double numbers would be left in the sample.

Table I was reproduced from Statistical Methods, by Allen L. Edwards, second edition, 1967, Holt, Rinehart and Winston, Inc.

3. Performance Objective and Test Collections

1. CTB/McGraw-Hill
   Del Monte Research Park
   Monterey, California 93940
   408/373-2932

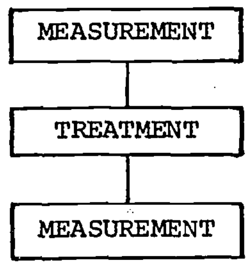
2. Educational Testing Service
   1947 Center Street
   Berkeley, California 94704
   415/849-0950

3. Instructional Objectives Exchange
   Box 24095
   Los Angeles, California 90024
   213/474-4531

4. Westinghouse Learning Corporation
   2680 Hanover Street
   Palo Alto, California 94304
   415/493-1360

4. Measurement Control Designs[1]

1. One Group - Pretest/Posttest Design



Factors Controlled:

1. Selection: the evaluator is only interested in students studied and does not plan a comparison with other groups.

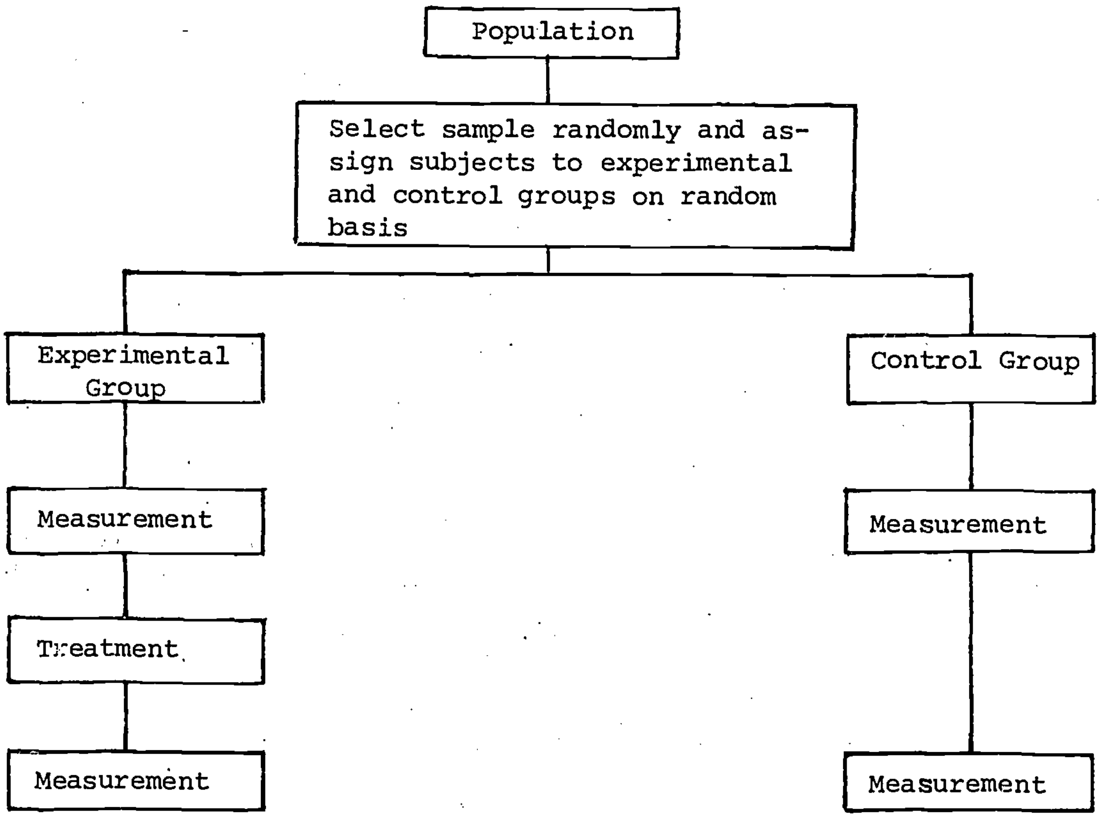
Factors Uncontrolled:

1. history
2. maturation
3. testing
4. instrumentation
5. regression

---

[1]After Evaluation Design, Educational Innovators Press, Tucson, Arizona, 1970, pp. 11-12.

Pretest/Posttest Control Group Design[1]



<u>Factors Controlled</u>:

1. history
2. maturation
3. testing
4. instrumentation
5. regression
6. selection

---

[1]After <u>Evaluation Design</u>, Educational Innovators Press, Tucson, Arizona, 1970, pp. 12-13.