DOCUMENT RESUME

ED 039 656 EC 005 463

TITLE	Training Materials for Gifted Evaluation Institute.
INSTITUTION	Cooperative Educational Research Lab., Inc., Northfield, Ill.; Illinois Univ., Urbana. Center for Instructional Research and Curriculum Evaluation.
SPONS AGENCY	Illinois State Office of the Superintendent of Public Instruction, Springfield. Dept. of Program Planning for the Gifted.
PUB DATE	68
NOTE	176p.
EDRS PRICE	EDRS Price MF-$0.75 HC-$8.90
DESCRIPTORS	Evaluation Methods, *Evaluation Techniques, *Exceptional Child Education, *Gifted, Inservice Programs, *Institutes (Training Programs), Instructional Materials, Measurement Instruments, Measurement Techniques, *Program Evaluation, State Programs, Statistics
IDENTIFIERS	Illinois, Illinois Gifted Program

ABSTRACT

Consisting of materials presented to the institute on evaluation of the Illinois gifted program, the document presents papers on educational evaluation and the evaluator's role in education. Also, the development of evaluation designs is considered and four evaluation plans are detailed. Questions regarding evaluation are listed; also listed and/or explained are workshop activities, exercises, and videotapes. Evaluative instruments discussed are rating and attitude scales and interview and observation schedules. Participant achievement tests, opinion surveys, and critiques are included; also included are materials on statistical computation, terminology, and references. A detailed text describing the 2-week institute in which the training package was tested by the Gifted Evaluation staff is available as EC 005 464. (JD)

TRAINING MATERIALS FOR GIFTED EVALUATION INSTITUTE

UNIVERSITY OF ILLINOIS

July 29 - August 9, 1968

SPONSORED BY:

COOPERATIVE EDUCATIONAL RESEARCH LABORATORY, INC. (CERLI)

Northfield, Illinois

CENTER FOR INSTRUCTIONAL RESEARCH AND CURRICULUM EVALUATION (CIRCE)

University of Illinois, Urbana

SUPPORTED BY:

DEPARTMENT OF PROGRAM DEVELOPMENT FOR GIFTED YOUTH

OFFICE OF THE SUPERINTENDENT OF PUBLIC INSTRUCTION

STATE OF ILLINOIS

TABLE OF CONTENTS

Attitude Inventory

Information Quiz (Achievement Test)

Information Quiz Key

Participant Interview Schedule Part I

Participant Interview Schedule Part II

Observation Schedules

Participant Opinionaire

Participant Critique Form

INDEX TO TRAINING MATERIALS FOR GIFTED EVALUATION INSTITUTE

July 29 - August 9, 1968

CERLI

Pre - Package

Description of Program and Description of Workshop Exercises

Robert Stake "The Countenance of Educational Evaluation"

List of References

List of Participants

Week 1 Package

Index

Revised list of participants

Viewing the Videotapes

Questions for videotape

Stufflebeam "Developing Evaluation Design"

Evaluation Plan I: "Does the gifted program in our school increase the student's ability to conduct independent study?"

Evaluation Plan II: "Which of these three sets of science materials should we use in our elementary science course?"

Evaluation Plan III: "Are we selecting the right students for our class in creative writing?"

Suggestions for Role Playing

Week 2 Package

Garth Sorenson "A New Role in Education: The Evaluator"

Exercise on Rating Scales

Attitude Scale Exercise

Glossary of Statistical Terms

Computation of Chi-Square

Computation of Pearson r

Computation of Spearman rho

Computation of t-test and Mann-Whitney U

Computation of Correlated t-test

One way and two way analysis of variance

Evaluation Plan IV: "Has the gifted program had an effect on the achievement of the participating students?"

Accessories

Video Tape I - Bob Stake and McQuarie

"Does the gifted program in our school increase the student's ability to conduct independent study?"

Video Tape II - Terry Denny and Nelson

"Which of these three sets of science materials should we use in our elementary science course?"

Video Tape III - Tom Hastings and Douglas Sjogren

"Are we selecting the right students for our class in creative writing?"

Video Tape IV - Bob Stake and School Personnel

"The Evaluator Role"

Evaluation Materials

Attitude Inventory on Evaluator Role

Institute Interview Schedule

Participant Critique Form

Lecture Observation Schedule

History of Institute Outline

Achievement Test

THE COUNTENANCE OF EDUCATIONAL EVALUATION

Robert E. Stake
Center for Instructional Research and Curriculum Evaluation
University of Illinois

President Johnson, President Conant, Mrs. Hull (Sara's teacher) and Mr. Tykociner (the man next door) are quite alike in the faith they have in education. But they have quite different ideas of what education is. The value they put on education does not reveal their way of evaluating education.

Educators differ among themselves as to both the essence and worth of an educational program. The wide range of evaluation purposes and methods allows each to keep his own perspective. Few see their own programs "in the round," partly because of a parochial approach to evaluation. To understand better his own teaching and to contribute more to the science of teaching, each educator should examine the full countenance of evaluation.

Educational evaluation has its formal and informal sides. Informal evaluation is recognized by its dependence on casual observation, implicit goals, intuitive norms, and subjective judgment. Perhaps because these are also characteristic of day-to-day, personal styles of living, informal evaluation results in perspectives which are seldom questioned. Careful study reveals informal evaluation of education to be of variable quality--sometimes penetrating and insightful, sometimes superficial and distorted.

Formal evaluation of education is recognized by its dependence on check lists, structured visitation by peers, controlled comparisons, and standardized testing of students. Some of these techniques have long histories of successful use. Unfortunately, when planning an evaluation, few educators consider even these four. The more common notion is to evaluate informally: to ask the opinion of the instructor, to ponder the logic of the program, or to consider the reputation of the advocates. Seldom do we find a search for relevant research reports or for behavioral data pertinent to the ultimate curricular decisions.

Dissatisfaction with the formal approach is not without cause. Few highly relevant, readable research studies can be found. The professional journals are not disposed to publish evaluation studies. Behavioral data are costly, and often do not provide the answers. Too many accreditation-type visitation teams lack special training or even experience in evaluation. Many check lists are ambiguous; some focus too much attention on the physical attributes of a school. Psychometric tests have been developed primarily to differentiate among students at the same point in training rather than to assess the effect of instruction on acquisition of skill and understanding. Today's educator may rely little on formal evaluation because its answers have seldom been answers to questions <u>he</u> is asking.

Potential Contributions of Formal Evaluation

The educator's disdain of formal evaluation is due also to his sensitivity to criticism--and his <u>is</u> a critical clientele. It is not uncommon for him to draw before

Reprinted from Teachers College Record, Vol. 68, No. 7, April, 1967.

him such curtains as "national norm comparisons," "innovation phase," and "academic freedom" to avoid exposure through evaluation. The "politics" of evaluation is an interesting issue in itself, but it is not the issue here. The issue here is the potential contribution to education of formal evaluation. Today, educators fail to perceive what formal evaluation could do for them. They should be imploring measurement specialists to develop a methodology that reflects the fullness, the complexity, and the importance of their programs. They are not.

What one finds when he examines formal evaluation activities in education today is too little effort to spell out antecedent conditions and classroom transactions (a few of which visitation teams do record) and too little effort to couple them with the various outcomes (a few of which are portrayed by conventional test scores). Little attempt has been made to measure the match between what an educator intends to do and what he does do. The traditional concern of educational-measurement specialists for reliability of individual-student scores and predictive validity (thoroughly and competently stated in the American Council on Education's 1950 edition of Educational Measurement)[1] is a questionable resource. For evaluation of curricula, attention to individual differences among students should give way to attention to the contingencies among background conditions, classroom activities, and scholastic outcomes.

This paper is not about what should be measured or how to measure. It is background for developing an evaluation plan. What and how are decided later. My orientation here is around educational programs rather than educational products. I presume that the value of a product depends on its program of use. The evaluation of a program includes the evaluation of its materials.

The countenance of educational evaluation appears to be changing. On the pages that follow, I will indicate what the countenance can, and perhaps, should be. My attempt here is to introduce a conceptualization of evaluation oriented to the complex and dynamic nature of education, one which gives proper attention to the diverse purposes and judgments of the practitioner.

Much recent concern about curriculum evaluation is attributable to contemporary large-scale curriculum-innovation activities, but the statements in this paper pertain to traditional and new curricula alike. They pertain, for example, to Title I and Title III projects funded under the Elementary and Secondary Act of 1966. Statements here are relevant to any curriculum, whether oriented to subject-matter content or to student process, and without regard to whether curriculum is general-purpose, remedial, accelerated, compensatory, or special in any other way.

The purposes and procedures of educational evaluation will vary from instance to instance. What is quite appropriate for one school may be less appropriate for another. Standardized achievement tests here but not there. A great concern for expense there but not over there. How do evaluation purposes and procedures vary? What are the basic characteristics of evaluation activities? They are identified in these pages as the evaluation acts, the data sources, the congruence and contingencies, the standards, and the uses of evaluation. The first distinction to be made will be between description and judgment in evaluation.

The countenance of evaluation beheld by the educator is not the same one beheld by the specialist in evaluation. The specialist sees himself as a "describer," one who describes aptitudes and environments and accomplishments. The teacher and school administrator, on the other hand, expect an evaluator to grade something or someone as to merit. Moreover, they expect that he will judge things against external standards, on criteria perhaps little related to the local school's resources and goals.

Neither sees evaluation broadly enough. Both description and judgment are essential--in fact, they are the two basic acts of evaluation. Any individual evaluator may attempt to refrain from judging or from collecting the judgments of others. Any individual evaluator may seek only to bring to light the worth of the program. But their evaluations are incomplete. To be fully understood, the educational program must be fully described and fully judged.

Towards Full Description

The specialist in evaluation seems to be increasing his emphasis on fullness of description. For many years he evaluated primarily by measuring student progress toward academic objectives. These objectives usually were identified with the traditional disciplines, e.g. mathematics, English, and social studies. Achievement tests--standardized or "teacher-made"--were found to be useful in describing the degree to which some curricular objectives are attained by individual students in a particular course. To the early evaluators, and to many others, the countenance of evaluation has been nothing more than the administration and normative interpretaion of achievement tests.

In recent years a few evaluators have attempted, in addition, to assess progress of individuals toward certain "inter-disciplinary" and "extracurricular" objectives. In their objectives, emphasis has been given to the integration of behavior within an individual; or to the perception of interrelationships among scholastic disciplines; or to the development of habits, skills, and attitudes which permit the individual to be a craftsman or scholar, in or out of school. For the descriptive evaluation of such outcomes, the Eight-Year Study[2] has served as one model. The proposed National Assessment Program may be another--this statement appeared in one interim report:

> "...all committees worked within the following broad definition of 'national assessment:'
>
> 1. In order to reflect fairly the aims of education in the U.S., the assessment should consider both traditional and modern curricula, and take into account all the aspirations schools have for developing attitudes and motivations as well as knowledge and skills..." [Italics added] (Educational Testing Service, 1965).[3]

In his paper, "Evaluation for Course Improvement,"[4] Lee Cronbach urged another step: a most generous inclusion of behavioral-science variables in order to examine the possible causes and effects of quality teaching. He proposed that the main objective for evaluation is to uncover durable relationships--those appropriate for guiding future educational programs. To the traditional description of pupil achievement, we add the description of instruction and the description of relationships between them. Like the instructional researcher, the evaluator--as so defined--seeks generalizations about educational practices. Many curriculum project evaluators are adopting this definition of evaluation.

The Role of Judgment

Description is one thing, judgment is another. Most evaluation specialists have chosen not to judge. But in his recent Methodology of Evaluation[5] Michael Scriven has charged evaluators with responsibility for passing upon the merit of an educational practice. (Note that he has urged the evaluator to do what the educator has expected the evaluator to be doing.) Scriven's position is that there is no evaluation until judgment has been passed, and by his reckoning the evaluator is best qualified to judge.

By being well experienced and by becoming well-informed in the case at hand in matters of research and educational practice the evaluator does become at least partially qualified to judge. But is it wise for him to accept this responsibility? Even now when few evaluators expect to judge, educators are reluctant to initiate a formal evaluation. If evaluators were more frequently identified with the passing of judgment, with the discrimination among poorer and better programs, and with the awarding of support and censure, their access to data would probably diminish. Evaluators collaborate with other social scientists and behavioral research workers. Those who do not want to judge deplore the acceptance of such responsibility by their associates. They believe that in the eyes of many practitioners, social science and behavioral research will become more suspect than it already is.

Many evaluators feel that they are not capable of perceiving, as they think a judge should, the unidimensional value of alternative programs. They anticipate a dilemma such as Curriculum I resulting in three skills and ten understandings and Curriculum II resulting in four skills and eight understandings. They are reluctant to judge that gaining one skill is worth losing two understandings. And, whether through timidity, disinterest, or as a rational choice, the evaluator usually supports "local option," a community's privilege to set its own standards and to be its own judge of the worth of its educational system. He expects that what is good for one community will not necessarily be good for another community, and he does not trust himself to discern what is best for a briefly-known community.

Scriven reminds them that there are precious few who can judge complex programs, and fewer still who will. Different decisions must be made--P.S.S.C. or Harvard Physics?--and they should not be made on trivial criteria, e.g. mere precedent, mention in the popular press, salesman personality, administrative convenience, or pedagogical myth. Who should judge? The answer comes easily to Scriven partly because he expects little interaction between treatment and learner, i.e., what works best for one learner will work best for others, at least within broad categories. He also expects that where the local good is at odds with the common good, the local good can be shown to be detrimental to the common good, to the end that the doctrine of local option is invalidated. According to Scriven the evaluator must judge.

Whether or not evaluation specialists will accept Scriven's challenge remains to be seen. In any case, it is likely that judgments will become an increasing part of the evaluation report. Evaluators will seek out and record the opinions of persons of special qualification. These opinions, though subjective, can be very useful and can be gathered objectively, independent of the solicitor's opinions. A responsibility for processing judgments is much more acceptable to the evaluation specialist than one for rendering judgments himself.

Taylor and Maguire[6] have pointed to five groups having important opinions on education: spokesmen for society at large, subject-matter experts, teachers, parents, and the students themselves. Members of these and other groups are judges who should be heard. Superficial polls, letters to the editor, and other incidental judgments are insufficient. An evaluation of a school program should portray the merit and fault perceived by well-identified groups, systematically gathered and processed. Thus, judgment data and description data are both essential to the evaluation of educational programs.

Data Matrices

In order to evaluate, an educator will gather together certain data. The data are likely to be from several quite different sources, gathered in several quite different ways. Whether the immediate purpose is description or judgment, three bodies of information should be tapped. In the evaluation report it can be helpful to distinguish between antecedent, transaction, and outcome data.

An antecedent is any condition existing prior to teaching and learning which may relate to outcomes. The status of a student prior to his lesson, e.g. his aptitude, previous experience, interest, and willingness, is a complex antecedent. The programmed-instruction specialist calls some antecedents "entry behaviors." The state accrediting agency emphasizes the investment of community resources. All of these are examples of the antecedents which an evaluator will describe.

Transactions are the countless encounters of students with teacher, student with student, author with reader, parent with counselor--the succession of engagements which comprise the process of education. Examples are the presentation of a film, a class discussion, the working of a homework problem, an explanation on the margin of a term paper, and the administration of a test. Smith and Meux studied such transactions in detail and have provided an 18-category classification system.[7] One very visible emphasis on a particular class of transactions was the National Defense Education Act support of audio-visual media.

Transactions are dynamic whereas antecedents and outcomes are relatively static. The boundaries between them are not clear, e.g. during a transaction we can identify certain outcomes which are feedback antecedents for subsequent learning. These boundaries do not need to be distinct. The categories should be used to stimulate rather than to subdivide our data collection.

Traditionally, most attention in formal evaluation has been given to outcomes--outcomes such as the abilities, achievements, attitudes, and aspirations of students resulting from an educational experience. Outcomes, as a body of information, would include measurements of the impact of instruction on teachers, administrators, counselors, and others. Here too would be data on wear and tear of equipment, effects of the learning environment, cost incurred. Outcomes to be considered in evaluation include not only those that are evident, or even existent, as learning sessions end, but include applications, transfer, and relearning effects which may not be available for measurement until long after. The description of the outcomes of driver training, for example, could well include reports of accident-avoidance over a lifetime. In short, outcomes are the consequences of educating--immediate and long-range, cognitive and conative, personal and community-wide.

Antecedents, transactions, and outcomes, the elements of evaluation statements, are shown in Figure 1 to have a place in both description and judgment. To fill in these matrices the evaluator will collect judgments (e.g. of community prejudice, of problem solving styles, and of teacher personality) as well as descriptions. In Figure 1 it is also indicated that judgmental statements are classified either as general standards of quality or as judgments specific to the given program. Descriptive data are classified as intents and observations. The evaluator can organize his data-gathering to conform to the format shown in Figure 1.

The evaluator can prepare a record of what educators intend, of what observers perceive, of what patrons generally expect, and of what judges value the immediate program to be. The record may treat antecedents, transactions, and outcomes separately within the four classes identified as Intents, Observations, Standards, and Judgments, as in Figure 1. The following is an illustration of 12 data, one of which could be recorded in each of the 12 cells, starting with an intended antecedent, and moving down each column until an outcome judgment has been indicated.

> Knowing that (1) Chapter XI has been assigned and that he intends (2) to lecture on the topic Wednesday, a professor indicates (3) what the students should be able to do by Friday, partly by writing a quiz on the topic. He observes that (4) some students were absent on Wednesday, that (5) he did not quite complete the lecture because of a lengthy discussion and that (6) on the quiz only about 2/3 of the class seemed to understand a certain major concept. In general, he expects (7) some absences but that the work will be made up by quiz-time; he expects (8) his lectures to be clear enough for perhaps 90 percent of a class to follow him without difficulty; and he knows that (9) his colleagues expect only about one student in ten to understand thoroughly each major concept in such lessons as these. By his own judgment (10) the reading assignment was not a sufficient background for his lecture; the students commented that (11) the lecture was provocative; and the graduate assistant who read the quiz papers said that (12) a discouragingly large number of students seemed to confuse one major concept for another.

Evaluators and educators do not expect data to be recorded in such detail, even in the distant future. My purpose here was to give twelve examples of data that could be handled by separate cells in the matrices. Next I would like to consider the description data matrix in detail.

Goals and Intents

For many years instructional technologists, test specialists, and others have pleaded for more explicit statement of educational goals. I consider "goals," "objectives," and "intents" to be synonymous. I use the category title Intents because many educators now equate "goals" and "objectives" with "intended student outcomes." In this paper Intents includes the planned-for environmental conditions, the planned-for demonstrations, the planned-for coverage of certain subject matter, etc., as well as the planned-for student behavior. To be included in this three-cell column are effects which are desired, those which are hoped for, those which are anticipated, and even those which are feared. This class of data includes goals and plans that others have, especially the students. (It should be noted that it is not the educator's privilege to rule out the study of a variable by saying, "that is not one of our objectives." The evaluator should include both the variable and the negation.) The resulting collection of Intents is a priority listing of all that may happen.

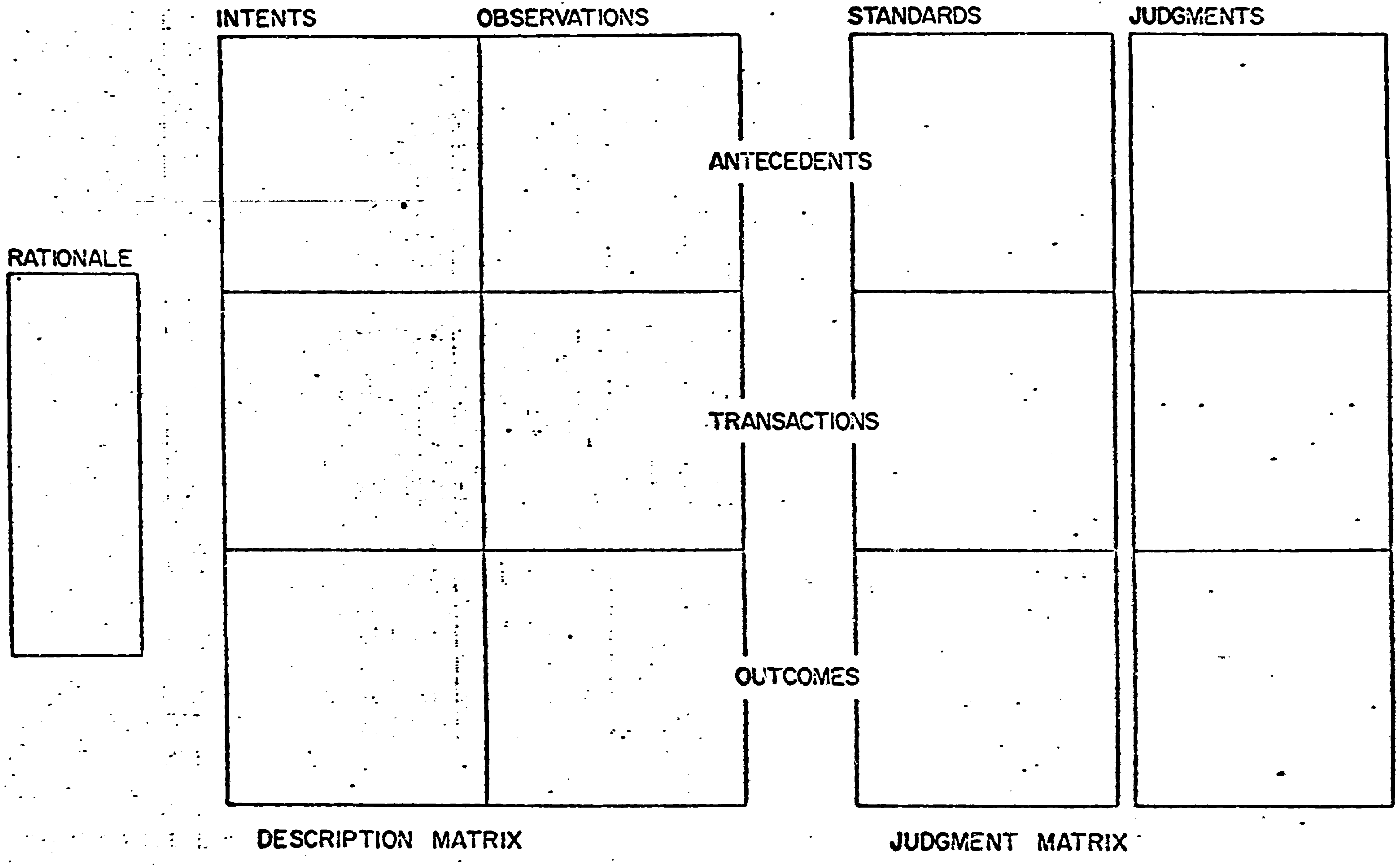

Figure 1. A layout of statements and data to be collected by the evaluator of an educational program.

The fact that many educators now equate "goals' with "intended student outcomes" is to the credit of the behaviorists, particula-ly the advocates of programmed instruction. They have brought about a small r form in teaching by emphasizing those specific classroom acts and work exercises which contribute to the refinement of student responses. The A.A.A.S. Science Project, for example, has been successful in developing its curriculum around behavioristic goals.[8] Some curriculum-innovation projects, however, have found the emphasis on behavioral outcomes an obstacle to creative teaching.[9] The educational evaluator should not list goals only in terms of anticipated student behavior. To evaluate an educational program, we must examine what teaching, as well as what learning, is intended. (Many antecedent conditions and teaching transactions can be worded behavioristically, if desired.) How intentions are worded is not a criterion for inclusion. Intents can be the global goals of the Educational Policies Commission or the detailed goals of the programmer.[10] Taxonomic, mechanistic, humanistic, even scriptural--any mixture of goal statements are acceptable as part of the evaluation picture.

Many a contemporary evaluator expects trouble when he sets out to record the educator's objectives. Early in the work he urged the educator to declare his objectives so that outcome-testing devices could be built. He finds the educator either reluctant or unable to verbalize objectives. With diligence, if not with pleasure, the evaluator assists with what he presumes to be the educator's job: writing behavioral goals. His presumption is wrong. As Scriven has said, the responsibility for describing curricular objectives is the responsibility of the evaluator. He is the one who is experienced with the language of behaviors, traits, and habits. Just as it is his responsibility to transform the behaviors of a teacher and the responses of a student into data, it is his responsibility to transform the intentions and expectations of an educator into "data." It is necessary for him to continue to ask the educator for statements of intent. He should augment the replies by asking, "Is this another way of saying it?" or "Is this an instance?" It is not wrong for an evaluator to teach a willing educator about behavioral objectives--they may facilitate the work. It is wrong for him to insist that every educator should use them.

Obtaining authentic statements of intent is a new challenge for the evaluator. The methodology remains to be developed. Let us now shift attention to the second column of the data cells.

Observational Choice

Most of the descriptive data cited early in the previous section are classified as Observations. In Figure 1 when he described surroundings and events and the subsequent consequences, the evaluator* is telling of his Observations. Sometimes the evaluator observes these characteristics in a direct and personal way. Sometimes he uses instruments. His instruments include inventory schedules, biographical data sheets, interview routines, check lists, opinionnaires, and all kinds of psychometric tests. The experienced evaluator gives special attention to the measurement of student outcomes, but he does not fail to observe the other outcomes, nor the antecedent conditions and instructional transactions.

*Here and elsewhere in this paper, for simplicity of presentation, the evaluator and the educator are referred to as two different persons. The educator will often be his own evaluator or a member of the evaluation team.

Many educators fear that the outside evaluator will not be attentive to the characteristics that the school staff has deemed most important. This sometimes does happen, but evaluators often pay too much attention to what they have been urged to look at, and too little attention to other facets. In the matter of selection of variables for evaluation, the evaluator must make a subjective decision. Obviously, he must limit the elements to be studied. He cannot look at all of them. The ones he rules out will be those that he assumes would not contribute to an understanding of the educational activity. He should give primary attention to the variables specifically indicated by the educator's objectives, but he must designate additional variables to be observed. He must search for unwanted side effects and incidental gains. The selection of measuring techniques is an obvious responsibility, but the choice of characteristics to be observed is an equally important and unique contribution of the evaluator.

An evaluation is not complete without a statement of the rationale of the program. It needs to be considered separately, as indicated in Figure 1. Every program has its rationale, though often it is only implicit. The rationale indicates the philosophic background and basic purposes of the program. Its importance to evaluation has been indicated by Berlak.[11] The rationale should provide one basis for evaluating Intents. The evaluator asks himself or other judges whether the plan developed by the educator constitutes a logical step in the implementation of the basic purposes. The rationale also is of value in choosing the reference groups, e.g. merchants, mathematicians, and mathematics educators, which later are to pass judgment on various aspects of the program.

A statement of rationale may be difficult to obtain. Many an effective instructor is less than effective at presenting an educational rationale. If pressed, he may only succeed in saying something the listener wanted said. It is important that the rationale be in his language, a language he is the master of. Suggestions by the evaluator may be an obstacle, becoming accepted because they are attractive rather than because they designate the grounds for what the educator is trying to do.

The judgment matrix needs further explanation, but I am postponing that until after a consideration of the bases for processing descriptive data.

Contingency and Congruence

For any one educational program there are two principal ways of processing descriptive evaluation data: finding the contingencies among antecedents, transactions, and outcomes and finding the congruence between Intents and Observations. The processing of judgments follows a different model. The first two main columns of the data matrix in Figure I contain the descriptive data. The format for processing these data is represented in Figure 2.

The data for a curriculum are congruent if what was intended actually happens. To be fully congruent the intended antecedents, transactions, and outcomes would have to come to pass. (This seldom happens--and often should not.) Within one row of the data matrix the evaluator should be able to compare the cells containing Intents and Observations, to note the discrepancies, and to describe the amount of congruence for that row. (Congruence of outcomes has been emphasized in the evaluation model proposed by Taylor and Maguire.) Congruence does not indicate that outcomes are reliable or valid, but that what was intended did occur.

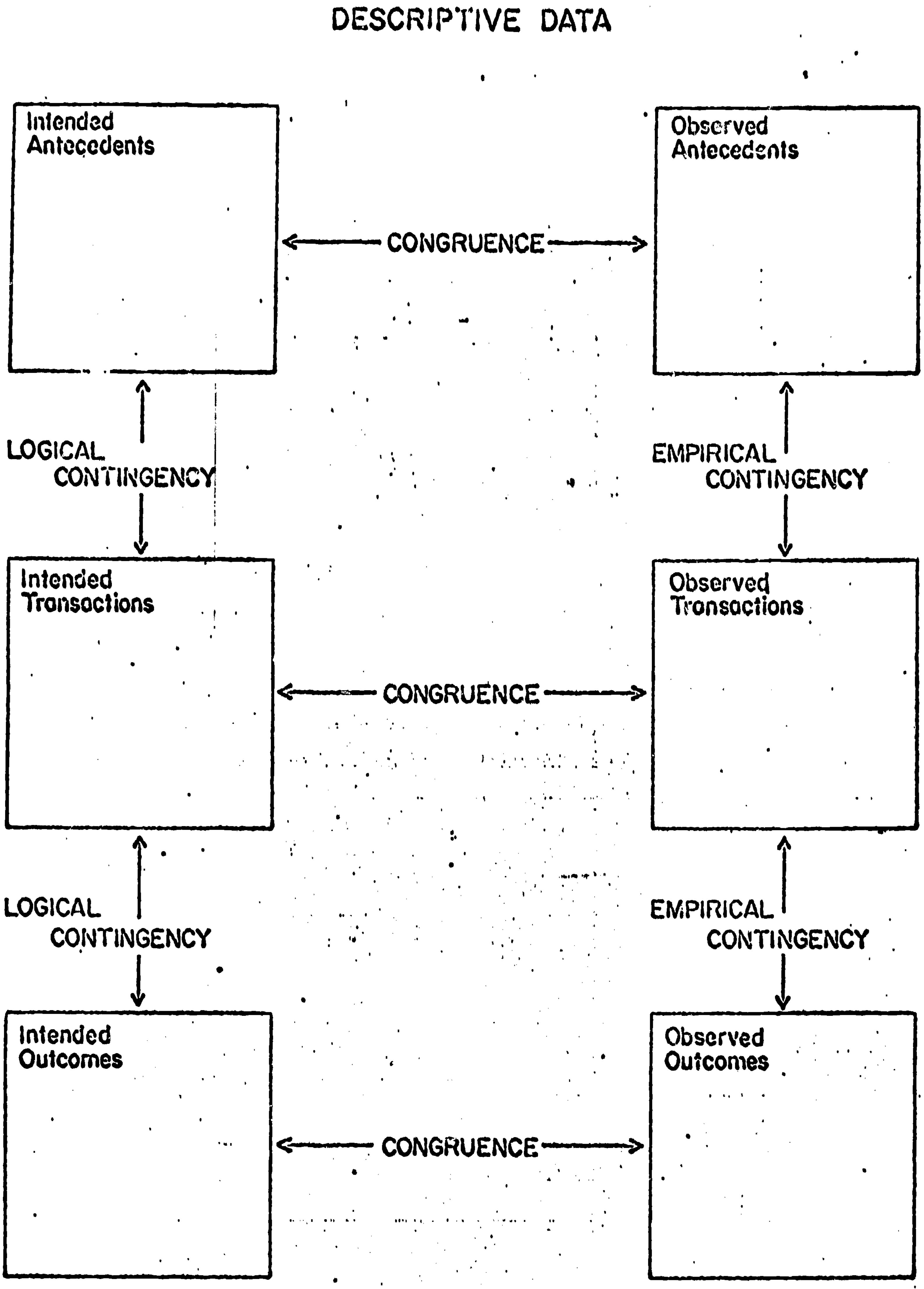

Figure 2. A representation of the processing of descriptive data.

Just as the Gestaltist found more to the whole than the sum of its parts, the evaluator studying variables from any two of the three cells in a column of the data matrix finds more to describe than the variables themselves. The relationships or contingencies among the variables deserve additional attention. In the sense that evaluation is the search for relationships that permit the improvement of education, the evaluator's task is one of identifying outcomes that are contingent upon particular antecedent conditions and instructional transactions.

Lesson planning and curriculum revision through the years has been built upon faith in certain contingencies. Day to day, the master teacher arranges his presentation and selects his input materials to fit his instructional goals. For him the contingencies, in the main, are logical, intuitive, and supported by a history of satisfactions and endorsements. Even the master teacher and certainly less-experienced teachers need to bring their intuited contingencies under the scrutiny of appropriate juries.

As a first step in evaluation it is important just to record them. A film on floodwaters may be scheduled (intended transaction) to expose students to a background to conservation legislation (intended outcome). Of those who know both subject matter and pedagogy, we ask, "Is there a logical connection between this event and this purpose?" If so, a logical contingency exists between these two Intents. The record should show it.

Whenever Intents are evaluated the contingency criterion is one of logic. To test the logic of an educational contingency the evaluators rely on previous experience, perhaps on research experience, with similar observables. No immediate observation of these variables, however, is necessary to test the strength of the contingencies among Intents.

Evaluation of Observation contingencies depends on empirical evidence. To say, "this arithmetic class progressed rapidly because the teacher was somewhat but not too sophisticated in mathematics" demands empirical data, either from within the evaluation or from the research literature.[12] The usual evaluation of a single program will not alone provide the data necessary for contingency statements. Here too, then, previous experience with similar observables is a basic qualification of the evaluator.

The contingencies and congruences identified by evaluators are subject to judgment by experts and participants just as more unitary descriptive data are. The importance of noncongruence will vary with different viewpoints. The school superintendent and the school counselor may disagree as to the importance of a cancellation of the scheduled lessons on sex hygiene in the health class. As an example of judging contingencies, the degree to which teacher morale is contingent on the length of the school day may be deemed cause enough to abandon an early morning class by one judge and not another. Perceptions of importance of congruence and contingency deserve the evaluator's careful attention.

Standards and Judgments

There is a general agreement that the goal of education is excellence--but how schools and students should excel, and at what sacrifice, will always be debated. Whether goals are local or national, the measurement of excellence requires explicit rather than implicit standards.

Today's educational programs are not subjected to "standard-oriented" evaluation. This is not to say that schools lack in aspiration or accomplishment. It is to say that standards--benchmarks of performance having widespread reference value--are not in common use. Schools across the nation may use the same evaluation checklist** but the interpretations of the checklisted data are couched in inexplicit, personal terms. Even in an informal way, no school can evaluate the impact of its program without knowledge of what other schools are doing in pursuit of similar objectives. Unfortunately, many educators are loathe to accumulate that knowledge systematically.[13, 14]

There is little knowledge anywhere today of the quality of a student's education. School grades are based on the private criteria and standards of the individual teacher. Most "standardized" test scores tell where an examinee performing "psychometrically useful" tasks stands with regard to a reference group, rather than the level of competence at which he performs essential scholastic tasks. Although most teachers are competent to teach their subject matter and to spot learning difficulties, few have the ability to describe a student's command over his intellectual environment. Neither school grades nor standardized test scores nor the candid opinions of teachers are very informative as to the excellence of students.

Even when measurements are effectively interpreted, evaluation is complicated by a multiplicity of standards. Standards vary from student to student, from instructor to instructor, and from reference group to reference group. This is not wrong. In a healthy society, different parties have different standards. Part of the responsibility of evaluation is to make known which standards are held by whom.

It was implied much earlier that it is reasonable to expect change in an educator's Intents over a period of time. This is to say that he will change both his criteria and his standards during instruction. While a curriculum is being developed and disseminated, even the major classes of criteria vary. In their analysis of nationwide assimilation of new educational programs, Clark and Guba[15] identified eight stages of change through which new programs go. For each stage they identified special criteria (each with its own standards) on which the program should be evaluated before it advances to another stage. Each of their criteria deserves elaboration, but here it is merely noted that there are quite different criteria at each successive curriculum-development stage.

Informal evaluation tends to leave criteria unspecified. Formal evaluation is more specific. But it seems the more careful the evaluation, the fewer the criteria; and the more carefully the criteria are specified, the less the concern given to standards of acceptability. It is a great misfortune that the best trained evaluators have been looking at education with a microscope rather than with a panoramic view finder.

**One contemporary check list is Evaluative Criteria, a document published by the National Study of Secondary School Evaluation (1960). It is a commendably thorough list of antecedents and possible transactions, organized mostly by subject-matter offerings. Surely it is valuable as a check list, identifying neglected areas. Its great value may be a catalyst, hastening the maturity of a developing curriculum. However, it can be of only limited value in evaluating, for it guides neither the measurement nor the interpretation of measurement. By intent, it deals with criteria (what variables to consider) and leaves the matter of standards (what ratings to consider as meritorious) to the conjecture of the individual observer.

There is no clear picture of what any school or any curriculum project is accomplishing today partly because the methodology of processing judgments is inadequate. What little formal evaluation there is is attentive to too few criteria, overly tolerant of implicit standards, and ignores the advantage of relative comparisons. More needs to be said about relative and absolute standards.

Comparing and Judging

There are two bases of judging the characteristics of a program, (1) with respect to absolute standards as reflected by personal judgments and (2) with respect to relative standards as reflected by characteristics of alternate programs. One can evaluate SMSG mathematics with respect to opinions of what a mathematics curriculum should be or with regard to what other mathematics curricula are. The evaluator's comparisons and judgments are symbolized in Figure 3. The upper left matrix represents the data matrix from Figure 2. At the upper right are sets of standards by which a program can be judged in an absolute sense. There are multiple sets because there may be numerous reference groups or points of view. The several matrices at the lower left represent several alternate programs to which the one being evaluated can be compared.

Each set of absolute standards, if formalized, would indicate acceptable and meritorious levels for antecedents, transactions, and outcomes. So far I have been talking about setting standards, not about judging. Before making a judgment the evaluator determines whether or not each standard is met. Unavailable standards must be estimated. The judging act itself is deciding which set of standards to heed. More precisely, judging is assigning a weight, an importance, to each set of standards. Rational judgment in educational evaluation is a decision as to how much to pay attention to the standards of each reference group (point of view) in deciding whether or not to take some administrative action.[‡]

Relative comparison is accomplished in similar fashion except that the standards are taken from descriptions of other programs. It is hardly a judgmental matter to determine whether one program betters another with regard to a single characteristic, but there are many characteristics and the characteristics are not equally important. The evaluator selects which characteristics to attend to and which reference programs to compare to.

From relative judgment of a program, as well as from absolute judgment we can obtain an overall or composite rating of merit (perhaps with certain qualifying statements), a rating to be used in making an educational decision. From this final act of judgment a recommendation can be composed.

Absolute and Relative Evaluation

As to which kind of evaluation--absolute or relative--to encourage, Scriven and Cronbach have disagreed. Cronbach[4] suggests that generalizations to the local-school situation from curriculum-comparing studies are sufficiently hazardous

[‡] Deciding which variables to study and deciding which standards to employ are two essentially subjective commitments in evaluation. Other acts are capable of objective treatment; only these two are beyond the reach of social science methodology.

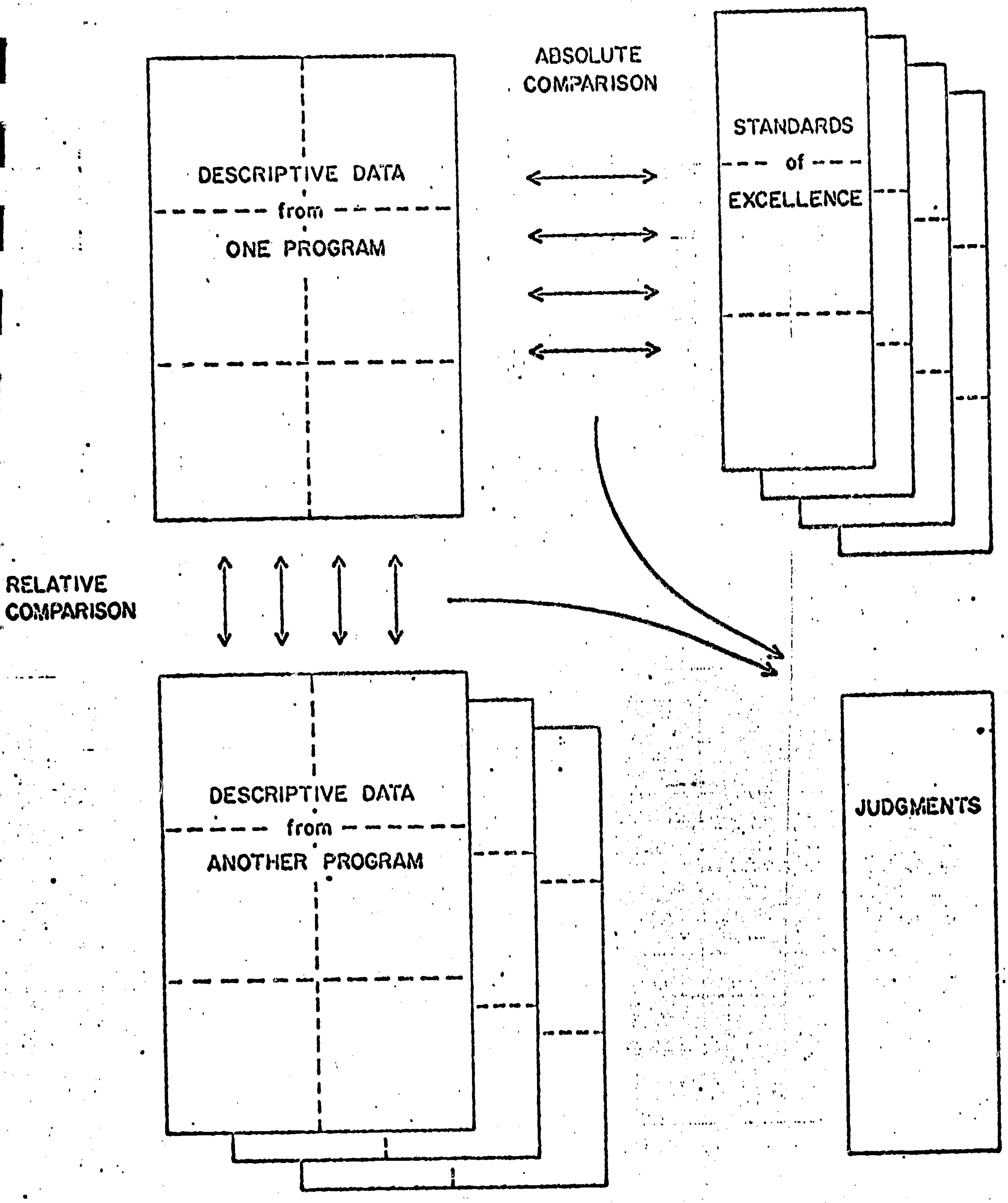
ABSOLUTE COMPARISON
STANDARDS of EXCELLENCE
DESCRIPTIVE DATA from ONE PROGRAM
RELATIVE COMPARISON
DESCRIPTIVE DATA from ANOTHER PROGRAM
JUDGMENTS

Figure 3. A representation of the process of judging the merit of an educational progr

(even when the studies are massive, well-designed, and properly controlled) to make them poor research investments. Moreover, the difference in purpose of the programs being compared is likely to be sufficiently great to render uninterpretable any outcome other than across-the-board superiority of one of them. Expecting that rarely, Cronbach urges fewer comparisons, more intensive process studies, and more curriculum "case studies" with extensive measurement and thorough description.

Scriven, on the other hand, indicates that what the educator wants to know is whether or not one program is better than another, and that the best way to answer his question is by direct comparison. He points to the difficulty of describing the outcomes of complex learning in explicit terms and with respect to absolute standards, and to the ease of observing relative outcomes from two programs. Whether or not Scriven's prescription is satisfying will probably depend on the client. An educator faced with an adoption decision is more likely to be satisfied, the curriculum innovator and instructional technologist less likely.

One of the major distinctions in evaluation is that which Scriven identifies as formative versus summative evaluation. His use of the terms relates primarily to the stage of development of curricular material. If material is not yet ready for distribution to classroom teachers, then its evaluation is formative; otherwise it is summative. It is probably more useful to distinguish between evaluation oriented to developer-author-publisher criteria and standards and evaluation oriented to consumer-administrator-teacher criteria and standards. The formative-summative distinction could be so defined, and I will use the terms in that way. The faculty committee facing an adoption choice asks, "Which is best? Which will do the job best?" The course developer, following Cronbach's advice, asks, "How can we teach it better?" (Note that neither are now concerned about the individual student differences.) The luator looks at different data and invokes different standards to answer these questions.'

The evaluator who assumes responsibility for summative evaluation--rather than formative evaluation--accepts the responsibility of informing consumers as to the merit of the program. The judgments of Figure 3 are his target. It is likely that he will attempt to describe the school situations in which the procedures or materials may be used. He may see his task as one of indicating the goodness-of-fit of an available curriculum to an existing school program. He must learn whether or not the intended antecedents, transactions, and outcomes for the curriculum are consistent with the resources, standards, and goals of the school. This may require as much attention to the school as to the new curriculum.

The formative evaluator, on the other hand, is more interested in the contingencies indicated in Figure 2. He will look for covariations within the evaluation study, and across studies, as a basis for guiding the development of present or future programs.

For major evaluation activities it is obvious that an individual evaluator will not have the many competencies required. A team of social scientists is needed for many assignments. It is reasonable to suppose that such teams will include specialists in instructional technology, specialists in psychometric testing and scaling, specialists in research design and analysis, and specialists in dissemination of information. Curricular innovation is sure to have deep and widespread effect on our society, and we may include the social anthropologist on some evaluation teams. The economist and philosopher have something to offer. Experts will be needed for the study of values, population surveys, and content-oriented data-reduction techniques.

The educator who has looked disconsolate when scheduled for evaluation will look aghast at the prospect of a team of evaluators invading his school. How can these evaluators observe or describe the natural state of education when their very presence influences that state? His concern is justified. Measurement activity--just the presence of evaluators--does have a reactive effect on education, sometimes beneficial and sometimes not--but in either case contributing to the atypicality of the sessions. There are specialists, however, who anticipate that evaluation will one day be so skilled that it properly will be considered "unobtrusive measurement."[16]

In conclusion I would remind the reader that one of the largest investments being made in U.S. education today is in the development of new programs. School officials cannot yet revise a curriculum on rational grounds, and the needed evaluation is not under way. What is to be gained from the enormous effort of the innovators of the 1960's if in the 1970's there are no evaluation records? Both the new innovator and the new teacher need to know. Folklore is not a sufficient repository. In our data banks we should document the causes and effects, the congruence of intent and accomplishment, and the panorama of judgments of those concerned. Such records should be kept to promote educational action, not obstruct it. The countenance of evaluation should be one of data gathering that leads to decision-making, not to trouble-making.

Educators should be making their own evaluations more deliberate, more formal. Those who will--whether in their classrooms or on national panels--can hope to clarify their responsibility by answering each of the following questions: (1) Is this evaluation to be primarily descriptive, primarily judgmental, or both descriptive and judgmental? (2) Is this evaluation to emphasize the antecedent conditions, the transactions, or the outcomes alone, or a combination of these, or their functional contingencies? (3) Is this evaluation to indicate the congruence between what is intended and what occurs? (4) Is this evaluation to be undertaken within a single program or as a comparison between two or more curricular programs? (5) Is this evaluation intended more to further the development of curricula or to help choose among available curricula? With these questions answered, the restrictive effects of incomplete guidelines and inappropriate countenances are more easily avoided.

REFERENCES

[1]American Council on Education. Educational Measurement. E. F. Lindquist (Ed.). Washington, D. C., 1951.

[2]Smith, E. R. and Tyler, Ralph W. Appraising and Recording Student Progress. New York: Harper and Row, 1942.

[3]Educational Testing Service. "A Long, Hot Summer of Committee Work on National Assessment of Education," ETS Developments, Vol. XIII, November, 1965.

[4]Cronbach, Lee. "Evaluation for Course Improvement," Teachers College Record, 64, 1963, pp. 672-683.

[5]Scriven, Michael. "The Methodology of Evaluation," AERA Monograph Series on Curriculum Evaluation, No. 1. Chicago: Rand McNally, 1967, pp.39-89.

[6]Taylor, Peter A. and Maguire, Thomas O. "A Theoretical Evaluation Model," The Manitoba Journal of Educational Research, I, 1966, pp.12-17.

[7]Smith, B. Othanel and Meux, M. O. A Study of the Logic of Teaching. Urbana: Bureau of Educational Research, University of Illinois. No date.

[8]Gagné, Robert M. "Elementary Science: A New Scheme of Instruction," Science, Vol. 151, No. 3706, pp. 49-53.

[9]Atkin, J. M., "Some Evaluation Problems in a Course Content Improvement Project," Journal of Research in Science Teaching, I, 1963, pp. 129-132

[10]Mager, R. F. Preparing Objectives For Programmed Instruction. San Francisco: Fearon Publishers, 1962.

[11]Berlak, Harold. Comments recorded in Concepts and Structure in the New Social Science Curricula. Irving Morrissett (Ed.). Lafayette, Indiana: Social Science Education Consortium, Purdue University, 1966, pp. 88-89.

[12]See Bassham, H. "Teacher Understanding and Pupil Efficiency in Mathematics: A Study of Relationship," Arithmetic Teacher, 9: 1962, pp. 383-387.

[13]Hand, Harold C. "National Assessment Viewed as the Camel's Nose" Phi Delta Kappan, 47, September, 1965, pp. 8-12.

[14]Tyler, Ralph W. "Assessing the Progress of Education," Phi Delta Kappan, 47, September, 1965, pp. 13-16.

[15]Clark, David L. and Guba, Egon G. "An Examination of Potential Change Roles in Education." Columbus: The Ohio State University, 1965. (Multilith).

[16]Webb, Eugene J., Campbell, Donald T., Schwartz, Richard D., and Sechrist, Lee. Unobtrusive Measures: Nonreactive Research in the Social Sciences. Chicago: Rand McNally, 1966.

EVALUATION INSTITUTE: "Gifted Program"
July 29 - August 9, 1968

This preliminary material has been prepared to provide participants with an overview of the workshop program and relevant detail about its continuity and content.

During the first week, all participants (either individually or in small groups) will work on developing a plan to collect information on some specific question. Presentations, during the initial week, will deal with ideas or concepts related to the problem.

During the second week participants will be working with individual problems. Presentations will deal with the development of skills that an evaluator uses to solve such problems.

The following descriptions of the daily and sequential activities project specific foci and suggestions you may wish to follow in determining the evaluation plan.

WEEK ONE

Monday

The morning sessions will be for introductions and an orientation to the workshop. In these sessions we hope to elicit from you the general and specific evaluation problems that concern you and the expectations you have of the workshop.

Bob Stake will make the first of two presentations of his evaluation model in the first afternoon session. These presentations (the second is on Tuesday a.m.) will provide you with a general overview of the model. Subsequent presentations will be relevant to specific components of the model, and they will assist you in translating the model into a specific plan. You should have read the article "The Countenance of Educational Evaluation" before hearing the presentations by Dr. Stake.

Three videotapes of consultation sessions between a school person and an evaluation expert have been prepared. Each tape projects a session that was held prior to the development of one of the three evaluation-plan examples. Showing of one of these tapes has been scheduled for the fourth session on Monday.

Tuesday

A continuation of the presentation on the Stake model will come in the first Tuesday session.

The second session will be a work session during which you can begin work on your evaluation plan You may want to study the example plans at this time.

Material of relevance to the observations cells of the model will be presented in the third session. Topics that will be considered in this session include operational definitions and principles of test selection.

The first part of the fourth session on Tuesday will be a presentation on the use of certain resource materials such as The Mental Measurements Yearbook, Research in Education, and Review of Educational Research. The remaining time in the session will be a work session.

Tuesday's fifth session is planned as the time when Dr. Stake will use one of the videotapes to discuss the role and task of the consultant.

Wednesday

A second presentation on conditions of observation is scheduled for the first Wednesday session. Dr. Denny will discuss classroom observation procedures.

The second and third sessions are scheduled as work sessions. Hopefully by the end of the third session you will have defined a rough outline of your evaluation plan. The rest of the videotapes will be shown on a schedule from 10:15 - 3:00.

The format of evaluation reports will be presented in the fourth session.

Thursday

The first session will be devoted to a presentation on ways to establish standards as bases for judgments. The content will be primarily on research designs that might be used in evaluation situations. The validity of the data obtained in the various designs will be stressed.

The second session will be a presentation on observational procedures not commonly used but with rich potential. The topic is unobtrusive measures.

The third and fourth sessions are planned as work sessions. Hopefully your plans will be ready to be submitted to the staff at the end of session four on Thursday The staff will develop some artificial data for your plan on Thursday night. These data will be used by you in the Friday exercise. Your plans will also be read by the staff and feedback provided you by Monday.

The fifth session will be a presentation on statistical problems. This presentation and the first one on Friday will provide an overview of properties of scales and of certain statistical techniques.

Friday

The first session is planned as a presentation on statistics.

The second session is a work session in which you will analyze the data provided you and prepare a report of your evaluation. Much of the report will have been written as the material prepared during the week. The reports will be duplicated over the weekend and distributed on Monday.

The third session is scheduled as a time for evaluation of the first week's activities in the workshop. You will be asked to complete some evaluation instruments, but we also hope you will discuss the strengths and weaknesses of the week as you perceived them. The planned schedule for the second week will be presented and perhaps revised on the basis of your expressed interests and desires.

WEEK TWO

Workshop Exercises

The work sessions in the second week have been set aside for you to work on a problem or problems that are of immediate concern to you. Some of you may want to develop a plan for gathering information on another aspect of your program than the one covered in the first week. An exercise like this would be desirable in the sense that it would add to your general evaluation plan for your program.

Another possibility is to develop one or more of the instruments that you expect to use in your evaluation. The instrument could be a questionnaire or interview schedule, rating scales, an attitude scale, or an achievement test.

Others may want some practice working with some of the statistical techniques. Exercises in this area are available for you to work on.

A list of suggested activities has been prepared and materials have been written that will be helpful for your work on the activity. You should not feel restricted to working on the suggested activities, however. Obviously some of the exercises will take more time than others, or some participants may work on three or four things during the week and some on one or two.

You will notice in the Wednesday and Thursday session that an interviewing exercise may be assigned. This will limit the amount of time that you will have on Thursday to work on your own problems.

The presentations in the second week are designed to develop skill and understanding in working with some common evaluation problems. Any one presentation may have only a peripheral relationship to the topic or problem on which you are working, but the topic is of central concern to the task of the evaluator in general.

Monday

The first session on Monday is an orientation to the second week. This will be needed especially if changes are suggested by points you raise in the Friday evaluation session. If time permits, part of the first session will be available for you to think about your individual activity for the second week. Feedback on your plans may also be provided.

The second session is planned as a presentation on judgments. Techniques for making judgments will be covered.

The third session is planned as a panel to bring up and discuss problems that confront an evaluator such as inadequate questionnaire returns, administrator interference, etc.

Tuesday

The Tuesday presentations are on test construction. Topics to be covered are measuring achievement, measuring higher-order mental processes, measuring attitudes and scaling. Principles of test construction will be stressed. The presentations will be in the first, third, and fifth sessions. The second and fourth sessions will be work sessions.

Wednesday

The first session on Wednesday is planned as an overview of survey procedures. The presentation will include material on kinds of information usually obtained in surveys and sampling considerations.

The construction of questionnaires and interview schedules is the topic assigned to the second Wednesday session.

The Wednesday afternoon and evening sessions are scheduled for an interview training exercise. Dr. Denny will conduct these sessions and will structure the situation for you.

Thursday

The first two sessions on Thursday are planned as work sessions.

The Thursday afternoon sessions are planned for the second part of the interview training activities. The content of these sessions will follow up the Wednesday afternoon presentation.

Friday

In the first session on Friday we have planned a critique and discussion of the evaluation-plan handout on the question "Has the gifted program had an effect on the achievement of the participating students?"

Dr. House will present some ideas on the establishment of an information pool in the second session. The information pool would be a central storage and clearinghouse of information on the gifted programs in the state.

The third Friday session is planned as an evaluation session and one in which the end-of-workshop administrative details are handled. The workshop will close at the end of this session.

You should regard the work sessions as your time to work on your problem in your own style. You may want to talk with staff members, read, talk with other participants, take a walk, go to the library, take a nap, etc. The point is that these sessions are open for you to structure or not structure as you desire. We believe such an environment is conducive to your getting what you need and want from the workshop.

WORKSHOP EXERCISE
(WEEK ONE)

The workshop is structured so that in the first week all participants will develop a plan for evaluating a specific aspect of their gifted program. This may be done individually or in small groups. Although it is anticipated that you will ultimately develop and institute an over-all evaluation plan for your program, such an ambitious project is not realistic for most of you for this first activity. We believe you will attain greater satisfaction from the activity if you complete a plan for a limited problem than if you get hung-up in working on a large-scale problem.

This should not be interpreted, however, that you cannot work on a general evaluation problem should you wish to do so. The scope of your selected problem will depend much on how much you already have done on evaluation. Should you complete a plan before the end of the week, you certainly may work on a second plan or you might work on some of the activities planned for the second week of the workshop.

Your evaluation plan should be focused at a situation in your school. It should also be a plan that can be done given the time, staff, and financial situation as it exists in your school.

We have developed three evaluation plans that you might use as resource material. These plans were developed to obtain evaluative information on three rather specific but common questions. Each plan was in the context of a unique school situation. Each of the plans was developed by first consulting with an expert on evaluation about the problem and then making the plan using his advice.

The consulting sessions were recorded on videotape. We have scheduled showings of these tapes during the first week. Observation of the videotapes should help clarify for you some of the important factors to consider in planning for an evaluation. You should view the tapes and then as you study the plans you will be aware of and understand the reasons for the procedures in the plan.

The three evaluation plans were developed for the following problems:

1. Does the gifted program in our school increase the students' ability to conduct independent study?

2. Are we selecting the right students for our class in creative writing?

3. Which of these three laboratory manuals should we use in our elementary science course?

The following list contains other specific problems that you might select for your week's work. The list is intended to be suggestive of the kinds and the scope of the problems. You should not feel you are restricted to something from this list, however.

1. What is the attitude of the people in the community toward a program for the gifted and our program in particular?

2. Do the students in the gifted program become isolated from other students in the school?

3. How well do the graduates of our gifted program do in college or other post high school activity?

4. How well have the students in this course learned the material covered?

5. Has the in-service program for the teachers of the gifted made them better teachers?

6. How useful are these materials for teaching these concepts?

7. Has this course helped develop a general problem solving ability in the student?

8. Has this literature course affected the student's attitude toward literature?

9. What are the occupational aspirations of the students in the gifted program?

10. How well do the students in the gifted program perform in their other classes?

Several kinds of resource material are available for you to use as you work on the problem. Among these are the videotapes and their associated plans, the books and articles on the reference list and available in the meeting room, other books and articles in the University Library, the staff, and the other participants. We would especially emphasize your use of each other as resources. There are at least three ways in which you can be very helpful to each other.

First, you can serve as reactors to one another's ideas. This is sort of a general "Here's what I've done, what do you think of it?" kind of role.

Second, each of you has had unique experiences which give you certain unique knowledge or skill. Some of you are administrators, others are

English teachers, some have knowledge about statistics, some have worked with evaluation, and so on. You will soon associate certain people in the group with certain competencies. Use these people as sources of information as well as reactors.

The third way you can interact with each other is in role playing kinds of situations. As you develop your plan ask others to play certain roles in reacting to your plan. For example, several of the participants have administrative positions. You might ask one of these people to question you about your plan from the administrator's point of view. Or you might be anticipating some resistance from a teacher in your school to some of the procedures you are planning. Ask one of the participants who is a teacher to play the role of a teacher who is resisting any kind of evaluation activity. Other roles that might be relevant would be an irate parent, an interested parent, a consultant, a student, a state department representative, or other kinds of people whom you feel would have some influence on the success of your evaluation effort. The questions and comments of the role player should be useful for you in identifying aspects of your plan that may be resisted or which need clarification.

An important benefit of your consulting and role playing with each other will be the knowledge you will gain about being a consultant. When you return to your school you will be more able to provide advice and consultation to members of your staff and to people from other schools on evaluation problems.

We hope that you will be able to have a plan done by Thursday evening. We will generate some data for you on the basis of your plan so that you can make a report on your evaluation. We intend to reproduce all of the reports and distribute them so that you can take these home with you and perhaps use them as reference material.

WORKSHOP EXERCISE
(WEEK TWO)

There are a number of things that you might do in the work sessions of the second week. An obviously desirable activity would be to develop plans for gathering other evaluative information than included in the problem of the first week. We have written some exercises that you may want to work on during the second week. These exercises are listed below.

1. Building rating scales

2. Building an attitude scale

3. A problem on each of the following statistical techniques:
   a. Chi squared
   b. Pearson coefficient of correlation (Pearson r)
   c. Spearman rank-order coefficient of correlation (Spearman rho)
   d. t-test and Mann-Whitney "U"
   e. Correlated t-test
   f. Analysis of variance

There are enough copies of each of the exercises so that you may have a copy of each and work on them after the workshop. If you have questions on the exercises, feel free to ask for help or clarification.

The following items are other suggested activities. We have not developed any materials for these, however.

1. Build an achievement test.
2. Build a questionnaire or interview schedule.
3. Study a number of evaluation models. (References for many of these are included in the reference list.)
4. Develop a scale such as scaling the importance of program objectives.
5. Review research and other writings on some topic.
6. Study several methods of classroom observation.

If you work on some special topic, we hope you will prepare a report of your work for distribution to the other participants.

	Monday	Tuesday	Wednesday	Thursday	Friday	Monday	Tuesday	Wednesday	Thursday	Friday
First Session 8:30-10:00	Orientation and Setting	Evaluation Model Stake	Classroom Observation Denny	Research Design Sjogren	Statistical Problems Sjogren	Second Week Orientation	Measuring Cognitive Outcomes Hastings	Survey Procedures Denny	Work Session	Evaluat Plan Critiqu Hasting
Second Session 10:15-11:45	Expectation	Work Session	Work Session (video tapes)	Unobtrusive Measures	Work Session	Judgments Stake	Work Session	Question-naire Construction Stake	Work Session	State Program
Lunch 11:45-1:30										
Third Session 1:30-3:00	Evaluation Model Stake	Observation and Testing Denny	Work Session	Work Session	Work Session	Panel on Evaluation Problems	Test Development Stake			Evaluat and Bookkee
Fourth Session 3:15-4:45	------- Videotape	Resource Material ------- Work Session	Report Format Stake	Work Session	Evaluation	Work Session	Work Session	Interview Training Denny	Interview Training Denny	End
Social and Dinner 5:00-7:30										
Fifth Session 7:30-9:00	Individual Work Session	Reactor Role Videotape Stake	Work Session	Statistical Problems Sjogren		Work Session	Scaling Stake			

VIEWING THE VIDEOTAPES

Each of the three videotapes presents a public school person's consultation with an evaluation expert about a specific evaluation problem.

Before viewing a tape, please read the first page of the corresponding evaluation plan that you may anticipate the situation and identify significant things to observe in the interview. (Each tape and corresponding plan are identically numbered.)

Primarily, you will find the videotapes helpful in emphasizing matters you must consider as you make an evaluation plan.

The tapes also project approaches and techniques you may find useful when other schools ask you to consult with them about their evaluation efforts. If possible, view all three tapes for each consultant preforms his unique and demanding role in his own style. Notice how each establishes rapport, generally does not judge, and always sensitively perceives the other person's response to a potentially threatening line of questioning.

Questions for Videotape

1. What is the problem?

2. For whom is the information intended?

3. What types of data are suggested for solving the problem?

4. What other data could be gathered to solve the problem?

Daniel L. Stufflebeam
January 1968

DEVELOPING EVALUATION DESIGNS

The logical structure of evaluation design is the same for all types of evaluation, whether context, input, process or product evaluation. The parts, briefly, are as follows:

A. Focusing the Evaluation
   1. Identify the major level(s) of decision making to be served, e.g., local, state or national.
   2. For each level of decision-making, project the decision situations to be served and describe each one in terms of its locus, focus, criticality, timing and composition of alternatives.
   3. Define criteria for each decision situation by specifying variables for measurement and standards for use in the judgment of alternatives.
   4. Define policies within which the evaluation must operate.

B. Collection of Information
   1. Specify the source of the information to be collected.
   2. Specify the instruments and methods for collecting the needed information.
   3. Specify the sampling procedure to be employed.
   4. Specify the conditions and schedule for information collection.
   5. Specify the definition of each item of information.

C. Organization of Information
   1. Provide a format for the information which is to be collected.
   2. Designate a means for coding, organizing, storing, and retrieving information.

D. Analysis of Information
   1. Select the analytical procedures to be employed.
   2. Designate a means for performing the analysis.

E. Reporting of Information
   1. Define the audiences for the evaluation reports.
   2. Specify means for providing information to the audiences.
   3. Specify the format for evaluation reports and/or reporting sessions.
   4. Schedule the reporting of information.

F. Administration of the Evaluation
   1. Summarize the evaluation schedule.
   2. Define staff and resource requirements and plans for meeting these requirements.
   3. Specify means for meeting policy requirements for conduct of the evaluation.
   4. Evaluate the potential of the evaluation design for providing information which is valid, reliable, credible, timely and pervasive.
   5. Specify and schedule means for periodic updating of the evaluation design.
   6. Provide a budget for the total evaluation program.

EVALUATION PLAN I

Problem: Does the gifted program in our school increase the student's ability to conduct independent study?

School situation:

In this high school, enrolling approximately 800 students, the gifted program offers special classes in 11th and 12th grade English and science.

In each grade, approximately 40 students enroll in the gifted program - usually for two years - but not every student takes both English and science.

One English and two science teachers are involved in the gifted program. The English teacher's schedule includes two special classes, three other classes, and one planning period. The science teachers' schedules include one special class, four other classes, and one planning period.

Problem situation:

The teachers in the gifted program would like to find out how the special classes affect their students. To determine this outcome, they select a significant objective - the development of a student's capacity for independent study - as an important aspect for evaluation.

The principal concurs and agrees to purchase some tests and provide a limited amount of clerical time. However, the teachers must do the evaluation on their own time and consider the task a part of their duty. One of the school's counselors consents to help with the project.

Videotape:

In this tape, Bob Stake is consulting with the teacher. As you view the conference, you will observe that the consultant is attempting to help the teacher clarify the evaluation problem. His questioning and probing take two tacks. One approach communicates the meaning of independent study; the other involves posing questions asking, "Is this all you really want or need to know?"

Note how, towards the conclusion of the tape, Stake stresses the point that planning an evaluation resembles planning any activity because decisions about priorities and alternatives must be made and governed by constraints that the inevitable limitations on resources always impose.

Evaluation Plan:

The following plan, made after the consultation, was developed for the "problem question" stated on page 1. One of an infinite variety that might have been drafted, this plan neither should be judged right or wrong nor regarded as an optimal plan.

Rationale

This school is committed to the belief that every student in the school should be provided maximum opportunity to develop his abilities and interests to the fullest extent. In order that the student may achieve this goal, the school necessarily must use all its available resources to offer its students the maximum number and variety of programs, materials and resources.

Among the students in this school, those who have special intellectual talents will benefit - as will society - from a program designed to develop their talents. Furthermore, the number of these gifted studente warrants the scheduling of special classes which will not increase the teacher-student ratio in other classes nor in any way impair the other students' educational opportunity and experience.

Purpose

The primary purpose of this evaluation is to determine whether students who participate in the special classes develop their ability to conduct independent study.

Predicated on the assumption that intellectually endowed students can work independently and learn to identify and work on new ideas in a conducive environment, a primary objective of the program is the development of this capacity that will maximize the student's productivity and potential contribution to society.

A second purpose is to attempt to identify characteristics of students who differ in their capacity for independent study when the program terminates.

A third purpose is to attempt to identify activities in the program or the school that enhance or inhibit a student's attainment of this objective.

Procedures

The following procedures will be used to measure the student's capacity for independent study.

1. Rating performance

   In each class, students will be given two intentionally ambiguous assignments and their performance rated: the first, during the first six weeks and the second, during the last six weeks of the year.

   Teachers will rate these assignments on such scales as:
   a. Amount of structure requested from the teacher
   b. Number of requests for help from the teacher
   c. Number of resources used
   d. Amount of interpretation included in the report
   e. Adequacy of procedures used as described in the report

   (To assure validity as they rate the second assignment, teachers will not have access to the initial ratings which will have been filed.)

   In the middle of the year, other appropriate teachers of the gifted students will be asked to complete rating scales indicating their impression of the student's sustained capacity for independent study in the class rather than in doing a specific assignment.

   In order to compare the gifted students with other students in their class, these teachers will randomly select an identical number of students and rate their capacity for independent study.

2. Identifying characteristics

   From the cumulative folder, characteristics including I.Q., sex, age, position in family, size and socio-economic status of the family will be obtained. These variables then will be related to the student's scores on the rating scales.

3. Identifying program activities

During the last six weeks, a counselor will interview students about the work they accomplished, the ways they would solve a problem, their capacity to work independently.

4. Reporting the program

At the end of the year, a report incorporating tabulated and analyzed data will be written. To provide meaningful context, the report will include a detailed description of the school and the community.

Time schedule

August 15 - September 15:	work on assignments and rating scales.
September 15 - November 1:	complete assignments and fill out rating scales.
November 1 - December 31:	develop interview schedule.
January 1 - February 28:	have other teachers fill out rating scales and collect data on student characteristics.
March 1 - March 31:	collect information on school and community.
April 1 - May 31:	complete second assignment and rating scales. Counselor interviews.
June 1 - June 30:	analyze data and write report.

EVALUATION PLAN II

Problem: Which of these three sets of science materials should we use in our elementary science course?

School situation:

In the school district, there are 30 K-6 elementary schools, most of which are four unit. Six of the schools offer gifted programs essentially based on grouping in the 5th and 6th grades and providing enrichment experiences in science, social studies and language arts.

Problem situation:

Several teachers have expressed dissatisfaction with the materials they are using in their science classes. Because of their limited backgrounds in science, the teachers feel insecure about doing the experiments which require materials that are difficult to procure and often are impracticable.

Aware of the teachers' concerns, the elementary coordinator has searched for and found three sets of suitable materials. However, he must make a final decision about selecting only one of the sets for general use because a school board policy stipulates that all schools must use the same basic materials. The coordinator is allowed and encouraged to purchase and try out materials. Thus, he wants to evaluate the three sets during the coming year before deciding which set to purchase for general use.

Videotape:

In this tape, Terry Denny functions as a resource person for the project, a role consultants often perform. Observe how Denny identifies with the project and offers much assistance. Note that he stresses the wide spectrum of data that may be used in evaluating materials. The way he uses the evaluation matrix (drawn on the board) effectively initiates such a project.

EVALUATION PLAN:

The following plan, developed after the consulting session, is only one of many that might have been developed and thus should not be considered the optimal plan.

Rationale

In a viable democratic society, each member must assume responsibility for participating in and contributing to that society. In order to effectively participate, each member must be educated and his special talents and abilities developed as much as possible.

This school believes that an individual's talent (music, art, physical skills, intellectual capacity) should be identified as early as possible and special programs instituted to develop such talent. Accordingly, in six of the district's elementary schools special 5th and 6th grade classes in science, social studies and language arts have been programmed for the intellectually gifted.

Purpose

The purpose of this study is to obtain information about elementary school science materials in order to decide what materials should be purchased for use in accelerated 5th and 6th grade classes.

Procedures

Although further review of the science materials might be desirable, only the three sets selected for try-out can be evaluated during the necessarily limited time for deciding which set should be purchased.

Information about these materials will be obtained from:

1. Producers of the materials

   a. Rationale describing objectives of materials and specific relevance of activities and problems

   b. Information about cost, durability, plans for future revision

   c. List of schools which have adopted the materials

2. Experts in science curriculum

   a. A literature search for reviews

   b. Request EPIE (Educational Products Information Exchange) for copies or sources of such reviews

   c. Request science methods' professors in the local College of Education to review or recommend someone to review the three sets of materials

3. Teachers

   a. Questionnaire

      Teachers in 20 schools, randomly selected from each list provided by the producers, will be asked:

      (1) Durability of item
      (2) Kinds of resource material needed
      (3) Science background of the teacher
      (4) Workability of experiments and exercises
      (5) Kinds of students in their classes
      (6) Rating of students' interest in materials
      (7) Indications of students' performance

   b. Trial use in classrooms

      (1) Each of the three sets randomly assigned to two classrooms during the next year

      (2) Each teacher will be asked to:

         (a) Keep a log describing both positive and negative incidents of significance that the materials motivated or created

         (b) Rate each unit on a series of rating scales and the total course when completed to determine: student performance, interest, amount of required preparation, availability of resources, feasibility of exercises and problems, level of difficulty

4. Students

a. Tests

(1) Administered as "pre" and "post" tests:

(a) Science test in California Achievement Tests
(b) Teacher-constructed test
Each teacher will write 30 items covering content in his set of materials and, after screening, a total of 120 items will be randomly incorporated in two 60-item forms. Each student will take one of these two tests.

(2) Comparison of student "pre" and "post" test performance on both tests

(3) Analysis of student performance on individual items in teacher-constructed test to determine extent to which specific learnings among the sets of materials differ.

b. Questionnaire

At the end of each unit, students will be asked to complete a brief questionnaire indicating:

(1) Interest in the unit
(2) Opinion about "difficulty" of the unit
(3) Opinion about "How much they learned"
(4) Specific things they liked and disliked in the materials

Time Schedule

August 15 - September 30: Prepare achievement test. Assign materials to teachers. Meet with teachers to discuss evaluation. Make log report forms, rating scales, and questionnaires.
October 1 - October 31: Administer achievement tests. Contact producers for information.
November 1 - December 31: Get experts' opinions. Send out questionnaires to other users. Remember to get logs and end of unit ratings.
January 1 - May 1: Analyze and synthesize information as it comes in.
May 1 - May 31: Administer achievement tests. Get final ratings.
June 1 - June 30: Complete analysis of information and prepare report.
July 1 - July 8: Meet with teachers and make decision regarding materials.

EVALUATION PLAN III

Problem: Are we selecting the right students for our class in creative writing?

School situation:

The gifted program, which consists of two 12th grade English classes in literature and creative writing, respectively, is offered in one of the district's two high schools enrolling approximately 1200 students each.

Problem situation:

Currently, 11th grade English teachers identify and recommend students (generally those who have performed well in their classes) for the gifted program.

The teacher of creative writing is dissatisfied with this procedure because he does not believe that excellent performance in a traditional composition course necessarily indicates interest in or aptitude for creative writing.

In order to find out if a different procedure might more validly identify potentially "creative writers", he wants to evaluate the present selection procedures.

Videotape:

In this situation, Tom Hastings is the consultant. Observe how Hastings' comments focus the initial problem and how he stresses the necessity for the teacher to more explicitly define the problem. Note the many different kinds and sources of information the consultant suggests to the teacher.

EVALUATION PLAN:

The following plan, written after only one session with the consultant, may not be a plan that the consultant would unequivocally endorse.

Purpose

In this study, the most effective way of determining a student's potential for creative writing should qualify the attempt to identify ways to select students for the class.

Before deciding on the procedures, the principal, counselors and most of the English teachers met to discuss and clearly define the general objectives for the gifted classes. The group agreed that a class in creative writing should develop the students' creative writing skills. This goal, however, does not state the basic problem of identifying such potential.

Procedures

In order to define a student's potential for creative writing, behaviors indicative of creative writing must be defined before antecedent behavior or characteristics validly predicting writing behaviors can be determined.

I. Define creative writing skill

A. The teacher will try to state his definition of creative writing skill.

B. Some writers, poets, teachers will be asked to comment on this definition and from their suggested additions and deletions a synthesized definition will be formulated.

II. Develop a series of rating scales

Based on the definition of creative writing skill, rating scales will be developed to:

A. Rate behavior of students in creative writing class

B. Adapt, if feasible, to selecting students for next year's creative writing class

III. Administer tests

A. At the beginning of the school year, the following tests (intuitively selected) will be administered to students in the creative writing class, the literature class, and to randomly selected students in other classes in 12 th grade English:

1. Association
2. Simile Interpretation
3. Plot Titles
4. Object Synthesis
5. Alternative Uses
6. Vocabulary

B. Correlation

1. At the conclusion of the semester, students' scores on these tests will be correlated with performance scores of students in creative writing to determine if tests do predict performance in creative writing.
   (Using rating scales (see IIA) the teacher and two other persons independently rate three samples of creative writing and the raters' respective scores are averaged.)

2. Scores of students in three groups (IIA) will be compared to determine differences which - if minimal - would invalidate test - performance correlations and the effectiveness of current method of selecting students.

IV. During the second semester, these findings will be studied in order to decide whether a new selection procedure should be recommended for try-out next year.

Time Schedule:

August 15 - September 15: Define creative writing skill. Conduct a review of literature on this. Administer tests, do not score.

September 16 - October 15: Survey writers', poets', and teachers opinions of definition. Try to get at least 30 responses.

October 15 - November 30: Build rating scales. Select assignments for rating.

December 1 - January 31: Using rating scales, rate the three assignments. Three people rate each paper. Have papers typed and rate them blind. Score tests and correlate test scores with rating scale scores.

February 1 - April 1: Consider results and make decision about recommending a new selection procedure.

May 1 - May 31: Administer what is needed for new selection procedure if one is inaugurated.

SUGGESTIONS FOR ROLE PLAYING

The following questions represent those that various people might ask about the gifted program and the evaluation plan. During the workshop, other participants may ask you to react to their plans as an administrator, teacher, parent or student might; or you may consult with a participant about your plan. In either situation, such provocative questions should guide and stimulate you to phrase further questions that will reflect your thinking about the ideas and points of view these "Suggestions for Role Playing" evoke.

School Administrator

1. Will this plan require additional staff?
2. Does the plan commit us to any future action?
3. What impact might this plan have on staff relations?
4. What kinds of materials will you need to buy?
5. Will the additional testing require a significant amount of counselor time?
6. What impact or repercussions can we expect from the parents and the community when you start interviewing?
7. What can we do with the results you will get?
8. What do we do if you get negative results?
9. How do you plan to get the other teachers and staff to cooperate?
10. I am afraid this evaluation will just stir up a hornet's nest. Why should we create problems?

Teacher

1. How much class time will this take?

2. I don't think standardized tests are any good. They don't get at the things I teach.

3. How can I get any teaching done if I have to spend all of my time testing?

4. Who will do the classroom observations? How can we arrange these observations so they don't disrupt the class?

5. I am glad you are doing this. How can I help? Have you thought of getting this kind of information?

6. What do you want in the log? How long should they be?

7. Who will score all of the tests? Will I know how the students do?

8. Are you going to compare the different classes? What happens if my class doesn't do so well?

9. What will happen after results are known? Who will get the report? Will I get a copy written so I can understand it?

10. Why do we want to do an evaluation? We know we are doing a good job.

11. I don't like the gifted program because the students who are in it can't take band, participate in athletics, etc.

Parent

1. Why isn't my child in the gifted program? or Why is he?

2. Will my child score well on college entrance tests if he is in the program?

3. Is it right to have special programs? Shouldn't all students

get the same program?

4. I think you're making a bunch of snobs at your school. What are you going to do about it?

5. What can we do to help the teacher do a better job?

6. Is it necessary for my child to have so much homework?

7. I don't understand why my child doesn't have homework?

8. Why don't my child's teachers crack down on him and make him work? He has the ability but he just doesn't work.

9. Why doesn't the school offer special programs in art, music, etc.?

10. These frills are costing too much money. Why don't you just teach the basics?

11. Why should I answer your questions? You won't do anything about it anyway.

12. How can the PTA help with your evaluation?

Student

1. Why do our assignments have to be so long?

2. The teachers in this school really don't care what the students think.

3. Will I have a better chance of getting into college if I take these courses?

4. Why do we have to take so many tests?

5. I get tired of filling out forms. Sometimes I just make fake answers on them.

6. Will we get to find out how we did on the tests?

7. How can I get out of this class?

8. How can I get into this class?

A NEW ROLE IN EDUCATION: THE EVALUATOR

G. Sorenson

With the increase of federal funds for education, a new professional is emerging - the evaluator. He is somewhat different from the expert in tests and measurements and in research design usually found working on a college faculty. Rather, he is a person who spends part or all of his working hours at research and development activities, thinking about and planning the evaluation of educational processes. Because his role is a new one on the educational scene, his functions and his relationship to other educational experts need to be more clearly defined. It is the aim of this article to present some ideas about that role.

Two papers on evaluation, one by Scriven (1965) and one by Stake (1966), contain a number of assertions and implicit assumptions about the evaluator's role which deserve examination. Among them are the following:

1. Scriven would assign evaluators the task of determining the effectiveness of instructional programs. But more than that, he would have them evaluate the goals of these programs as well. It is not enough for the evaluator to find out whether the teacher of mathematics or English or physical education has taught the students what he intended to teach them. The evaluator must also decide, Scriven believes, whether the specific course content was appropriate and worthwhile; for, as Scriven sees it, the evaluator is the person best qualified to judge.

2. Scriven holds that the relative goodness of different educational goals is to be determined by applying a set of absolute standards which will somehow be obvious to the evaluator. Apparently, Scriven doubts that it is possible for intelligent, informed, and well-intentioned people seriously to disagree about what should be taught, for he asserts that arguments over criteria turn out to be mainly "disputes about what is to be counted as good, rather than arguments about the straightforward 'facts of the situation,' i. e. about what is in fact good." (Page 13)

3. Continuing his argument, Scriven implies that without absolute standards, evaluation is in fact probably impossible. "The process of relativism has not only led to over-tolerance for over-restrictive goals, it has led to incompetent evaluation of the extent to which these have been achieved..." (Page 18)

4. Stake seems to imply that since absolute standards exist, it is not necessary to take the individual teacher's nor the individual school's goals into account. He seems to believe that such standards should be applied even if they relate only slightly or not at all to the local school's resources and goals. "It should be noted that it is not the educator's privilege to rule out the study of a variable by saying, 'That is not one of our objectives.'" (page 4, 11)

5. Both Scriven and Stake believe that it is possible and perhaps desirable to appraise teaching and other instructional programs independent of their effects on the students. Stake (page 11) says, "The educational evaluator should not list goals only in terms of anticipated student behavior. To evaluate an educational program, emphasis must be given to what teaching as well as what learning is intended..."; and, "It is not wrong to teach a willing educator about behavioral objectives - they may facilitate his work. It is wrong to insist on them..." (page 12). Scriven further comments that "...pressure on a writer (curriculum maker) to formulate his goals, to keep to them, and to express them in testable terms, may enormously alter his product in ways that are certainly not always desirable." (page 21)

6. It may be inferred that Scriven believes that teachers who feel threatened by evaluators holding such absolute values should be ignored or at least discounted. "A little toughening of the moral fibre is required if we are not to shirk the social responsibilities of the educational branch of our culture." (page 5)

7. While it appears that he endorses most of Scriven's assertions, Stake would qualify at least one of them. If an individual evaluator were less than fully qualified, Stake would substitute a team of specialists as the appropriate determiners of educational goals and practices. The team would consist of experts in "instructional technology...psychometric testing and scaling...research design and analysis...the dissemination of information...(and perhaps) a social anthropologist" (page 23). He does not include historians, philosophers, businessmen, labor leaders, legal experts, or even non-behavioral scientists.

To be sure, the assertions listed above do not constitute a summary of what Scriven and Stake have said in their papers. Nevertheless, it appears that they represent, at least roughly, some of the beliefs of Scriven and Stake and a point of view resembling that of a number of writers on public education.

In spite of the fact that a number of brilliant and famous men support a position similar to that just described, I believe that if evaluators generally were to take an absolutist position, a number of unfortunate consequences would follow.

For one thing, teachers would be unwilling to cooperate and work with these evaluators. An evaluator who insists on evaluating in terms of his own goals while ignoring what the school people are trying to do, an evaluator who criticizes them and the school for failing to do what they had not intended to do in the first place would certainly be viewed as threatening. It can be safely predicted that teachers who feel threatened will resist and will devote their time and energies to defending old practices rather than to examining and improving them.

A second unfortunate consequence would be that evaluators would not get the support they need from powerful groups in the community who have a legitimate interest in what goes on in the school. Evaluation requires large amounts of time, money, and other commodities that evaluators cannot get without a good deal of public support-especially if they already have alienated the teachers and school administrators. Many of the individuals and groups in this country whose support is needed believe that the schools were invented to serve the needs of society and ultimately are answerable to the taxpayers, or at least to someone other than professional evaluators.

These individuals and groups do not always agree with one another about how the schools can best serve society, but they do agree that the schools are not autonomous. Many of these individuals - for example, Paul Goodman, Robert Hutchins, Sidney Hook, James Conant, John Goodlad, Roald Campbell, Ralph Tyler, Clark Kerr, Admiral Rickover, Harold Taylor, Paul Woodring, Jerome Bruner, David Ausubel, Myron Lieberman, Lawrence Cremin, Benjamin Bloom, to name only a few, as well as many groups - have given a good deal of thought and study to questions about the goals and methods of education. They are likely to regard individuals whose main qualification for the prescribing of educational goals is that they are experts in psychometry, research design, or social anthropology, but who are ignorant of the philosophical and political issues in education, as naive, arrogant, parochial, and, therefore, unworthy of assistance.

A third possible consequence - an evaluation program based upon the absolutistic assumption that "good" educational programs exist independent of persons and their preferences and independent of what students learn - is bound to fail. Its results are certain to be inconclusive and meaningless.

An analogy can be found in the attempts to evaluate teacher effectiveness. After surveying the results of half a century of research, investigators like Anderson and Hunka (1963) and Turner and Fattu (1960) have concluded that research in this area has been unproductive and has reached a dead end because of problems encountered in developing suitable criterion variables. In statistical terms, the variables lack reliability. It is my contention that the reason for the failure to develop usable criterion variables is a basic error in the way in which the researchers conceptualized the problem - more specifically, in their reliance on an absolute model of teacher effectiveness. Virtually all the investigators assumed either implicitly or explicitly the existence of sets of behaviors that objectively define the teacher-behaviors which exist as an absolute, independent of any particular observer and which would be recognized by an experienced educator when he encountered them, even though he might not be able to verbalize them in advance. Those researchers were failing to recognize and take into account the fact that any two observers are likely to differ in their beliefs about the ideal traits of the good teacher.

Ryans (1960) found that even when two observers were simultaneously watching the same teacher, they did not agree about him in their independent ratings unless they had had considerable training in Ryan's rating system - and sometimes not even then. It was probably his observers' differing notions about the ideal teacher they were observing. Analogously, any two evaluators are likely to disagree about the goals of education and can, therefore, be expected to disagree about the "goodness" of whatever actual method or program they may at a specific time be seeking to evaluate. The point is, there never has been and never will be general agreement on the goals of education any more than there is agreement on the qualifications and characteristics of the ideal teacher. Though particular groups of people will agree on particular goals, we must live with the fact that there is a welter of conflicting ideas on the subject in the society as a whole.

Following is a set of assumptions which may provide a reasonable alternative to those selected from Scriven and Stake.

1. Educational institutions should serve the needs of society and of the individuals who comprise it; these needs are complementary and interdependent.

2. A society's needs can best be defined by the members of that society through discussion, persuasion, and, ultimately, through voting. To insure that the goals of education will correspond with the citizens' views of their needs, the goals should be defined in a process of interaction between professionals and representatives of the society.

3. Every society changes; its needs and values are in a constant state of flux. Because of increases in population, knowledge, and technology, our society is very different from what it was even a decade ago. We now need new classes of workers, e.g., technicians who can build and operate computers. And because, as Gerard Piel (1961) has pointed out, we are no longer a society characterized by scarcity of goods, values based on dearth, such as hard work, thrift, etc., are less salient. Concomitantly as our needs and values change, we must expect our educational goals to change.

4. Even though many of our values seem to be changing, we continue to prize diversity. Ours is a pluralistic society with different religions, political viewpoints, subcultures, and values. We believe that our hetergeneity makes our society richer, more interesting, and stimulating. What is even more critical, we believe that heterogeneity makes our society viable. To accommodate such a diverse population, we must expect our educational goals and practices to be varied.

5. The goals of our educational institutions are not and never have been limited to purely academic objectives. Most people want the schools to do more than to teach the traditional academic subjects: they want individual and societal objectives included. For example, a century ago, the McGuffey Readers attempted to inculcate moral principles. More recently, James B. Conant (1953, page 62) said that the schools should provide a basis for the growth of mutual understanding between the different cultural, religious, and occupational groups in our country. "If the battle of Waterloo was won on the playing fields of Eton, it may well be that the ideological struggle with Communism in the next fifty years will be won on the playing fields of the public high schools of the United States."

6. We can tell if an educational program or teaching method is working only by observing whether hoped-for changes are occurring in the students - while at the same time making certain that damaging changes are not occurring, e.g., learning to hate a particular subject or learning to believe one cannot learn arithmetic even if he works at it. We cannot properly evaluate an instructor or a program without assessing the effects wanted and unwanted, on students. To evaluate a schedule of events within a school, or a series of teacher activities, or any array of teacher characteristics while neglecting the product is to examine intentions without considering consequences.

7. Educational goals must be stated in descriptive rather than in interpretive language. We have learned that it is not useful to define educational goals in the terms formerly used by professional educators and still used by their critics. We know that instead of such high-sounding slogans as "transmitting the cultural heritage," "educating citizens for democracy," and "developing the individual's potential," we must develop objectives defined in terms of changes in pupils' behavior or in the products of student behaviors. We must also be careful that, in rigorously setting behavioral goals, we do not slip into triviality. We must be prepared to defend each behavioral goal in terms of value assumptions and to answer the question why one particular behavioral goal is better than another. These points do not represent new thinking. They describe a trend, which according to Ralph Tyler (1954, 1956) began about 1935, a trend of which many public school teachers still are unaware. Tyler stated that it is more important to evaluate the educational process than the structure of the school and that it is more important to evaluate the product than the process. I would rephrase this point: the proper way to evaluate both the educational process and the structure of the schools is to find out whether they are in fact producing the hoped-for product.

The function of the professional evaluator should be to help teachers and administrators in a given school to do such things as the following:

1. Define their goals in terms of pupil performance. John McNeil (1966), director of Supervised Teaching at UCLA, and I both have found that many experienced teachers are not able to define their objectives in language which describes observable changes in pupil behavior. It is easy to be critical of such teachers, and it is easy to state educational goals behaviorally - if we limit ourselves to role learning. For example, "students will be able to name the bones of the body" is a goal stated in behavioral terms. While this goal may be important in some contexts, it is a very limited one. The behavioral definition of higher order goals is much more difficult. At the end of a course, teachers want their students to perform in such a manner as to warrant the inference that the students have learned to "know," "understand," "appreciate," and "think" about what the teacher has tried to teach. Merely to tell teachers that they should state these goals behaviorally is far from sufficient. What would be more helpful would be to show them how, and to invent more sophisticated instruments for them to use.

2. Learn how systematically to discover differences among pupils that require particular kinds of instruction. Teachers need appraisal devices that will do more than reveal differences in what students already have learned. They need instruments that will also reveal barriers to, or interferences with learning, among them (a) misconceptions; (b) particular habits, such as failure to pay attention; (c) certain needs that the child is satisfying at the expense of learning, e.g., need for group approval or sensitivity to peer pressures; and (d) attitudes deriving from class and ethnic background, etc. Some important differences among students are so subtle that, without sophisticated instruments, the child who has not learned to attend to the teacher's instructions may be mistaken for a dull child, or an angry one, or perhaps one with a constitutional impairment.

3. Design and administer evaluation programs. More importantly, professional evaluators should help individual teachers to find out which of their instructional procedures are paying off and which are not. With guidance, it is possible for the teachers themselves to try out and to evaluate alternative instructional methods on the job. For example, Bartlett (1960) demonstrated that when an instructor spent part of his time in an algebra class teaching study habits, the students learned more than when he spent the entire time teaching algebra.

Public school people do not need more critics - critics abound. What these educators do need is someone to help them find and test alternative solutions to the complex problems they face daily. For the most part, university personnel who have the knowledge to perform the kinds of evaluation functions described above have not been taking their knowledge to the schools. They have been publishing their findings in professional journals, but they have failed to make explicit to teachers the relevance of those findings for the teachers' work. Hopefully, the research and development evaluator will bridge the gap between the laboratory and the field.

(Evaluation Comment, Vol. I, No. 1, January, 1968.)

EXERCISE ON RATING SCALES

Rating scales are generally used in situations where we are observing the behavior of a person and want to be as objective in our observations as possible. Rating scales take many different forms. A good overview of the different kinds of rating scales is provided in the chapter by H. H. Remmers entitled "Rating Methods in Research on Teaching" in the Handbook of Research on Teaching.

The U. S. Air Force has done much research on construction of rating scales. (Most of this work has been done at the research center at Lackland Air Force Base in San Antonio, Texas.) An important finding of this research is that rating scales that have their points defined by behavioral statements are generally more reliable than scales in which the points are defined by numbers or by adjectives.

This exercise is designed to have you work through the development of rating scales that use behavioral statements to define the points. We suggest that you work on a scale or scales that you might use by following the procedures we have outlined in a sample problem.

When behavior is to be observed and rated in some situation, the first consideration in building the rating devices is to decide what components of the behavior are going to be observed. For example, in building an observation device for grading essays we would first decide what things we will consider in grading the essays. The following list contains examples of the kinds of things that might be considered in grading an essay.

1. Vocabulary level
2. Sentence structure
3. Paragraph construction

4. Format

5. Quality of argument

6. Use of references

7. Writing style

Certainly those aspects of the essay that will be considered in the grading of the essay will be determined to a great extent by the purposes of the essay assignment. The objectives of the specific assignment should probably not be the only criteria employed in judging the essay, however. For example, vocabulary building may not be an objective of an assignment, but vocabulary level would be an appropriate aspect to consider in grading most essays.

When the decision has been made regarding the components, we can then build a rating scale for each of these components. The rating scale for each component will define a continuum of which we can rate the person's performance from low to high.

An early decision in building a rating scale is that of how many points to have on the scale. Research on the behavior of raters indicates that with few points on the scale the raters are uncomfortable because they would like to make finer discriminations than the few points allow. On the other hand, there is a limit to the fineness of discrimination that most raters can make. The research tends to indicate that a scale should not have fewer than five points nor more than fifteen or sixteen points. A seven or nine point scale seem to be the preference of most raters.

The Air Force research alluded to above has indicated that a scale in which the points are defined by statements is superior to numerical or adjective scales in terms of reliability. The research of this group has also studied whether each point should be defined or whether a fewer number of statements than points yields results comparable to having every point defined. The results suggest that a minimum of three points should be defined, the two end-points and the mid-point, and that definition of five points seems to be sufficient. Defining more points than five does not seem to increase reliability of ratings.

The research on rating scales indicates that a desirable format for scales would be like the one shown below.

```
/________/_________/_________/_________/_________/_________/________ /________/________ /
Statement                Statement                Statement                Statement           Statement
defining                 defining                 defining                 defining            defining
low end                  mid-point                mid-point                mid-point           high end
                         of low half                                       of high half
```

This is a nine point scale with five of the points defined by statements. This exercise is to build such a scale.

The important concern in constructing such a scale is to write statements that define well the points along the continuum. The statements should both define the point very well and also reflect the distance between the points as accurately as possible.

A usual procedure for building this kind of rating scale is to just write five statements, one to define each point, and the scale is done. Such a procedure, although quick, has obvious limitations. The statements may be ambiguous and there is little confidence that they reflect the distances along the continuum.

There is a procedure for scaling the statements that will yield a scale on which the statements are likely to be good definitions of the points. Furthermore, the procedure provides a basis for assigning scale values to the statements so that we can select those statements that most nearly define those points that we want to define.

The procedure that is described below is referred to as the "Method of Equal-Appearing Intervals." It was originally developed by Thurstone and Chove in 1929. A complete description of the method is in the book by Edwards entitled Techniques of Attitude Scale Construction. This book is listed on the reference listing.

The steps in the procedure are as described below.

1. Identify the characteristic that is to be rated.

2. Write a number of statements that are descriptive of behavior along the continuum. The statements should be relevant to the context in which the behavior will be observed. The statements should be written so that they cover the continuum with about equal numbers at all points along the continuum. For a rating scale it is suggested that at least 20 statements be written. The "Suggested Criteria for Writing Attitude Statements" paper attached to this exercise contains some points that should be considered in writing the statements.

3. Have a number of judges (at least 15) judge the statements as to where they belong on the continuum. One way to do this is to type each statement on a 3 X 5 card. Prepare nine other cards with the letters A to I on them. That is, each card will have a letter on them. Arrange the lettered cards along a table and indicate to the judge that the letter A indicates the low end of the continuum, the letter I the high point, and the letter E the midpoint.

   Each judge is then asked to judge the point on the continuum which each statement defines by placing the statement under the appropriate lettered card.

4. When the judge has placed all of the statements to his satisfaction, record the pile in which the judge placed the statement by writing the letter on the back of the statement card. It is more convenient to record the letters as numbers. Thus, write a one on the back of card in pile A, a two on cards in pile B, and so on.

5. After all judges have completed the judging, compute the median scale value for each statement and the Q value. Q is the interquartile range. Procedures for computing these values are described in Edwards. The following example problem describes the procedures also.

Suppose a statement has been judged by 15 judges. Four placed it under B, five under C, and six under D. Arbitrarily assign the number two to represent the B category, three to C, and four to D. Thus, if the assigned numbers represent scale values for the statements, then four judges gave the statement a scale value of two, five judges gave it a scale value of three, and six judges gave it a scale value of four.

Scale value	Number of judges
2	4
3	5
4	6

The median scale value is that scale value below which 50% of the judges placed the statement and above which 50% of the judges placed the statement. The following formula is used to compute this value.

$$s = 1 + \left(\frac{.50-P_b}{P_w} \right) i$$

where s - the median scale value.

1 = the lower limit of the interval within which the median value occurs.

P_b = the proportion of cases below the interval where the median is.

P_w = the proportion of cases within the interval where the median is.

i = the height of the interval.

For the problem:

1 = 2.5 because the scale value of three really represents the interval of 2.5 to 3.5. The median value falls in this interval.

$P_b = 4/15 = .27$

$P_w = 5/15 = .33$

$i = 1$

so

$$s = 2.5 + \left(\frac{.50-.27}{.33}\right) .1$$

$$s = 2.5 + \frac{.23}{.33}$$

$$s = 2.5 + .70$$

$$s = 3.2 = \text{the median scale value}$$

The procedure for obtaining Q is very similar to that for obtaining the median. Q is the interquartile range which is the range covered by the middle 50% of the cases. To obtain Q it is necessary to find the first quartile and the third quartile. The first quartile is that point that divides the scores into 25% below and 75% above, and the third quartile divides the scores in the proportion 75=25. The difference between these points is Q.

$$Q_1 = 1 + \left(\frac{.25-P_b}{P_w}\right) i$$

and
$$Q_3 = 1 + \left(\frac{.75-P_b}{P_w}\right) i$$

Computing Q_1

$1 = 1.5$

$P_b = 0$

$P_w = .27$

$i = 1$

so $Q_1 = 1.5 + \left(\frac{.25-0}{.27}\right) 1$

$Q_1 = 1.5 + \frac{.25}{.27}$

$Q_1 = 1.5 + .94$

$Q_1 = 2.44$ = first quartile

For Q_3

$1 = 3.5$

$P_b = .60$

$P_w = .40$

$i = 1$

so

$Q_3 = 3.5 + \left(\frac{.75-.60}{.40}\right) 1$

$Q_3 = 3.5 + \frac{.15}{.40}$

$Q_3 = 3.5 + .38$

$Q_3 = 3.88$ = third quartile

$Q = Q_3 - Q_1$

$Q = 3.88 - 2.44$

$Q = 1.44$

For the statement then, the median scale value is 3.2 and the Q is 1.44.

6. When the median scale value and Q have been computed for each statement, the five statements for the rating scale can then be selected. The criteria for statement selection are to select those five statements that most nearly have the scale values of the points to be defined and that have the smallest Q value. These criteria are not absolute because of two statements one might have a scale value a little nearer the point than the other

but have a larger Q. Your judgment or hunch about the two statements will have to prevail in such a situation. It is well to remember that a large Q reflects ambiguity in the statement.

If a nine point scale is being constructed with five defined points, the five statements should have scale values as near one, three, five, seven, and nine as possible.

SUGGESTED CRITERIA FOR WRITING ATTITUDE STATEMENTS

-Chas. K. A. Wang

1. An attitude statement must be debatable. It must represent only an opinion which has no general acceptance.

   Bad: It is hard on the children to have the mother working.

   Better: Women with children should not work.

2. All statements on a given issue should belong, as nearly as can be judged, to the same attitude variable. That is, they must be not only relevant to the issue but belong to the linear continuum that is being measured.

   Statement: In an ideal society there would be no law. (From a scale on attitude toward law, where the variable being measured is from complete respect to utter disrespect for law.)

3. An attitude statement must not be susceptible to more than one interpretation.

4. Avoid "double-barreled" statements.

   Statement: Athletic conditions are bad, but officials are trying to improve them.

5. An attitude statement should be short. It should rarely exceed fifteen words in length.

   In writing attitude statements, it is well to try to shorten the length of each sentence written. In doing so, one usually also avoids the violation of many of the other rules here mentioned.

6. Each attitude statement should be complete in denoting a definite attitude toward a specific issue. Do not assume that the issue in question can be understood without specific reference to it.

7. Each attitude statement should contain only one complete thought.
"The church was established to serve a useful purpose but it has outlived its time; therefore, it is doing more harm than good."

8. Avoid grouping two or more complete sentences as one attitude statement. Do not transplant quotations by the paragraph en bloc, but rewrite them into one single sentence or several separate statements.

9. An attitude statement should be clear-cut and direct. Avoid statements, which are not directly an attitude but from which an attitude is to be inferred, unless the inference is clear and unquestionable.

10. Use with care and moderation such words as "only," "mere," "just," (in the sense of only), "merely," etc.
Statement: Only by taking the money out of football can it be made really amateur.

11. Avoid colorless expressions or statements lacking effect.
"The unions (or anything else) are all right."

12. Whenever possible, write an attitude statement in the form of a simple rather than a complex or compound sentence. The simple kind of statement reduces the chance for a wrong interpretation.

13. When a statement cannot be made in the form of a simple sentence, write it as a complex rather than a compound one.

14. It is usually better to use the active rather than the passive voice.

15. In general, use the term of the issue as the subject of a statement. This is desirable in order to secure proper emphasis and attention. Hence it is permitted even in violation of Rule 14.

16. Avoid high-sounding words, uncommon words or expressions, technical terms not ordinarily understood, etc. When a scale is being prepared

for use in a specific age, school, or sociological group, the vocabulary of that group should be borne in mind.

In addition to the foregoing criteria, there may be mentioned several general rules, based largely upon good usage in English. These rules improve sentence structure although they are not necessarily concerned with the scale values or the Q-values of the statements.

1. Avoid a negative expression whenever a positive one can be substituted. Thus, use "disagree" instead of "not agree", "difficult" instead of "not easy" etc. Exceptions, of course, are permitted when the negative effect is desired.
2. Avoid double infinitives, especially in a short statement. For example, instead of saying, "To work on Sunday is to be immoral," say "Working on Sunday is immoral."
3. Do not use redundant phrases. To illustrate:
   Bad: We should not knock but boost our public officials.
   Better: We should boost our public officials.
4. Avoid excessive use of such phrases as "I think that..."; "I believe that ..."; "I feel ..."; etc., to precede a statement.
5. Avoid double negatives.

ATTITUDE SCALE EXERCISE

There are two commonly used techniques of attitude scale construction sometimes referred to as the Thurstone and Chove method and the Likert method. This exercise will be with the Likert method. The Thurstone and Chove method is used in the exercise on rating scales.

The Likert method is described in the chapter entitled "The Method of Summated Ratings" in Edwards, Techniques of Attitude Scale Construction.

The usual definition of an attitude is that it is a feeling held by a person toward some psychological object. Thus an attitude refers to feelings about specific things. The fact that attitudes are specific is a reason why few attitude scales are available commercially, and why they often have to be built specifically for a project. The book by Shaw and Wright that is listed in the references does contain descriptions of many attitude scales that have been used, but few of these are commercially available. Many may be obtained directly from the author, however.

The following paragraphs describe the steps to follow in constructing an attitude scale with the Likert method.

1. The first thing is to select the psychological object. This object must be something that evokes positive or negative responses from people. It should be quite specific. It is difficult to develop a scale measuring attitudes toward education, business, government, etc., because it is difficult to find people who have negative or positive feelings about such general and pervasive concepts. Who is against education? Something like a gifted program, the War on Poverty, television instruction, etc. are specific enough and do evoke positive and negative responses.

2. After deciding what it is you want to measure, then develop a content outline. The content outline will define the factors that contribute to a person's having positive or negative feelings about the program. For example, a content outline for an attitude toward the gifted program scale might include (1) cost, (2) the problem of equality of education, (3) the problem of social relationships, and (4) the definition of who the gifted are.

3. The content outline provides a guide for the content of the statements you will write. The next step is to write about 40 statements for the attitude scale. A paper by Wang entitled "Suggested Criteria for Writing Attitude Statements" is attached to this exercise. It contains some helpful ideas to follow in writing your statements.

4. Prepare the 40 statements in a format in which a respondent can react to each statement in one of five categories of strongly agree, agree, neutral, disagree, and strongly disagree.

5. Administer this 40 statement scale to a sample of persons similar to those you will use in the study. This try-out sample should be as large as feasible. A sample size of thirty is minimal, and it is preferable to have 100 or more. For the exercise you should be able to get 30 from the participants and staff.

6. Score the scale by assigning weights to the response categories on the basis of your judgment of whether the statement reflects a favorable or unfavorable feeling. It is recommended that you assign a weight of 4 to strongly agree down to a 0 for strongly disagree to all favorable statements and a weight of 0 for strongly agree up to a 4 for strongly disagree to all unfavorable statements.

When you sum the scores for a person across the items then a high score will indicate a favorable attitude toward the object and a low score an unfavorable attitude.

7. Obtain the total score on the 40 items for each person in the sample.

8. With small samples (less than 100) divide the group at the median. Those above the median will hereafter be referred to as high scorers and those below as low scorers. If there are a number of scores in the median interval assign these papers at random so that there are 50% of the total group in the high scorer and low scorer groups. If your sample size is 100 or more then find the top 27% and the bottom 27%. The middle 46% will be ignored for subsequent analysis.

9. For each of the 40 items compute the proportion of people in the high group and the proportion of people in the low group who scored high on the item. This would be the proportion who scored four or three on the item plus one half of those who scored two.

10. Using an abac, a copy is attached, find the point biserial correlation between the item score and the total score.

11. Select the 20 or 25 items with the highest correlation as the items for the attitude scale.

12. Make the final form of the attitude scale using the same format as for the try-out version.

13. Before using the scale, an attempt should be made to validate it. A common procedure for doing this is to have two groups, whom you feel should have different attitudes toward the object,

take the scale. If the groups do score differently on the scale you have evidence that the scale is measuring the attitude you want to measure. You will probably not be able to do the validity step during the workshop.

SUGGESTED CRITERIA FOR WRITING ATTITUDE STATEMENTS

-Chas. K. A. Wang

1. An attitude statement must be debatable. It must represent only an opinion which has no general acceptance.

   Bad: It is hard on the children to have the mother working.

   Better: Women with children should not work.

2. All statements on a given issue should belong, as nearly as can be judged, to the same attitude variable. That is, they must be not only relevant to the issue but belong to the linear continuum that is being measured.

   Statement: In an ideal society there would be no law. (From a scale on attitude toward law, where the variable being measured is from complete respect to utter disrespect for law.)

3. Ar attitude statement must not be susceptible to more than one interpretation.

4. Avoid "double-barreled" statements.

   Statement: Athletic conditions are bad, but officials are trying to improve them.

5. An attitude statement should be short. It should rarely exceed fifteen words in length.

   In writing attitude statements, it is well to try to shorten the length of each sentence written. In doing so, one usually also avoids the violation of many of the other rules here mentioned.

6. Each attitude statement should be complete in denoting a definite attitude toward a specific issue. Do not assume that the issue in question can be understood without specific reference to it.

7. Each attitude statement should contain only one complete thought. "The church was established to serve a useful purpose but it has outlived its time; therefore, it is doing more harm than good."

8. Avoid grouping two or more complete sentences as one attitude statement. Do not transplant quotations by the paragraph en bloc, but rewrite them into one single sentence or several separate statements.

9. An attitude statement should be clear-cut and direct. Avoid statements, which are not directly an attitude but from which an attitude is to be inferred, unless the inference is clear and unquestionable.

10. Use with care and moderation such words as "only," "mere," "just," (in the sense of only), "merely," etc.
Statement: Only by taking the money out of football can it be made really amateur.

11. Avoid colorless expressions or statements lacking effect. "The unions (or anything else) are all right."

12. Whenever possible, write an attitude statement in the form of a simple rather than a complex or compound sentence. The simple kind of statement reduces the chance for a wrong interpretation.

13. When a statement cannot be made in the form of a simple sentence, write it as a complex rather than a compound one.

14. It is usually better to use the active rather than the passive voice.

15. In general, use the term of the issue as the subject of a statement. This is desirable in order to secure proper emphasis and attention. Hence it is permitted even in violation of Rule 14.

16. Avoid high-sounding words, uncommon words or expressions, technical terms not ordinarily understood, etc. When a scale is being prepared

for use in a specific age, school, or sociological group, the vocabulary of that group should be borne in mind.

In addition to the foregoing criteria, there may be mentioned several general rules, based largely upon good usage in English. These rules improve sentence structure although they are not necessarily concerned with the scale values or the Q-values of the statements.

1. Avoid a negative expression whenever a positive one can be substituted. Thus, use "disagree" instead of "not agree", "difficult" instead of "not easy" etc. Exceptions, of course, are permitted when the negative effect is desired.
2. Avoid double infinitives, especially in a short statement. For example, instead of saying, "To work on Sunday is to be immoral," say "Working on Sunday is immoral."
3. Do not use redundant phrases. To illustrate:

   Bad: We should not knock but boost our public officials.

   Better: We should boost our public officials.
4. Avoid excessive use of such phrases as "I think that..."; "I believe that ..."; "I feel ..."; etc., to precede a statement.
5. Avoid double negatives.

COMPUTATION OF CHI-SQUARED

The chi-squared statistical technique is often used when we want to know whether two variables are related to each other, but we are only able to classify the object observed rather than measure it on one or both of the variables. The example problem illustrates a situation in which chi-squared might be used. The problem is presented as a series of steps in computation. As you work through each step of the problem you can check your computation with the correct answers that are provided on the last sheet. The problems you are to work are numbered.

A program evaluator has conducted a survey which was designed to find out the feelings of the community about the gifted program. His questionnaire obtained information about the education level of the respondent and a response to a question of what the school should do with the gifted program. He decided to see whether the education level of the respondent was related to the way they answered the question about the gifted program. The table contains the results. The number in each cell is the number of people who are classified into that category by their responses. Thus, there were 18 people with some college education who believed the gifted program should be dropped.

Response to Question on What the School Should do with the Gifted Program

Education Level	Drop it	Deemphasize	Keep as is	Expand	Total
Some College	18	29	70	115	232
High School Graduate	17	28	30	41	116
Less Than High School	11	10	11	20	52
Total	46	67	111	176	400

An examination of the table suggests that people with more education tend to be more supportive of the program than those with lower levels of education. Before we make such a conclusion, however, we would like to determine how confident we can be that the results did not occur just by chance and there really is no relationship between the education level of the respondents and their response to the question. The chi squared technique is an appropriate way to establish our degree of confidence in judging that there is a relationship.

To compute chi squared we have to first determine what the table would be if there were absolutely no relationship between the variables. To do this we compute what are called expected values for each cell. The expected values are the values we would expect to occur were there no relationship. To obtain the expected values we use the totals column and row. Notice that 232 of the 400 people went to college, 116 graduated from high school, and 52 had less than high school education.

1. What percent of the people went to college?

2. What percent graduated from high school?

3. What percent had less than high school?

If there were absolutely no relationship between the variables then these same percentages should occur in each of the columns. Thus, 58% of the people who said the gifted program should be dropped should have had some college, 29% should have been high school graduates, and 13% should have had a less than high school education.

58% of 46 is 27 rounded off

29% of 46 is 13 rounded off

13% of 46 is 6 rounded off

The values 27, 13, and 6 are the expected values for that column because these are the values we would expect to get if there were no relationship between the variables.

4. Compute the expected values for the rest of the cells in the table. Round off to the nearest whole number. Remember, multiply the percentages by the column total.

The values you have just computed are the expected values within rounding for the situation of no relationship between the variables. Notice that all of the column values are in the proportion of 232:116:52:400. Likewise the rows are the same proportion as 46:67:111:176:400. Notice also that the row and column totals are the same for the expected table as for the table with the actual data. This always is the case.

The next step is to compute the chi squared value. The formula is:

$$X^2 = \Sigma \frac{(O-E)^2}{E}$$

O means observed frequencies, that is, the actual number of cases in each cell in the data.

E means expected frequencies, that is, the expected values that you computed.

The formula indicates that we should subtract the expected frequency from the observed frequency for each cell. We then square this difference and divide the squared value by the expected frequency. We then add all these results together and have our chi squared value.

5. Finish this equation. You do not use the totals columns and rows.

$$X^2 = \frac{(18-27)^2}{27} + \frac{(29-39)^2}{39} + \frac{(70-64)^2}{64} +$$

$$+ \quad + \quad + \quad + \quad + \quad + \quad +$$

$$+ \frac{(23-20)^2}{20}$$

6. Finish this equation. Subtract the expected from the observed.

$$X^2 = \frac{-9^2}{27} + \frac{-10^2}{39} \text{ -------------------- } + \frac{-3^2}{23}$$

7. Finish this equation. Square the numerators obtained in six.

$$X^2 = \frac{81}{27} + \frac{100}{39} \text{ -------------------- } + \frac{9}{23}$$

8. Finish this equation. Change the fractions in seven to decimals. Round to two decimals.

$$X^2 = 3.00 + 2.56 \text{ --------------- } +.39$$

9. Compute X^2 by adding the decimals in eight together.

$X^2 =$

The next step is to determine the probability of getting a chi squared value as large as this by chance. To do this, we use a table that is found in most statistics books and is called the table of Chi squared.

To use the table we have to compute the degrees of freedom (d.f.) for our table. The degrees of freedom for a table are equal to the number of rows minus one times the number of columns minus one.

$$d.f. = (r-1)(c-1)$$

With the table in the problem we have 3 rows and 4 columns so:

$$d.f. = (3-1)(4-1)$$

10. or d.f. =

Entering the Chi squared table with six degrees of freedom we find the number 16.812 under the .01 or 1% column. This means that if the proportions we used in computing the expected frequencies were the actual situation in some large population, and we were to draw many random samples from this population, we could expect the proportions in the samples to differ from the population proportions sufficiently to yield a Chi squared value of 16.812 or greater only one time in one hundred samples. In other words a Chi squared this large wouldn't occur very often by chance. Our Chi squared value of 20.81 is even larger than 16.812 so we can conclude that it is very unlikely that the results we got occurred by chance. We can be quite confident that there is a relationship between the education level of the respondents and their response to the question. In statistical jargon we would say that the relationship is significant at the .01 level of confidence.

II. To give you some practice here is another problem. The question is whether students in the gifted program differed from students in the regular program in their participation in school activities. A random sample of 100 students from the regular program was used.

	Participate	Do Not Participate	Total
Gifted Students	55	45	100
Regular students	35	65	100
Total	90	110	200

1. Compute the expected frequencies. Note the proportion of gifted students and regular students and use these.

2. Compute X^2

3. Compute d.f. d.f. = (r-1)(c-1)

4. Evaluate the results

Answers on chi-squared problems:

1. 58%

2. 29%

3. 13%

4.

27	39	64	102	232
13	19	33*	51	116
6	9	14	23	52
46	67	111	176	400

*This cell was computed as 32, but rounded up to 33 to keep the row and column totals the same.

5.

$$X^2 = \frac{(18-27)^2}{27} + \frac{(29-39)^2}{39} + \frac{(70-64)^2}{64} + \frac{(115-102)^2}{102} + \frac{(17-13)^2}{13} + \frac{(28-19)^2}{19} + \frac{(30-33)^2}{33} + \frac{(41-51)^2}{51} + \frac{(11-6)^2}{6} + \frac{(10-9)^2}{9} + \frac{(11-14)^2}{14} + \frac{(20-23)^2}{23}$$

6. $X^2 = \frac{-9^2}{27} + \frac{-10^2}{39} + \frac{6^2}{64} + \frac{13^2}{102} + \frac{4^2}{13} + \frac{9^2}{19} + \frac{-3^2}{33} + \frac{-10^2}{51} + \frac{5^2}{6} + \frac{1^2}{9} +$

$\frac{-3^2}{14} + \frac{-3^2}{23}$

7. $X^2 = \frac{81}{27} + \frac{100}{39} + \frac{36}{64} + \frac{169}{102} + \frac{16}{13} + \frac{81}{19} + \frac{9}{33} + \frac{100}{51} + \frac{25}{6} + \frac{1}{9} + \frac{9}{14} + \frac{9}{23}$

8. $X^2 = 3.00 + 2.56 + .56 + 1.66 + 1.23 + 4.26 + .27 + 1.96 + 4.17 +$

$.11 + .64 + 39$

9. $X^2 = 20.81$

10. d.f. = (2)(3)

d.f. = 6

II. 1.

45	55	100
45	55	100
90	110	200

2. $X^2 = \frac{(55-45)^2}{45} + \frac{(35-45)^2}{45} + \frac{(45-55)^2}{55} + \frac{(65-55)^2}{55}$

$X^2 = \frac{10^2}{45} + \frac{-10^2}{45} + \frac{-10^2}{55} + \frac{10^2}{55}$

$X^2 = \frac{100}{45} + \frac{100}{45} + \frac{100}{55} + \frac{100}{55}$

$X^2 = 2.22 + 2.22 + 1.82 + 1.82$

$X^2 = 8.08$

3. d.f. = 1

4. The observed frequencies are highly unlikely to have occurred by chance. The gifted students are significantly more likely to have participated in school activities. Significant at .01 level of confidence.

COMPUTATION OF PEARSON r

The Pearson r is the most commonly used coefficient of correlation. It is an index of the extent to which there is a linear relationship between two sets of numbers, and by inference between two variables on which the numbers are measures. The Pearson r will generally be used when we have two sets of scores for a group of people and we want to see whether the variables that have been measured to yield the scores are related.

For example, an evaluator was interested in whether the students' performance on the Torrance Test of Creativity was related to their performance on a problem solving task in a science class. The scores for the students on the two tests were as follows:

Student	Torrance Test		Problem Solving Test		
	X	X^2	Y	Y^2	XY
1.	28		12		
2.	30		23		
3.	32		30		
4.	13		12		
5.	16		17		
6.	18		7		
7.	14		13		
8.	12		14		
9.	18		16		
10.	22		11		
11.	23		10		
12.	25		25		
13.	31		18		
14.	19		22		
15.	18		12		
16.	10		6		
17.	29		30		
18.	18		21		
19.	19		19		
20.	23		10		
21.	25		21		
22.	27		15		

The Pearson r would often be the statistic employed in a situation such as this to indicate the extent to which the variables are related.

To compute the Pearson r with the given data go through the following steps. You can che:k your results with those on the attached sheet.

1. Square each X score, each Y score, and multiply each X score by its corresponding Y score. Use a table of squares to square the scores.

2. Sum the five columns, that is, add up the X scores, the X^2 scores, the Y scores, the Y^2 scores, and the XY scores. The five values you will have are:

$\Sigma X=$
$\Sigma X^2=$
$\Sigma Y=$
$\Sigma Y^2=$
$\Sigma XY=$

3. The next three values we will get are the Σx^2, Σy^2, and Σxy. Notice that these are the lower case letters and are the symbols for what are often called the sum of squares and the sum of cross-products. The formulas for these are:

$$\Sigma x^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N}; \quad \Sigma y^2 = \Sigma Y^2 - \frac{(\Sigma Y)^2}{N}; \quad \Sigma xy=\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N}$$

N in all cases is the number of pairs of scores.

Lets compute each separately.

$$\Sigma x^2 = 10{,}914 - \frac{(470)^2}{22}$$

a. First square 470

$(470)^2 =$

b. Divide the obtained value by N which is 22.

$\frac{(470)^2}{22} =$

c. Subtract the quotient obtained in b from 10,914

$\Sigma x^2 = 10{,}914 - \frac{(470)^2}{22} =$

You now have the Σx^2.

Next compute Σy^2

d. Substitute the known values into the equation.

e. Square the ΣY

f. Divide the value obtained in step e by N

g. Subtract the quotient obtained in f from ΣY^2

$\Sigma y^2=$

Next compute Σxy

h. Substitute the known values into the equation

i. Multiply the ΣX times the ΣY

j. Divide the product obtained in i by N

k. Subtract the quotient obtained in j from ΣXY

$\Sigma xy=$

You now have the values needed to solve for Pearson r. Copy them here.

$\Sigma x^2=$

$\Sigma y^2=$

$\Sigma xy=$

A commonly used formula for obtaining the Pearson r is:

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

l. Substitute the values into the formula

r=

m. Multiply the Σx^2 times the Σy^2

n. Find the square root of the product obtained in m. You can estimate this quite well by using a table of squares.

o. Divide the Σxy by the root obtained in n

r=

After obtaining the correlation we want to know how much confidence we have that the coefficient indicates a real relationship or whether we would be quite likely to get a correlation of that size purely by chance. Tables are available in most statistics books to help us with this. They are usually labeled something like "Values of r at the 5 and 1 Per Cent Levels of Significance."

We enter the table under the column headed N and find the N that corresponds to our situation, in this case 22.

p. What is the r under the 5% column? Under the 1% column? in this row.

These figures mean that only 5 times in 100 would we get an r of .423 or greater by chance and only 1 time in 100 would we get an r of .537 or greater by chance. Our correlation is larger than .537 so we would conclude that the variables are related because it is highly unlikely that we would have gotten as large an r as we did were they not related. In fact, we would get such an r fewer than 1 time in 100 by chance. In research jargon we would say that there is a statistically significant relationship between the variables and the significance is at the .01 level of confidence. In gambler jargon we would say that the odds are large for betting that there is a significant relationship between the variables.

Answers on Pearson r problem

Student	Torrance Test		Problem Solving Test		
	X	X^2	Y	Y^2	XY
1.	28	784	12	144	336
2.	30	900	23	529	690
3.	32	1024	30	900	960
4.	13	169	12	144	156
5.	16	256	17	289	272
6.	18	324	7	49	126
7.	14	196	13	169	182
8.	12	144	14	196	168
9.	18	324	16	256	288
10.	22	484	11	121	242
11.	23	529	10	100	230

Student	Torrance Test		Problem Solving Test		
	X	X^2	Y	Y^2	XY
12.	25	625	25	625	625
13.	31	961	18	324	558
14.	19	361	22	484	418
15.	18	324	12	144	216
16.	10	100	6	36	60
17.	29	841	30	900	870
18.	18	324	21	441	378
19.	19	361	19	361	361
20.	23	529	10	100	230
21.	25	625	21	441	525
22.	27	729	15	225	405
	$\Sigma X=470$	$\Sigma X^2=10,914$	$\Sigma Y=364$	$\Sigma Y^2=6978$	$\Sigma XY=8296$

a. $(470)^2 = 220,900$

b. $\frac{(470)^2}{22} = \frac{220,900}{22} = 10,040.91$

c. $\Sigma x^2 = 10,914 - 10,040.91$

$\Sigma x^2 = 873.09$

d. $\Sigma y^2 = 6978 - \frac{(364)^2}{22}$

e. $(364)^2 = 132,496$

f. $\frac{132,496}{22} = 6022.55$

g. $\Sigma y^2 = 6978 - 6022.55$

$\Sigma y^2 = 955.45$

h. $\Sigma xy = 8296 - \frac{(470)\ (364)}{22}$

i. $(470)\ (364) = 171,080$

j. $\frac{171,080}{22} = 7,776.36$

k. $\Sigma xy = 8296 - 7776.36$

$\Sigma xy = 519.64$

l. $r = \frac{519.64}{\sqrt{(873.09)(955.45)}}$

m. $(873.09)(955.45) = 834,193.8405$

n. $\sqrt{834,193.8405} = 913.34$

o. $r = \frac{519.64}{913.34} = .57$

p. .423, .537

COMPUTATION OF SPEARMAN rho

The Pearson r is the index that is most commonly used to indicate the strength of the linear relationship between two variables. Many times the Pearson r is not really appropriate because the data do not meet the assumptions for the Pearson r. An r can be computed between any two sets of numbers, but the kinds of interpretation that are made of r might be misleading if the assumptions are not met. The four assumptions for the Pearson r are

1. The two measures are obtained independently
2. The sample is drawn randomly from a parent population
3. The characteristics (variables) are normally distributed in the parent population
4. The scales used to measure the variables are interval scales.

The Spearman rho or the rank-order correlation is a statistic that can be used when assumptions three and four are not met. Assumptions one and two above are made for the Spearman rho, but interpretation of the rho does not assume anything about how the characteristics are distributed in the population and rho only requires that the scales are ordinal scales. The Spearman rho is interpreted as the Pearson r, that is, it is an index of the strength of the linear relationship between two variables.

The following problem is an exercise for you to work a Spearman rho. The answers to the steps are provided on the answersheet at the end of the exercise. The problem is the same one that is on the Pearson r work sheet. The correlation being obtained will indicate the relationship between the Torrance Test of Creativity and performance on a problem solving task with an N of 22.

Student	Torrance Test		Problem Solving Test			
	X	Rx	Y	Ry	D(Rx-Ry)	D^2
1.	28		12			
2.	30		23			
3.	32		30			
4.	13		12			
5.	16		17			
6.	18		7			
7.	14		13			
8.	12		14			
9.	18		16			
10.	22		11			
11.	23		10			
12.	25		25			
13.	31		18			
14.	19		22			
15.	18		12			
16.	10		6			
17.	29		30			
18.	18		21			
19.	19		19			
20.	23		10			
21.	25		21			
22.	27		15			

a. The first step is to rank the scores on each variable. Assign a one to the highest score and the lowest score should have the rank of N. When you have tie scores, assign each the average rank of the tied scores. Be sure to give the next score the next rank. For example:

X	R
8	1
9	2.5
9	2.5
10	4
11	5

Rank the X scores and the Y scores in the problem. Use the Rx and Ry columns.

b. Compute the difference between each Rx and Ry.

$D = Rx - Ry$

c. As a check compute the ΣD, that is, add all the D scores. They should sum to zero.

d. Square each D value and put in the D^2 column.

e. Get the sum of the D^2 column.

$\Sigma D^2 =$

f. The formula for the Spearman rho is:

$$\text{rho} = 1 - \frac{6\Sigma D^2}{N(N^2-1)}$$

Substitute the values for this problem into the equation.

g. Multiply 6 times ΣD^2.

h. Multiply N times $(N^2 - 1)$

i. Divide the product obtained in g by the product obtained in h.

j. Subtract the quotient obtained in i from 1.000.

rho =

Our confidence in stating that the obtained rho indicates a definite relationship between the variables can be determined from a table. Not all statistics books have tables of significance for rho but Popham does on Page 397. Entering the table with our N of 22 we see that .508 is the value under the .01 column. Because the table is a "one-tailed" table we should double the column headings in this situation and think of that column as the .02 column. This table indicates that only two times in one hundred would we get a rho of .508 or larger by chance. Our rho is larger than .508 so we conclude that it is very likely that there is a relationship between the variables. In statistical jargon we are confident at the .02 level that there is such a relationship.

Answers on Spearman rho problem

Student	Torrance Test		Problem Solving Test			
	X	Rx	Y	Ry	D	D^2
1.	28	5	12	16	-11	121
2.	30	3	23	4	- 1	1
3.	32	1	30	1.5	- .5	.25
4.	13	20	12	16	4	16
5.	16	18	17	10	8	64
6.	18	15.5	7	21	- 5.5	30.25
7.	14	19	13	14	5	25
8.	12	21	14	13	8	64
9.	18	15.5	16	11	4.5	20.25
10.	22	11	11	18	- 7	49
11.	23	9.5	10	19.5	-10	100
12.	25	7.5	25	3	4.5	20.25
13.	31	2	18	9	- 7	49
14.	19	12.5	22	5	7.5	56.25
15.	18	15.5	12	16	- .5	.25
16.	10	22	6	22	0	0
17.	29	4	30	1.5	2.5	6.25
18.	18	15.5	21	6.5	9	81
19.	19	12.5	19	8	4.5	20.25
20.	23	9.5	10	19.5	-10	100
21.	25	7.5	21	6.5	1	1
22.	27	6	15	12	- 6	36
					$\Sigma D=0$	$\Sigma D^2 = 861$

f. $\text{rho} = 1 - \frac{(6)(861)}{22(22^2 - 1)}$

g. (6)(861) = 5166

h. (22)(484-1) = (22)(483) = 10,626

i. $\frac{5166}{10{,}626} = .486$

j. rho = 1 - .486

rho = .514

COMPUTATION OF "t-test" AND MANN-WHITNEY "U"

These two tests are used in situations in which we have two groups of people and we want to determine whether they differ on some variable. For the "t-test" we assume that the groups are random samples drawn independently from a population, that the characteristic is normally distributed in the population, and that the scale used to measure the variable is an interval scale. The Mann-Whitney "U" is based on the assumptions that the groups are random samples drawn independently from a population and that the scale used is an ordinal scale. No assumption is made about how the characteristic is distributed in the population for the Mann-Whitney "U".

We will work through a problem doing the "t-test" first and then do the Mann-Whitney "U" for the same data.

An evaluator is comparing the achievement on a lesson in elementary science in which two different sets of materials have been used. One group has a programmed lab exercise and the other group used the regular lab manual. The materials were assigned to the students on a random basis. The achievement test scores of the two groups are shown below. There were 21 students in each group.

Program		Manual	
X_1	X_1^2	X_2	X_2^2
10		18	
17		16	
15		20	
17		25	
21		24	
18		19	
12		19	
10		16	
19		22	
8		16	
26		26	
21		26	
13		21	
24		26	

Program		Manual	
X_1	X_1^2	X_2	X_2^2
18		26	
12		20	
22		21	
13		17	
25		28	
19		17	
15		16	

The steps for computing t are listed below. As you work each step you can check your results with the answers on the answer sheet.

a. Square each X value in both groups. Use a table of squares.

b. Add the four columns. This will give you the ΣX_1, $\Sigma X_1{}^2$, ΣX_2, and $\Sigma X_2{}^2$

c. Divide the ΣX_1 by N_1 and the ΣX_2 by N_2. This will give you the mean score for each group.

$$\overline{X}_1 = \frac{\Sigma X_1}{N_1} = \qquad ; \quad \overline{X}_2 = \frac{\Sigma X_2}{N_2} =$$

d. Compute the $\Sigma x_1{}^2$ and $\Sigma x_2{}^2$. Notice these are the lower case letters and refer to the sum of deviation scores squared. These values are sometimes called the sum of squares. The general formula is:

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

For the $\Sigma x_1{}^2$ then it is

$$\Sigma x_1{}^2 = \Sigma X_1{}^2 - \frac{(\Sigma X_1)^2}{N_1}$$

Substitute the known values of X_1 into the equation.

e. Let's go through the equation step by step. First square the ΣX_1

$(\Sigma X_1)^2 =$

f. Divide the $(\Sigma X_1)^2$ obtained in e by N_1.

$$\frac{(\Sigma X_1)^2}{N_1} =$$

g. Subtract the quotient obtained in f from ΣX_1^2

$\Sigma x_1^2 =$

h. Now solve for the Σx_2^2 following steps d, e, b, and g using the values of X_2

i. Write the indicated values.

$\overline{X}_1 =$

$\overline{X}_2 =$

$\Sigma x_1^2 =$

$\Sigma x_2^2 =$

$N_1 =$

$N_2 =$

j. A formula for "t" is:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\Sigma x_1^2 + \Sigma x_2^2}{N_1 + N_2 - 2} \left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

Substitute the values into the equation.

k. Let's solve the denominator. First carry out all additions and subtractions in the denominator.

l. Next multiply the obtained fractions.

m. Now change the fraction to a decimal.

n. Rewrite the equation for t using the value obtained in m.

o. Find the square root of 2.0372

p. Subtract $\overline{X}_2$ from $\overline{X}_1$

q. Rewrite the "t" equation

r. Compute "t" as a decimal

The fact that the "t" value is negative need not concern you. It is negative because we subtracted the largest mean from the smallest. We could have just as well subtracted the means the other way and the "t" would have been positive. In the rest of the problem we will discuss the "t" as though it were positive.

Obviously there is a difference in the means of the two groups. How confident can we be that the observed difference is a real difference and not just chance? Our "t" value can help us establish our degree of confidence. To do this we will use a table found in most statistics books. This table is usually titled something like "Distribution of t".

To use the table we first need to get the degrees of freedom (d.f.). Degrees of freedom are equal to $N_1 + N_2 - 2$.

s. Now find the degrees of freedom for the problem.

$$\text{d.f.} = N_1 + N_2 - 2 =$$

Now enter the table with the obtained d.f. We see in the table, using the levels for a two-tail test that the "t" associated with .02 is 2.423 and with .01 it is 2.704. This means that the likelihood of getting means that differ so much that we get at t of 2.423 is only 2 in 100 by chance alone. In other words, such a large t wouldn't occur very often by chance. Our "t" is greater than 2.704. Consequently we would conclude that our groups do differ because it isn't very likely that we would have gotten such a large "t" if they didn't differ. We would conclude that they differ at the .01 level of confidence.

Computation of Mann-Whitney "U"

The data below are the same as were used for the t-test

Program		Manual	
X_1	R_1	X_2	R_2
10		18	
17		16	
15		20	
17		25	
21		24	
18		19	
12		19	
10		16	
19		22	
8		16	
26		26	
21		26	
13		21	
24		26	
18		26	
12		20	
22		21	
13		17	
25		28	
19		17	
15		16	

a. The first thing to do is rank all the scores together giving the rank of 1 to the highest score and the rank of $N_1 + N_2$ to the lowest score, in this case 42. For tie scores, assign them all the average rank. The next score should be the next rank. For example:

X_1	R_1	X_2	R_2
8	8	9	7
10	5	10	5
10	5	11	2.5
11	2.5	12	1

b. Next get the sum of the two ranks columns.

ΣR_1 =

ΣR_2 =

c. Next compute U and U^1

$$U = N_1N_2 + \frac{N_1(N_1+1)}{2} - \Sigma R_1$$

d. $U^1 = N_1N_2 - U$

For the rest of the computation we use the smaller value of U and U^1. In this case we use U because it is smaller than U^1.

e. Now we compute z where

$$z = \frac{1/2\ (U-(N_1N_2))}{\sqrt{\frac{(N_1)(N_2)(N_1+N_2+1)}{12}}}$$

We then go to a table of the normal curve to get an indication of the likelihood that the groups are the same. z values of ±2.73 includes about 99.36% of the area of the curve between them. 64% or less than 1% of the area of the curve lies beyond z values of ±2.73. This means that the likelihood of getting a z as large as 2.73 by chance is less than 1 in 100. Since the likelihood of this outcome occurring by chance is so small we conclude that the difference between the two groups is a real difference.

There were many tie scores in this problem. There is a procedure for correcting for ties that increases the sensitivity of the Mann-Whitney U. This procedure is described in Siegel, Nonparametric Statistics.

Answers on t-test:

Program		Manual	
X_1	X_1^2	X_2	X_2^2
10	100	18	324
17	289	16	256
15	225	20	400
17	289	25	625
21	441	24	576
18	324	19	361
12	144	19	361
10	100	16	256

Program		Manual	
X_1	X_1^2	X_2	X_2^2
19	361	22	484
8	64	16	256
26	676	26	676
21	441	26	676
13	169	21	441
24	576	26	676
18	324	26	676
12	144	20	400
22	484	21	441
13	169	17	289
25	625	28	784
19	361	17	289
15	225	16	256
$\Sigma X_1=355$	$\Sigma X_1^2=6531$	$\Sigma X_2=439$	$\Sigma X_2^2=9503$

c. $\overline{X}_1 = \frac{355}{21} = 16.90; \quad \overline{X}_2 = \frac{439}{21} = 20.90$

d. $\Sigma x_1^2 = 6531 - \frac{(355)^2}{21}$

e. $(\Sigma X_1)^2 = (355)^2 = 126{,}025$

f. $\frac{(X_1)^2}{N_1} = \frac{(355)^2}{21} = \frac{126{,}025}{21} = 6001.19$

g. $\Sigma x_1^2 = 6531 - 6001.19$

$\Sigma x_1^2 = 529.81$

h. $\Sigma x_2^2 = 9503 - \frac{(439)^2}{21}$

$(439)^2 = 192{,}721$

$\frac{192{,}721}{21} = 9177.19$

$\Sigma x_2^2 = 9503 - 9177.19$

$\Sigma x_2^2 = 325.81$

i. $\overline{X}_1 = 16.90$

$\overline{X}_2 = 20.90$

$\Sigma x_1^2 = 529.81$

$\Sigma x_2^2 = 325.81$

$N_1 = 21$

$N_2 = 21$

j. $t = \dfrac{16.90 - 20.90}{\sqrt{(\frac{529.81 + 325.81}{21+21-2})(\frac{1}{21} + \frac{1}{21})}}$

k. $(\frac{529.81 + 325.81}{21+21-2})(\frac{1}{21} + \frac{1}{21}) = (\frac{855.62}{40})(\frac{2}{21})$

l. $(\frac{855.62}{40})(\frac{2}{21}) = \frac{1711.24}{840}$

m. $\frac{1711.24}{840} = 2.0372$

n. $t = \dfrac{16.90 - 20.90}{\sqrt{2.0372}}$

o. $\sqrt{2.0372} = 1.43$

p. $\overline{X}_1 - \overline{X}_2 = 16.90 - 20.90 = -4.00$

q. $t = \dfrac{-4.00}{1.43}$

r. $t = - 2.80$

s. d.f. = 40

Answers on Mann-Whitney U

Program		Manual	
X_1	R_1	X_2	R_2
10	40.5	18	24
17	27.5	16	31.5
15	34.5	20	17.5
17	27.5	25	7.5
21	14.5	24	9.5
18	24	19	20.5
12	38.5	19	20.5
10	40.5	16	31.5
19	20.5	22	11.5
8	42	16	31.5
26	4	26	4
21	14.5	26	4
13	36.5	21	14.5
24	9.5	26	4

Program		Manual	
X_1	R_1	X_2	R_2
18	24	26	4
12	38.5	20	17.5
22	11.5	21	14.5
13	36.5	17	27.5
25	7.5	28	1
19	20.5	17	27.5
15	34.5	16	31.5
	$\Sigma R_1 = 547.5$		$\Sigma R_2 = 355.5$

c. $$U = N_1N_2 + \frac{N_1(N_1 + 1)}{2} - \Sigma R_1$$

$$U = (21)(21) + \frac{(21)(21+1)}{2} - 547.5$$

$$U = 441 + \frac{462}{2} - 547.5$$

$$U = 441 + 231 - 547.5$$

$$U = 672 - 547.5$$

$$U = 124.5$$

d. $$U^1 = 441 - 124.5$$

$$U^1 = 316.5$$

e. $$z = \frac{1/2(124.5) - (21 \cdot 21)}{\sqrt{\frac{(21)(21)(21+21+1)}{12}}}$$

$$z = \frac{1/2(124.5-441)}{\sqrt{\frac{(441)(43)}{12}}}$$

$$z = \frac{1/2\ (-316.5)}{\sqrt{\frac{18963}{12}}}$$

$$z = \frac{-108.25}{\sqrt{1580.25}}$$

$$z = \frac{-108.25}{39.7}$$

$$z = -2.73$$

COMPUTATION OF CORRELATED t-test

The correlated t-test is used when we want to determine whether the means of two sets of scores differ and the scores are correlated. This statistic is very commonly used when one has given a pretest before a lesson or unit and then the same test as a posttest. Since each person has two scores, one on the pretest and one on the posttest, we can expect the two sets of scores to be correlated. We can allow for this correlation in the scores with the correlated t-test.

The following data were obtained by giving an attitude scale to a class before and after a unit. The attitude scale was designed to measure attitude toward programmed instruction which was the method for presenting the material. The question is whether the students' attitude toward programmed instruction changed from pretest to posttest.

Student	Pretest (X)	Posttest (Y)	D=(Y-X)	D^2
1.	76	81		
2.	71	85		
3.	57	52		
4.	49	52		
5.	70	70		
6.	69	72		
7.	26	33		
8.	65	83		
9.	59	58		
10.	42	56		

a. First compute D for each pair of scores. Subtract Y-X.

b. Square each D value.

c. Get the sum of each column.

$\Sigma X=$

$\Sigma Y=$

$\Sigma D=$

$\Sigma D^2=$

d. Compute the mean of X and the mean of Y

$$\overline{X} = \frac{\Sigma X}{N} =$$

$$\overline{Y} = \frac{\Sigma Y}{N} =$$

e. Compute the Σd^2 where

$$\Sigma d^2 = \Sigma D^2 - \frac{(\Sigma D)^2}{N}$$

Substitute the known values into the equation.

f. $(\Sigma D)^2 =$

g. Divide the value obtained in f by N.

$$\frac{(\Sigma D)^2}{N} =$$

h. Subtract the quotient obtained in g from ΣD^2

$\Sigma d^2 =$

i. A formula for the correlated t-test is

$$t = \frac{\overline{Y} - \overline{X}}{\sqrt{\frac{\Sigma d^2}{N(N-1)}}}$$

Substitute the known values into the formula

j. Subtract $\overline{Y} - \overline{X}$

k. Solve the fraction under the square root and make it a decimal. Carry to four places.

l. Take the square root of the answer obtained in k.

m. Rewrite the t formula with value obtained in j as numerator and value obtained in 1 as denominator.

n. Convert the fraction in m to a decimal

$t =$

To evaluate this "t" we use the table of t found in most statistics books. We first find the degrees of freedom (d.f.). The degrees of freedom for a correlated t-test are the number of pairs of scores minus one.

o. Compute the d f for this problem.

With nine degrees of freedom we enter the table and observe a value of 2.262 under the .05 column and a value of 4.032 under the .01 column. Our obtained t is between these two values. This means that the likelihood is somewhat less than 5 in 100 that we would have gotten differences as great as those obtained purely by chance. We conclude that there was a change in attitude scores at the .05 level of confidence.

Answers for correlated "t" test

Student	Pretest (X)	Posttest (Y)	D(Y-X)	D^2
1.	76	81	5	25
2.	71	85	14	196
3.	57	52	-5	25
4.	49	52	3	9
5.	70	70	0	0
6.	69	72	3	9
7.	26	33	7	49
8.	65	83	18	324
9.	59	58	-1	1
10.	42	56	14	196
	$\Sigma X=584$	$\Sigma Y=642$	$\Sigma D=58$	$\Sigma D^2=834$

d. $\overline{X} = \frac{584}{10} = 58.4$

$\overline{Y} = \frac{642}{10} = 64.2$

e. $\Sigma d^2 = 834 - \frac{(58)^2}{10}$

f. $(\Sigma D)^2 = (58)^2 = 3364$

g. $\frac{(\Sigma D)^2}{N} = \frac{3364}{10} = 336.4$

h. $\Sigma d^2 = 834 - 336.4$

$\Sigma d^2 = 497.6$

i. $t = \frac{64.2 - 58.4}{\sqrt{\frac{497.6}{10(10-1)}}}$

j. $\overline{Y} - \overline{X} = 5.8$

k. $\frac{\Sigma d^2}{N(N-1)} = \frac{497.6}{90} = 5.5289$

l. $\sqrt{5.5289} = 2.35$

m. $t = \frac{5.8}{2.35}$

n. $t = 2.468$

o. d.f. = 9

COMPUTATION OF ONE-WAY AND TWO-WAY ANALYSIS OF VARIANCE (ANOVA)

The analysis of variance technique is used when we want to compare two or more groups of people on one variable. When we have two groups to compare, the analysis of variance and the "t" test give the same results. Most statistics books will indicate situations where "t" might be used instead of ANOVA, but generally either can be used when two groups are compared An advantage of ANOVA is that we can use it to determine whether there is a difference among three or more groups on some variable with a single test. This is impossible with the "t" test. Another feature of the ANOVA technique is that we can classify our students in two or more ways and determine not only whether there are differences among the groups on each classification, but also whether there are dependencies or interactions between the classifications.

The first problem will be a situation where we have three groups of people and we want to know whether they differ on some variable.

The data were achievement test scores over a unit on biology in an elementary science course for the gifted. The three groups were made up of students who used three different sets of materials in studying the unit. The evaluator was concerned with evaluating the materials.

Set I		Set II		Set III	
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
10		18		10	
13		11		12	
17		13		19	
15		18		16	
16		15		13	
15		15		14	
9		16		14	
12		16		14	
14		14		12	
14		17		13	

a. First square each score Use a table of squares Check your results with the answer sheet

b Get the sum of all six columns

$\Sigma X_1 =$ $\Sigma X_2 =$ $\Sigma X_3 =$

$\Sigma X_1^2 =$ $\Sigma X_2^2 =$ $\Sigma X_3^2 =$

c. Compute the mean for each group

$\bar{X}_1 = \frac{\Sigma X_1}{N_1} = \frac{135}{10} = 13.5$ $\bar{X}_2 =$ $\bar{X}_3 =$

d. Add $\Sigma X_1 + \Sigma X_2 + \Sigma X_3 =$ $= \Sigma X_t$

$\Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 =$ $= \Sigma X_t^2$

$N_1 + N_2 + N_3 =$ $= N_t$

The first sum is the sum of the raw scores for the total group.

The second sum is the sum of the raw scores squared for the total group.

The third sum is the total number of people

e. Next compute Σx_t^2. Notice that the letter is lower case. This is the sum of deviation scores squared and is often referred to as the sum of squares for total. The formula is:

$$\Sigma x_t^2 = \Sigma X_t^2 - \frac{(\Sigma X_t)^2}{N_t}$$

Substitute the known values into the equation.

f. Solve the equation

$\Sigma x_t^2 =$

g. Next we will solve for a value called sum of squares among groups. The formula is:

$$\Sigma x_a^2 = \frac{(\Sigma X_1)^2}{N_1} + \frac{(\Sigma X_2)^2}{N_2} + \frac{(\Sigma X_3)^2}{N_3} + \frac{(\Sigma X_t)^2}{N_t}$$

Substitute into the equation

h. Solve the equation

Σx_a =

i. The next thing to compute is the sum of squares for within. The formula is:

$\Sigma x_w^2 = \Sigma x_t^2 - \Sigma x_a^2$

Substitute into the equation and solve.

j. We now have almost everything we need for the ANOVA. The ANOVA table looks like this:

Source of Variance	Degrees of Freedom	Sum of Squares	Mean Square	F
Among groups		19.47		
Within groups		156.70		
Total		176.17		

k. The above table contains what has been computed thus far. The next thing we need to get is the degrees of freedom (d.f.).

The degrees of freedom for total is equal to $N_t - 1$. Write it in.

The degrees of freedom for among groups is the number of groups minus one. Write it in.

The degrees of freedom for within groups is the degrees of freedom for total minus the degrees of freedom for among groups. Write it in.

l. Next we compute the Mean Square. We do this only for Among and Within groups. The Mean Square is equal to the Sum of Squares divided by the degrees of freedom. Thus for Among groups the Mean Square = $\frac{19.47}{2}$ = and for Within groups the Mean Square =

Compute and write in the table.

m. Finally we get the F value which is obtained by

$$F = \frac{\text{Mean Square Among Groups}}{\text{Mean Square Within groups}}$$

$F =$

Compute and write in the table

The F value is what is used to determine the significance of the results. For this we use a table found in most statistics books which is usually called the F table To use the table we first have to determine our degrees of freedom and with the F table it needs to be two different degrees of freedom. The degrees of freedom that we use are the degrees of freedom for among groups and the degrees of freedom for within groups. These are 2 and 27 respectively for our problem. We use the 2 or d.f. for among groups to determine our column in the F table and the 27 or d.f. for within groups for our row. Looking at the intersection of that column and row we find two numbers, 3.35 and 5.49. These numbers mean that only 5 times in 100 would we get an F value of 3.35 or larger by chance and only 1 time in 100 would it be 5.49 or larger by chance. The f value in the problem is considerably less than 3.35. Consequently, we would conclude that even though the means of the three groups did differ, this difference isn't large enough for us to assert that it is a real difference. Such a difference could occur by chance quite often and it would be too risky to bet that there is any difference among the three groups.

Computation of two-way ANOVA

The data below are from a two-way ANOVA.

An evaluator had developed an attitude scale for adults to complete which would measure their attitude toward the gifted program. He was interested in finding out whether parents of students in the program differed from parents of students not in the program in their attitudes and also whether fathers differed from mothers. He was careful to get data from only one parent. The data were as follows:

	Parents of Children in Gifted Program		Parents of Children Not in Gifted Program	
	X_1	X_1^2	X_3	X_3^2
Fathers	66		48	
	33		25	
	84		74	
	80		87	
	72		68	
	57		44	
	82		78	
	51		58	
	X_2	X_2^2	X_4	X_4^2
Mothers	12		15	
	57		57	
	82		62	
	62		49	
	22		17	
	56		37	
	38		21	
	68		61	

a. First square each score. Use a table of squares.

b. Next compute the following

ΣX_t^2 = Add all the squared scores together =

ΣX_t = Add all the raw scores =

ΣX_1 =

ΣX_2 =

ΣX_3 =

ΣX_4 =

ΣX fathers = $\Sigma X_1 + \Sigma X_3$ =

ΣX mothers = $\Sigma X_2 + \Sigma X_4$ =

ΣX parents in gifted = $\Sigma X_1 + \Sigma X_2$ =

ΣX parents not in gifted = $\Sigma X_3 + \Sigma X_4$ =

c. Next get the sum of squares for total.

$$\Sigma x_t^2 = \Sigma X_t^2 - \frac{(\Sigma X_t)^2}{N_t}$$

d. Now get the sum of squares between father and mother or for sex.

$$\Sigma x_s^2 = \frac{(\Sigma X_f)^2}{N_f} + \frac{(\Sigma X_m)^2}{N_m} - \frac{(\Sigma X_t)^2}{N_t}$$

e. Now get the sum of squares between parents of gifted and other parents.

$$\Sigma x_p^2 = \frac{(\Sigma X_g)^2}{N_g} + \frac{(\Sigma X_{ng})^2}{N_{ng}} - \frac{(\Sigma X_t)^2}{N_t}$$

f. The next sum of squares is the sum of squares for interaction. An interaction would be a situation where the fathers and the mothers would have a different pattern of responding. For example, if the fathers of the gifted had higher scores than the fathers of the other students, but the mothers' pattern were opposite then there would be an interaction.

$$\Sigma x_{sxp}^2 = \frac{(\Sigma X_1)^2}{N_1} + \frac{(\Sigma X_2)^2}{N_2} + \frac{(\Sigma X_3)^2}{N_3} + \frac{(\Sigma X_4)^2}{N_4} - \frac{(\Sigma X_t)^2}{N_t} - \Sigma x_s^2 - \Sigma x_p^2$$

g. The within sum of squares is

$$\Sigma x_w^2 = \Sigma x_t^2 - (\Sigma x_p^2 + \Sigma x_s^2 + \Sigma x_{sxp}^2)$$

h. Now we can build the ANOVA table.

Source of Variance	d.f.	Sum of Squares	Mean Square	F
Parent group				
Parent sex				
Group x sex interaction				
Within				
Total				

i. Put in the Sum of Squares

Next put in d.f.

df for total = total number of cases minus one

df for group = number of parent groups minus one

df for parent sex = number of sexes minus one

df for interaction = df for group times df for sex

df for within = df total - (df group + df sex + df interaction)

Next compute the Mean Squares for group, sex, interaction, and within by dividing the Sum of Squares by its corresponding df.

Next compute the F for group, sex, and interaction by dividing each Mean Square by the Mean Square for within.

We now evaluate the F by using the F table and finding the intersection of the column headed with a 1 and the row headed with a 28. At this place we find the numbers 4.20 and 7.64. These numbers mean that with 1 and 28 degrees of freedom we could expect to get an F of 4.20 or larger 5 times in 100 by chance and an F of 7.64 or larger 1 time in 100 by chance. Two of the obtained F values are much smaller than these, which means that the parents of the gifted and the parents of the other children seem to hold very similar attitudes toward the gifted program. Also there is no interaction of parental group with the sex of the parent. The F value for sex group was 6.22, however, which falls between the tabled values. This means that the fathers and the mothers differed enough in their scores that we wouldn't consider it a chance difference. The fathers apparently were more favorable toward the program than the mothers. We would conclude that the fathers differ from the mothers and our confidence level would be .05.

Answers to one-way ANOVA

Set I		Set II		Set III	
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
10	100	18	324	10	100
13	169	11	121	12	144
17	289	13	169	19	361
15	225	18	324	16	256
16	256	15	225	13	169
15	225	15	225	14	196
9	81	16	256	14	196
12	144	16	256	14	196
14	196	14	196	12	144
14	196	17	289	13	169
$\Sigma X_1=135$	$\Sigma X_1^2=1881$	$\Sigma X_2=153$	$\Sigma X_2^2=2385$	$\Sigma X_3=137$	$\Sigma X_3^2=1931$

c. $\overline{X}_2 = \frac{153}{10} = 15.3, \ \overline{X}_3 = \frac{137}{10} = 13.7$

d. $\Sigma X_t = 425$

$\Sigma X_t^2 = 6197$

$N_t = 30$

e. $\Sigma x_t^2 = 6197 - \frac{(425)^2}{30}$

f. $\Sigma x_t^2 = 6197 - \frac{180,625}{30}$

$\Sigma x_t^2 = 6197 - 6020.83$

$\Sigma x_t^2 = 176.17$

g. $\Sigma x_a^2 = \frac{(135)^2}{10} + \frac{(153)^2}{10} + \frac{(137)^2}{10} - \frac{(425)^2}{30}$

h. $\Sigma x_a^2 = \frac{18,225}{10} + \frac{23,409}{10} + \frac{18,769}{10} - \frac{180,625}{30}$

$\Sigma x_a^2 = 1822.5 + 2340.9 + 1876.9 - 6020.83$

$\Sigma x_a^2 = 6040.3 - 6020.83$

$\Sigma x_a^2 = 19.47$

i. $\Sigma x_w^2 = 176.17 - 19.47$

$\Sigma x_w^2 = 156.70$

j.

Source of Variance	Degrees of Freedom	Sum of Squares	Mean Square	F
Among groups	2	19.47	9.78	1.686
Within groups	27	156.70	5.80	
Total	29	176.17		

Answers to two-way ANOVA

	Parents of Children In Gifted Program		Parents of Children Not in Gifted Program	
	X_1	X_1^2	X_3	X_3^2
	66	4356	48	2304
	33	1089	25	625
	84	7056	74	5476
Fathers	80	6400	87	7569
	72	5184	68	4624
	57	3249	44	1936
	82	6724	78	6084
	51	2601	58	3364
	X_2	X_2^2	X_4	X_4^2
	12	144	15	225
	57	3249	57	3249
	82	6724	62	3844
Mothers	62	3844	49	2401
	22	484	17	289
	56	3136	37	1369
	38	1444	21	441
	68	4624	61	3721

b. ΣX_t^2 = 107,829

ΣX_t = 1723

ΣX_1 = 525

ΣX_2 = 397

ΣX_3 = 482

ΣX_4 = 319

ΣX fathers = 1007

ΣX mothers = 716

ΣX gifted = 922

ΣX not gifted = 801

c. $\Sigma x_t{}^2 = 107{,}829 - \frac{(1723)^2}{32}$

$\Sigma x_t{}^2 = 107{,}829 - \frac{2{,}968{,}729}{32}$

$\Sigma x_t{}^2 = 107{,}829 - 92{,}772.78$

$\Sigma x_t{}^2 = 15{,}056.22$

d. $\Sigma x_s{}^2 = \frac{(1007)^2}{16} + \frac{(716)^2}{16} - \frac{(1723)^2}{32}$

$\Sigma x_s{}^2 = \frac{1{,}014{,}049}{16} + \frac{512{,}656}{16} - \frac{2{,}968{,}729}{32}$

$\Sigma x_s{}^2 = \frac{1{,}526{,}705}{16} - 92{,}772.78$

$\Sigma x_s{}^2 = 95{,}419.06 - 92{,}772.78$

$\Sigma x_s{}^2 = 2{,}616.28$

e. $\Sigma x_p{}^2 = \frac{(922)^2}{16} + \frac{(801)^2}{16} - \frac{(1723)^2}{32}$

$\Sigma x_p{}^2 = \frac{850{,}084}{16} + \frac{641{,}601}{16} - \frac{2{,}968{,}729}{32}$

$\Sigma x_p{}^2 = \frac{1{,}491{,}685}{16} - 92{,}772.78$

$\Sigma x_p{}^2 = 93{,}230.31 - 92{,}772.78$

$\Sigma x_p{}^2 = 457.53$

f. $\Sigma x_{sxp}{}^2 = \frac{(525)^2}{8} + \frac{(397)^2}{8} + \frac{(482)^2}{8} + \frac{(319)^2}{8} - \frac{(1723)^2}{32} - 2{,}646.28 - 457.53$

$\Sigma x_{sxp}{}^2 = \frac{275{,}625}{8} + \frac{157{,}609}{8} + \frac{232{,}324}{8} + \frac{101{,}761}{8} - 92{,}772.78 - 3103.81$

$\Sigma x_{sxp}{}^2 = \frac{767{,}319}{8} - 95{,}876.59$

$\Sigma x_{sxp}{}^2 = 95{,}914.88 - 95{,}876.59$

$\Sigma x_{sxp}{}^2 = 38.29$

g. $\Sigma x_w{}^2 = 15{,}056.22 - (2.646.28 + 457.53 + 38.29)$

$\Sigma x_w{}^2 = 15{,}056.22 - 3142.10$

$\Sigma x_w{}^2 = 11{,}914.12$

h.

Source of Variance	d f	Sum of Squares	Mean Squares	F
Parent group	1	457 53	457 53	1.075
Parent sex	1	2 646 28	2,646.28	6.219
Group X sex	1	38 29	38.29	.090
Within	28	11 914 12	425 50	
Total	31	15 056 22		

EVALUATION PLAN IV

Evaluation Question: Has the gifted program had an effect on the achievement of the participating students?

Situation

A school situation as described in the following paragraphs is the context in which this plan might be operational.

The school system (let's call it Great) has one senior high school with about 800 students and 100 teachers, staff, and support people. The school district might be considered to be toward the progressive side of some progressive-traditional continuum because of its attempts to work with special programs such as gifted, handicapped, tutorial sessions for culturally disadvantaged, special art classes, and other things. The school board and the administration have decided that the total school program should be "evaluated" during the 1968-69 school year. On the basis of this evaluation a new policy and program statement will be developed for implementation in the 1970-71 school year. The gifted program that is operated in the school will be included in the evaluation.

The gifted program in the school is essentially a program for accelerating the progress of certain students in mathematics, English, and science. The students are identified in the ninth grade and enrolled in one or more of the accelerated classes in Grade 10. The students will usually remain in the accelerated class through their high school career. There are three accelerated classes at each grade level, one for each of the subject matter areas. It is intended that no more than 25 students be in any accelerated class.

There are approximately 50 students in the accelerated program at each grade level, with some students being in all three accelerated classes, others in two, and others in one.

The teachers of the accelerated classes teach one or two such classes as part of their teaching load. The assignments are as follows:

English

Miss Jones - 10th and 11th grade

Mrs. Bright - 12th grade

Science

Mr. Carson - 10th grade

Mr. Steel - 11th grade

Mrs. White - 12th grade

Mathematics

Miss Pearl - 10th grade

Mr. Wilson - 11th and 12th grade

Mr. Wilson has the responsibility for coordinating the program. He has a planning period for this responsibility in addition to his regularly assigned planning period. He has two students who do routine clerical work for him on a work-study program. He can get a limited amount of clerical time from the office (average 3 hours a week), and the three counselors in the school assist with testing and keeping the cumulative records.

Mr. Wilson, as program coordinator, has been given the responsibility for providing information about the gifted program for the overall evaluation.

He has not been given any additional time for this function, but he has been assured of some additional clerical help (4 hours a week) and a budget of $750 for the purchase of supplies and tests, test scoring, and data processing.

Naturally Mr. Wilson wants to provide as complete information about the program as possible to the board. He does have limits to what he can do so he must decide on some priorities. He considers that information about the achievement of the participants in the program should have a high priority so he decides to develop a plan for obtaining this information first. The amount of additional information that he will plan to obtain will depend on the time and money that remain after the achievement information plan has been developed.

Mr. Wilson is taking graduate work at the University of Illinois. He took a course with a fellow named Bob Stake who talked a lot about evaluation of educational programs and who had developed a model for such evaluation. This model seemed useful to Mr. Wilson as a guide to follow in planning evaluation so he tried to plan the evaluation of the gifted program with the model. His plan follows.

Rationale

Every educational activity has a rationale and/or a set of assumptions on which it is based. Usually the rationale and the assumptions are not stated. One wonders what educational activities would be thrown out immediately if the rationale and assumptions were stated and, when expressed, their irrationality becomes obvious.

A statement of program rationale is important for the evaluator. It helps him know the purposes and the assumptions of the program and provides a base for the evaluation effort as well as the program. The statement of rationale for the Great program is as follows:

"The essence of a true democracy is that every individual in that society be free and have the opportunity to develop his abilities and talents to the maximum. Provision and exploitation of such opportunities enhance the well-being of the individual and the society. As individuals are able to develop to their maximum capability, so will the society of these individuals develop to its maximum. The less-talented in the society will benefit if the more-talented are allowed to develop maximally because the less-talented will benefit from the creative endeavors of the more-talented that are facilitated by the exploitation and development of the talents. Likewise, the more-talented will benefit if the less-talented develop maximally because the less-talented, if developed fully, will be able to implement the advances in knowledge with minimal direction and supervision.

The above paragraph implies and this school subscribes to the belief that education in a democracy must continually strive toward a system in which each person is educated to a level and at a pace that is suited to him. This in contrast to a mistaken belief that democractic education means all receive equal education at the same pace. This school interprets equality of education to mean equality of educational opportunity rather than an equal education for all.

Although completely individualized instruction is considered the optimal educational situation, the technology for attaining this ideal is not yet available. A workable approximation to individualized instruction is to identify students who have certain talents, handicaps, and interests and provide instruction for them in groups formed on the basis of exhibited common factors.

Persons with special talents and abilities are a valuable resource to society. In this technological age, it is of paramount importance that those talents and abilities that are especially relevant to technology be developed. It is for this reason that the Great schools have developed a special accelerated program in science, mathematics, and English in the high school. This program is designed to allow students with a high level of ability and interest in these areas to develop deeper understandings and work at a faster pace than they would be able to do in the regular program.

Group instruction implies teaching to the norm of the groups. If the group is homogeneous, however, the norm is very representative of each person in the group so that the instruction approximates individual instruction. By grouping students of high ability together, by providing an excellent teacher for the group, and by providing teaching materials commensurate with the ability and interests of the group, an effective program for developing the unique abilities of the persons in the group can be developed. This is the basis for the Great High School program for the gifted."

Intents

Most textbooks on evaluation stress the point that the first step in evaluation of a program is to state the objectives of the program in behavioral terms. The Stake model implies a similar starting point in the Intents column. Certainly the intended outcomes of a program would be the same thing as the objectives of the program.

The model provides for other kinds of intents than objectives, however. The intended antecedents, transactions and the contingencies between and among the three cells are as important to specify as the objectives. Stake also places less emphasis on behavioral terms than do many other writers.

The intents for the Great High School program were specified as follows:

Intended Antecedents

1. The students who participate in the program will have a high level of ability in dealing with cognitive tasks and an interest in participating in the program.
2. The teachers in the program will be highly knowledgeable of the subject matter, will have demonstrated an unique ability to work with gifted children, will be innovative and adaptive. and will have indicated a high level of interest in working in the program.
3. The administration and school board will have made a commitment to support the program.
4. Adequate facilities and materials will be available to the program.
5. The community will have been informed about the program and will have indicated acceptance.
6. The State Department of Education will have approved the program for support in the reimbursed program.

Intended Transactions

1. The students will interact with each other, with the teachers, and with other resource persons in a variety of ways. Discussions among students will be encouraged to exchange and challenge ideas. The teacher will lecture, converse, and advise students as the situation suggests. Resource persons will be brought to the class to present material and students will be encouraged to interact with available resource people in the school and community.

2. A large collection of instructional materials will be readily available for student and teacher usage. Books, kits, programs, films, journals, maps, charts, syllabi, etc. are the kinds of materials. These will be stored so that student access is easily attained.

3. Adequate equipment will be available to allow optimal use of materials. Laboratory and audio-visual equipment are examples of this. Provision for easy student usage will be made.

4. The classroom procedures will be designed to maximize individual problem solving activity.

5. The teachers in the program will meet regularly to discuss the operation of the program, participate in in-service activities either individually or as a group, and continue to translate and interpret

the program to the school staff and the community.

6. The students in the program will participate at a level consistent with their ability and interest in classes and school activities other than those in the program.

Intended Outcomes

1. The participating students will learn the material that is presented as the required material in each course.
2. The participating students will each exhibit a capability for independent study.
3. Each participating student will learn and use problem solving techniques.
4. Each participating student will exhibit the attitudes that each course is designed to develop.
5. The students will develop and maintain normal social relationships.
6. The teachers in the program will exhibit increased understanding of and resourcefulness in working with gifted students.
7. The patrons of the school and the administration will continue to support the program.
8. The students and teachers of the school who are not in the program will have positive feelings toward the program and the participants.

Obviously, some of the stated intents are not directly related to the achievement question except in a very tangential manner. They were written to illustrate the variety of intents that might be included in a complete evaluation plan. The listed intents are not exhaustive of all possible intents, however. The rest of the plan will deal only with the achievement question.

The following statements are attempts to specify the logical contingencies among the three intent types.

1. Individuals with a high level cognitive ability when provided the opportunity to interact with competent resources under the guidance of a competent teacher will learn well the material of the course.
2. Individuals with a high level of cognitive ability and interest in the area when provided adequate resources will develop a capability for independent study.
3. Teachers who are capable of independent problem solving can structure materials and situations to develop the problem solving skills of students.
4. Competent and enthusiastic teachers when interacting with competent students will have an effect on the attitudes of the students.

Observations

The observations cells contain the specification of the ways by which the intents will be observed or measured. In effect the observations are the operational definitions of the intents. A specification of kinds of observation for each of the intents related to achievement is listed below:

Observed Antecedents

1. The cognitive ability of the students is determined by their scores on the California Test of Mental Maturity and an estimate of ability from performance in Junior High School. The latter is obtained from grades and teacher judgement. The CTMM is used by the schools in the testing program and seems to provide a quite reliable indication of ability of ninth grade students. The CTMM is administered in the first quarter of the second semester of the ninth grade.

2. The interest of the students in participating in the program is determined by their expressed interest when working on their high school schedules with the junior high counselor. The Kuder Personal Preference Scale is another indicator of interest areas that is used. This scale is administered in the ninth grade.

3. The teachers' knowledge of the subject matter is determined by an examination of their college transcripts in terms of courses taken and grades. The ability of the teacher and his interest in

working with gifted children is presently based on testimony of students and self-report of the teachers. Mr. Wilson is interested in obtaining better data on these factors because some of the teachers now in the program do not seem to be as effective as desired.

4. A record of the facilities and materials is available from the office records. An inventory of teacher-owned materials, materials brought by students, and materials and resources used from the community will be obtained with a questionnaire to the teachers.

Observed Transactions

1. Usage of materials will be determined by examining check-out records such as in the library, wear and tear on materials and equipment with rating scales, and observations by the teacher. Self-reports of material and equipment usage will be obtained from the students.
2. The interaction of students with each other, teachers, and other people will be obtained by:
   a. Observing the classroom situation on a random basis for 1/2 hour each month.
   b. Having the teachers keep logs of the activity in the class one day each week. The day will

be randomly assigned and a schedule will be developed for the teacher.

c. Reports of the activities of the class by each teacher using a common report form.

d. Self-reports of the students.

3. Participation of the students in activities other than the gifted program will be obtained by self-reports on a common reporting form.

Observed outcomes

1. The learning outcomes need to be discussed in terms of individual courses in that there are nine separate sets of learning outcomes. Mr. Wilson decides he will have to rely heavily on the information that can be obtained from the final testing period. He asks each teacher to make out or select the measuring instruments to be used for measuring learning. He does assist in selection and/or development of the instruments. The measures for the learning outcomes in each course are as follows:

a. 10th Grade English - This is a composition course. Each student writes two major papers for this course and several short themes. Analysis of these papers will be one basis for measuring outcomes. The papers will be analyzed with rating scales and with some "natural language" measures. English usage is emphasized in the course so another instrument for assessing outcomes to be used in the ITED test on English usage. This is a reliable instrument for this purpose.

b. 11th Grade English - This is an American Literature and creative writing course.

Each student writes a minimum of three short stories or poems in the class. The papers will be examined with the "natural language" system and other appropriate rating scales. Achievement in the literature portion of the course will be measured with a teacher-made test because no appropriate commercial test is available.

c. 12th Grade English - This is a "Great Books" type of course. Achievement in this course will be determined by a teacher-made test and by observation of analyses the students will write of the books they read. In addition all students in this course will take the CEEB tests as part of the school testing program.

d. 10th Grade Math - 2nd year algebra.

The essential difference between this course and the same course taught in the regular program is that this course treats each topic more completely using more complex problems and special topics. None of the commercial test in this area appears to sample adequately the complex problems used in the course. It was decided that student performance on the

Cooperative Math Test for Algebra II students would be useful to determine a level of mastery, however, and that a teacher-made test designed to sample the more complex areas would serve to obtain discrimination for grading purposes.

e. 11th Grade Math - Trigonometry and Geometry.

This course is quite unique in that it was built to introduce analytic geometry concepts early in connection with the study of plane geometry. The course syllabus is unique to the teacher as are many of the materials. Because of the uniqueness of the course it was decided that the achievement test had to be unique also, that is, a teacher-made test.

f. 12 Grade Math - Advanced Mathematics.

This course is designed so that the student will be ready to enter the first course in calculus in college. In addition, many special topics have been developed as units of study for individuals. Such topics include symbolic logic, computer programming, theory of numbers, history of mathematics, mathematics in problem solving, mathematical models in the sciences, probability theory, sampling theory, etc. The uniqueness of the topics indicates that teacher-made test will be necessary as achievement measures. All students will take the CEEB tests as part of the school testing program.

g. 10th Grade Science - This is the BSCS course in

Biology, Blue series, Molecules and Man.

A complete set of achievement measurement devices is available with this course and will be the measurer to be used.

h. 11th Grade Science - This is a physics course in which the PSSC materials are the basic materials. Achievement in this course will be measured with the measures developed for the PSSC materials.

i. 12th Grade Science - This is a chemistry course which is based on the CHEM materials. The tests constructed for these materials will be used as the achievement measures. The students will also take the CEEB tests.

2. Capability for independent study will be measured by having the teacher complete a series of rating scales on each student. The scales will be built to obtain proficiency ratings on the various components of independent study. The teachers will be asked to develop the scales.

3. The problem solving ability of the students will be measured by a series of rating scales completed by the teachers and by the Watson-Glaser Critical Thinking. Appraisal which appears to yield reliable scores on six factors often associated with problem-solving.

4. Each teacher will specify the attitudes that might be included among the objectives of the class. Attitude or rating scales will be built to measure these attitudes.

Unobtrusive measures may be used to measure attitudes also. For example, appreciation of literature might be considered an attitudinal outcome. Library check-out data might be an indicator of the extent to which this attitude was developed.

5. Mr. Wilson is cognizant of the possibility that unique situations will come up during the year in which evidence of learning and attitude change or lack of same is apparent. Measures cannot be anticipated for such situations, but he does emphasize to the teachers that anecdotal reports of such situations should be written when the situation occurs.

The empirical contingencies among the observations cells parallel the logical contingencies among the intents cells. Essentially the empirical contingencies are that the selected students will interact with the materials, each other, the teachers, and other resources and will exhibit changes in behavior consistent with the objectives of the courses. Several after-the-fact examinations of contingencies among the cells are intended such as looking at the relationship between the sex of the student and the kinds of transactions and outcomes that are observed or studying the relationship between class participation time and the learning outcomes. The extent to which such after-the-fact studies are done will likely depend on the interests of the teachers, their hunches and anecdotal evidence, and the time available for such things.

Standards

Several ways of establishing standards are possible. In some situations absolute standards might be established, e.g. achieving a certain level of performance such as being able to solve all of the quadratic equations in this set of problems. Relative standards might also be used, e.g. determining whether an individual can solve more complex equations at the end of the course than at the beginning.

The standards for judging congruency between the intents and observations are described in the following paragraphs.

1. The standards for determining whether the selected students are meeting the definition of gifted will be established by reviewing the writings of authorities on gifted children.
2. Standards with respect to teacher qualifications will also be established by reviewing the writings of authorities.
3. The standards for facilities and materials will be based on writings of authorities. Another basis will be to compare the facilities and materials at the start of the year with those available at the end of the year.
4. Standards on the usage of materials will be established by comparing usage at the start of the year with usage at the end of the year.
5. Standards on interaction will also be established by determining a baseline at the start of the year and observing changes in relation to the baseline.
6. Standards on participation will be by the baseline procedure and also by determining participation patterns for students not in the program for comparison.
7. The basic procedure for establishing standards for the learning and attitudinal outcomes will be to employ a pre-post test research design. The final tests in the

courses or a sample of items from the final tests will be administered early in the course. These same tests will be administered at the end of the course and the extent to which there was change will be determined. This procedure will be used with all teacher-made tests, student writings, attitude scales, rating scales, and tests for special curricula.

In evaluating learning outcomes there is also a concern for determining to what extent students who are not in the gifted program have achieved the learnings thought unique to the gifted program and the extent to which the students in the gifted program may not have learned some things that were taught in the regular program. To obtain data of relevance to these questions students in the comparable regular program will take a test made up of a sample of items from the gifted program. Tests and students in the gifted program will take a test made up of a sample of items from the regular program tests. These tests will be administered as part of the final testing procedures. The performances of the groups will be compared on the tests.

The above procedure will not provide a very definitive answer to the question of whether the differences may be due to the programs because of the obvious ability differences. It will be useful, however, for determining to some extent whether the learnings in the gifted program are unique, and, more importantly, whether being in the gifted program decreases the likelihood of attaining certain learnings emphasized in the regular program.

Judgments

Ultimately the judgments about the program must be made by the administration and the school board. This does not absolve the evaluator of the responsibility for also making judgments. The program evaluator should accept responsibility for judging the extent to which the intents and observations are congruent and the extent to which the standards have been met or behavior has changed.

The judgments implied by this evaluation plan are discussed in the following paragraphs.

1. The judgment of whether the students in the program meet the criteria established in the intents can be made by comparing the descriptive information about the students with the stated criteria. Furthermore these descriptive data can be compared with criteria established by authorities to determine whether those students classified as gifted for the program are similar to authoritiative definitions of giftedness.
2. Judgments with respect to teacher qualification will be made on the basis of information about schooling in the teacher's personnel file and from ratings of teacher behavior made during classroom observation. These will be compared with the criteria established in the intents column with respect to teacher qualifications.
3. Facilities and materials will be inventoried at the start and the end of the year. A comparison of the

inventories will be a basis for judging the change in quality of the program as it is affected by facilities and materials. The inventories will also be compared with criteria for facilities and materials indicated in authoritative sources.

4. Records of student usage of materials in the classroom, audo-visual materials, library, etc. will be kept. The record of material usage, and changes in same, will provide data for making judgments in this category.

5. The judgments about classroom interaction will be based on data obtained in the classroom observations. Changes that take place during the year will be the basis for judgment. Appropriate statistical analyses will be made of these data. Appropriate statistics may be the correlated "t" test, sign test, or a correlation index,

6. Rates of participation in various school activities will be determined for students in the program and students not in the program. These rates will be compared with the Chi squared technique to determine whether being in the program is related to participation in school activities.

7. Judgments about learning and attitude changes will be based on comparisons between pre and post-test scores. These will be analyzed with the correlated "t" or sign test. Comparisons between the gifted class students and the regular students on the common tests will be done with the separate group "t" test or the Mann-Whitney "U".

> Performance of the senior students on the CEEB tests will be judged by determining the likelihood that the seniors in Great High School can be considered to be from the population on which the CEEB norms were established.

Several analyses will be made to investigate the kinds of contingencies that may exist among the antecedents, transactions, and outcomes. These analyses will generally be correlational in nature. The various analyses to be done will be dependent on interesting clues that appear in the data.

Time Schedule

It is important that a time schedule be established for an evaluation plan. The time schedule serves as a reminder of things to do as well as providing an indication of how well the schedule is being maintained.

The time schedule for this plan follows.

September 1, 1968 to September 30, 1968.

Work on instrumentation for the project. Develop and/or select the achievement measures, attitude scales, and rating scales for making observations. The achievement tests should be pretty well developed already by the teachers from their prior experiences in the courses. All commerical tests should be ordered in this time. Work with teachers on developing attitude and rating scales.

October 1, 1968 to October 31, 1968

Administer all pre-tests during the first two weeks of October. Work with counselors on this. Prepare self-report forms for students and teachers. Develop inventory of facilities and materials. Have counselors check the cumulative folders of students for missing data. Administer make-up tests if necessary. Have teachers fill out rating scales on problem solving and independent study November 1, 1968 to April 30, 1969.

Score tests, writing samples, rating scales, etc. Continue to gather report forms where appropriate. Do classroom observations. Be sensitive to unexpected outcomes. Gather data on participation in school activities. Prepare file on information about each teacher. Code data on sheets from cumulative files, tests, etc. Much of this is clerical activity that student assistants and office staff can help with.

May 1, 1969 to May 31, 1969.

Prepare and administer final tests, gather final writing samples, rating scales, attitude scales, etc. Coordinate with teachers and counselors.

June 1, 1969 to June 30, 1969.

Finish coding and analyze data.

July 1, 1969 to July 31, 1969.

Prepare evaluation report.

The above plan appears very ambitious. Much of the work can be done by the clerical staff, however, and much will need to be done by the teachers and counselors. This plan would likely take most of the time, money, and staff that was specified as being available in the situation description. Once a plan like this was started, however, it would take less time and other kinds of evaluation could be started.

GLOSSARY: STATISTICAL TERMINOLOGY

Level of confidence

A term indicating the statistician's degree of confidence that the obtained results reflect a true difference or relationship. (In layman's terms, this resembles a person's willingness to wager that something is so because he is that confident.)

The .05 and .01 levels of confidence often are used because they reflect the probabilities that the outcome occurred by chance. For example, if the probability that a correlation coefficient occurred by chance is .01, this means that only once in 100 times would you expect to observe such a strong relationship purely by chance and thus the "chance" probability is so small that one can consider the outcome to be true.

Ordinal scale

A measurement scale for arranging the measured items from most to least - usually by ranking the highest item as 1, the next highest as 2, etc.

For example, five people posting scores of 80, 68, 60, 92 40 on a test would be ranked as numbers 2, 3, 4, 1, 5. Note that once the scores have been ranked, differences between and among the scores becomes invisible.

Interval scale

A measurement scale that not only ranks items but indicates the precise difference and relative amount of difference between and among posted scores.

If the scores (cited in the previous definition) were assumed to be an interval scale, the distance between any two points on the scale can be assumed to equal the distance between any other two points. Thus, the interval scale indicates not only that 92 was higher than the score of 80 but that it was twelve points higher and that these twelve points indicate the same amount of difference as the twelve points between the score of 80 and the score of 68.

One-tail and two-tail tests

These terms refer to the tails on the ends of the typical bell-shaped curve or normal curve. Since only a lengthy and complex description adequately defines these terms, one can accept them "on faith" or refer to a book on statistics.

Σ Sigma

A Greek letter symbolizing "the sum of".

If we have five scores such that $X = 1$, $X_2 = 2$, $X_3 = 10$, $X_4 = 6$ and $X_5 = 9$ then the ΣX is equal to $1+2+10+6+9$ or $\Sigma X = 28$.

$\overline{X}$ or $\overline{Y}$

The symbol of the mean or arithmetic average of a set of numbers.

The $\overline{X}$ of the numbers cited in the definition of sigma is

$$\overline{X} = \frac{\Sigma X}{N} = \frac{28}{5} = 5.6$$

NOTE: The following sequence in working the problems seems most feasible and logical:

Chi-squared

Pearson r

Spearman rho

t-test and Mann-Whitney U

Correlated t-test

Analysis of Variance

SUGGESTEP REFERENCES:

Rationale Development

H. Berlak and A. Tom. Toward Rational Curriculum Decisions in the Social Studies. (Mimeo paper). Metropolitan St. Louis Social Studies Center. Washington University, St. Louis, 1967.

This paper includes a discussion of a rationale for a social studies curriculum which is generalizable to other curricula and programs. Pages 10 through 20 are especially helpful for this purpose.

H. S. Broudy, B. O. Smith, and J. R. Burnett. Democracy and Excellence in American Secondary Education. Chicago: Rand McNally and Company, 1964.

The material in the first five chapters in this book is relevant to the task of thinking through a rationale.

M. P. Hunt and L. E. Metcalf. Teaching High School Social Studies. New York: Harper and Row, 1955.

Chapter 10 contains a rationale for a social studies curriculum that is a good example of a rationale for a program. Chapter 4 is a good discussion of a methodological considerations that are relevant for evaluators.

- - - - - - - - - - - - - - - - - - - -

Evaluation Models

R. E. Stake. "The Countenance of Educational Evaluation." Teachers College Record, 68, April 1967, 7.

R. W. Tyler, R. M. Gagne, and M. Scriven. Perspectives of Curriculum Evaluation. Chicago: Rand McNally and Company, 1967.

R. W. Tyler. "The Functions of Measurement in Improving Instruction." In E. F. Lindquist (Ed). Educational Measurement. Washington: American Council on Education, 1950.

P. A. Taylor and T. O. Maguire. "A Theoretical Evaluation Model." The Manitoba Journal of Educational Research, 1, 1966, 12-17.

E. A. Suchman. Evaluative Research. New York: Russell Sage Foundation, 1967.

D. L. Stufflebeam. Evaluation as Enlightenment for Decision-Making. Evaluation Center, Ohio State University, 1968.

- - - - - - - - - - - - - - - - - - - -

Stating Objectives

R. F. Mager. Preparing Instructional Objectives. Palo Alto: Fearon Publishers, 1962.

This small paperback presents a very behavioral point of view on stating objectives.

D. R. Krathwohl. "The Taxonomy of Educational Objectives - Use of the Cognitive and Affective Domains." In C. M. Lindvall (Ed.), Defining Educational Objectives. Pittsburgh: University of Pittsburgh Press, 1964. pp. 19-36. Also in N. E. Gronlund (Ed.), Readings in Measurement and Evaluation. New York: The Macmillan Company, 1968. pp. 18-36.

This article describes the two well-known taxonomies and describes their use in evaluation. The taxonomies themselves might also be read. References for the taxonomy handbooks can be found in this article.

A. D. Woodruff. A Map of Classroom Conditions Required for Producing Behavioral Change in Students. (Mimeo Paper) Salt Lake City, Utah: University of Utah, 1968.

- - - - - - - - - - - - - - - - - - - -

Operational Definition

F. N. Kerlinger. Foundations of Behavioral Research. New York: Holt, Rinehart, and Winston, Inc., 1964.

Chapter three is especially relevant to this topic.

May Brodbeck. "Logic and Scientific Method in Research on Teaching."

In N. L. Gage (Ed.), Handbook of Research on Teaching. Chicago: Rand McNally and Company, 1963. pp. 44-93

This is rather difficult reading. Pages 55 to 67 are the most relevant.

- -

Resource Materials

O. K. Buros (Ed.) The Fifth Mental Measurements Yearbook. Highland Park, New Jersey: Gryphon Press, 1959.

Actually the sixth yearbook is available, but you will want to become familiar with the fourth, fifth, and sixth.

Research in Education - A monthly publication for the Educational Resources Information Center (ERIC).

Review of Educational Research - A quarterly publication of the American Educational Research Association (AERA) in which research is reviewed by educational areas on a periodic interval basis.

N. L. Gage (Ed.) Handbook of Research on Teaching. Chicago: Rand McNally and Company, 1963.

There is a wealth of information in this book.

The EPIE Forum - A monthly publication of the Educational Products Information Exchange which includes information about educational materials.

- -

Test Construction and Selection

Any of several Tests and Measurements textbooks would be useful in this area. Part Five of the Gronlund book of readings (cited above) would be helpful as well as No. 34 in Part Six. The two articles cited below are from the Handbook of Research on Teaching.

B. S. Bloom. "Testing Cognitive Ability and Achievement."

G. G. Stern. "Measuring Noncognitive Variables in Research on Teaching."

Other references in this area are:

H. Gulliksen. Theory of Mental Tests. New York: John Wiley and Sons, Inc., 1950.

M. E. Shaw and J. M. Wright. Scales for the Measurement of Attitudes. New York: McGraw-Hill, 1967.

Unobtrusive Measures and Observation

E. J. Webb, et. al. Unobtrusive Measures: Nonreactive Research in the Social Sciences. Chicago: Rand McNally and Company, 1966.

"Must" reading for the evaluator.

D. M. Medley and H. E. Mitzel. "Measuring Classroom Behavior by Systematic Observation." Chapter 6 in Gage.

H. H. Remmers. "Rating Methods in Research on Teaching." Chapter 7 in Gage.

Most Tests and Measurements Texts have material on observation, but little or nothing on unobtrusive measures.

Anita Simon and E. G. Boyer. Mirrors for Behavior: An Anthology of Classroom Observation Instruments. Philadelphia: Research for Better Schools, Inc., 1967.

Twenty-six classroom observation instruments are reviewed. It is a very complete listing and review.

Research Design

Part Four of Kerlinger (cited above) is an excellent source.

D. B. Van Dalen. Understanding Educational Research. (2nd edition). New York: McGraw-Hill, 1967.

The chapters on descriptive and experimental research, especially the latter, are quite good. Be sure to read the latest edition because the 1962 edition is only so-so.

D. T. Campbell and J. C. Stanley. Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally and Company, 1963.

This is Chapter 5 in the Gage (cited above) Handbook and is also available in reprint form. This article has already become a classic for persons working in educational research.

Statistical Techniques

There are many books in this area and you have a wide choice of level of sophistication and comprehensiveness. A very readable recent book is:

W. J. Popham. *Educational Statistics*. New York: Harper and Row, 1967.

The book by Kerlinger (cited above) is also an excellent source. Other popular statistics books are:

A. L. Edwards. *Statistical Methods for the Behavioral Sciences.*

New York: Holt, Rinehart and Winston, 1964. and

J. P. Guilford. *Fundamental Statistics in Psychology and Education.*

New York: McGraw-Hill, 1965.

Judgment and Scaling

The most commonly used reference in this area is:

W. S. Torgerson. *Theory and Methods of Scaling*. New York: John Wiley and Sons, Inc., 1958.

A more superficial but easier to read discussion is found in:

J. C. Nunnally, Jr. *Tests and Measurements*. New York: McGraw-Hill, 1959.

A discussion that is sort of half-way between the above two in difficulty of reading and length is in:

J. P. Guilford. *Psychometric Methods.* New York: McGraw-Hill, 1954.

Measuring Achievement

Most of the references under the topic of Test Construction above are relevant. Part Two of Gronlund (cited above) is relevant. No reference list on testing is complete without a listing of the following book which is now rather old but still very useful.

E. F. Lindquist (Ed.) *Educational Measurement*. Washington: American Council on Education, 1951.

- -

Measuring Higher Order Mental Processes

Parts Six and Eight of Gronlund (cited above) contain very relevant material, especially Nos. 29, 30, and 43.

The Gulliksen book (cited above) is also relevant.

The following reference may also be helpful.

E. P. Torrance. Torrance Tests of Creative Thinking. Princeton,

New Jersey: Personnel Press, 1966.

- -

Measuring Attitudes

The classic in this area is:

A. L. Edwards. Techniques of Attitude Scale Construction New York:

Appleton-Century Crofts, Inc., 1957.

The Shaw and Wright book (cited above) is another reference on measuring attitudes.

- -

Survey Research

Some handout material will be provided on this topic. The four books listed below are excellent references.

L. Festinger and D. Katz. Research Methods in the Behavioral Sciences.

New York: Holt, Rinehart and Winston, 1953.

S. L. Payne. The Art of Asking Questions. Princeton, New Jersey:

Princeton University Press, 1951.

A. N. Oppenheim. Questionnaire Design and Attitude Measurement. New

York: Basic Books, 1966.

C. Y. Glock (Ed.) Survey Research in the Social Sciences. New York:

Russell Sage Foundation, 1967.

Evaluation Institute
Urbana, Illinois
July 29-August 9, 1968

ATTITUDE INVENTORY

Directions: For each of the statements below, mark the letter which indicates your agreement or disagreement with the statement according to the following code:

SA = I strongly agree with the statement
A = I am in slight agreement with the statement
? = I am undecided
D = I am in slight disagreement with the statement
SD = I strongly disagree with the statement

1. The role of the evaluator should be that of a describer rather than a grader. SA A ? D SD

2. The evaluator should determine whether the goals of a program are worthwhile. SA A ? D SD

3. Most decisions made in the public schools today are based on hunches, hearsay, and individual beliefs. SA A ? D SD

4. Findings from laboratory studies seldom are applicable to regular classroom activities. SA A ? D SD

5. One of the first things an evaluator must do is obtain a list of behavioral objectives. SA A ? D SD

6. A major role of the evaluator is to make explicit the standards by which an educational program is judged. SA A ? D SD

7. Evaluators often pay too much attention to what they have been urged to look at, and too little attention to other facets. SA A ? D SD

8. The kind of data gathered in an evaluation should seldom be determined by what the groups are like that will receive the results of the evaluation. SA A ? D SD

9. As long as hoped for outcomes occur, it is not important that objectives be stated clearly. SA A ? D SD

10. The most important use of evaluation findings is to change the program. SA A ? D SD

11. The evaluator is the person best qualified to judge an educational practice. SA A ? D SD

12. It is possible to evaluate a program without knowing the goals of the individual teachers. SA A ? D SD

13. The personal characteristics of the evaluator are a major determinant of the evaluation. SA A ? D SD

14. It is not practical to draw conclusions in evaluating a program prior to the programs completion. SA A ? D SD

15. We can tell if an educational program is successful only by observing whether hoped for changes are occurring in the students. SA A ? D SD

16. In order to evaluate a program, equal resources should be devoted to what teaching is occurring as well as what learning is occurring. SA A ? D SD

17. It is up to the local educator to rule out the study of a variable because it is not one of his objectives. SA A ? D SD

18. No school can evaluate the impact of its program without knowledge of what other schools are doing. SA A ? D SD

19. The most appropriate instruments for evaluating educational programs are standardized tests. SA A ? D SD

20. Joyous distrust is a sign of health. Everything absolute belongs to pathology. SA A ? D SD

21. An evaluator has the right to decide what to evaluate. SA A ? D SD

22. The task of describing curricular objectives is the responsibility of the evaluator. SA A ? D SD

23. The evaluator should identify unanticipated outcomes of the program. SA A ? D SD

24. It is more important to compare local data with national norms than to compare it with local norms. SA A ? D SD

25. Absolute standards, e.g. the judgments of people, should not be applied to a program. SA A ? D SD

26. In selecting variables for evaluation, the evaluator must make a subjective decision. SA A ? D SD

27. The most important use of evaluation findings is to justify the program to other groups. SA A ? D SD

ACHIEVEMENT TEST

True False

____ ____ 1. Formative evaluation is aimed more at long-range generalizations about instruction than is summative evaluation.

____ ____ 2. One critical task for the evaluator is to combine the judgments of merit and shortcoming into a single consensus of program value

____ ____ 3. The educational program having goals that are clearly understood and stable is a better program than one having goals that are only implicit and changing.

____ ____ 4. Educational evaluation is essentially the same as educational research in terms in techniques used and in terms of questions to be answered.

____ ____ 5. The value of a model such as Bloom's Taxonomy or Stake's "countenance model" comes in using the categories to sort the different items or data after they have been collected.

____ ____ 6. It is wrong for the evaluator to try to get the educator to state his objectives in terms of student behaviors.

____ ____ 7. Item discriminability coefficients should exceed .50 if a 30-item test is to have the usually acceptable amount of reliability.

____ ____ 8. Questionnaire information is the least reliable and useful information evaluators collect.

____ ____ 9. Interviewing as a method of inquiry is universal in the social sciences.

____ ____ 10. The literature of anthropology serves as an example of the products obtained through interviewing informants.

11. The following may be obtained from empirical studies and used to appraise survey results:

____ ____ Estimates of variation between elements in the population and between various groupings of these elements.

____ ____ Cost factors and analyses, cost relationships.

____ ____ Data of established accuracy for use in testing and correcting ordinary procedures.

____ ____ 12. The size of samples, method of drawing it, and other features of the survey design will not be affected by the kind of analysis to be made of the results.

True False

___ ___ 13 The best starting point for any design is to be found in the aims that the survey is to fulfill.

___ ___ 14. The simplest and most satisfactory test of the accuracy of an estimate from a sample survey is not a direct comparison of the estimate with the true value of the variable being estimated.

___ ___ 15. A study of attrition rates will be of little help in identifying sources of bias.

___ ___ 16. Sampling variability is the amount of variability that arises through repeated application of a given sampling procedure.

___ ___ 17. We cannot ordinarily expect to get very substantial gains in accuracy in the estimation of a population proportion through the use of stratification.

___ ___ 18. Unobstrusive measures complete with formal experimental design to provide information to educational decision makers. That is, one must choose which has the higher likelihood of reducing error in collecting data.

___ ___ 19, Quality of teaching as a source of error can be controlled by Flander's interaction analysis for the four groups of sixth graders.

___ ___ 20. Archives might include examining science-teacher-of-the-year candidates careers.

___ ___ 21. Sampling conversation in the teachers' lounge is an example of simple observation.

Choice

___ 22. Which of the following is the outstanding obstacle to representing a program's objectives and priorities?

a. teachers are not oriented to student behaviors
b. goal statements and indicators are oversimplifications
c. no educationally meaningful unit of "investment" exists
d. goals cannot be represented by numbers, spatial areas, vextors pie-graph sectors, etc.

___ 23. Interviews typically yield subjective data--descriptions of the world of experience--for which of the following?

a. goals
b. perceptions
c. attitudes
d. all of the above
e. none of the above

___ 24. The Chi square technique is commonly used for

a. describing groups in terms of "fine measurement" data
b. testing hypotheses regarding "fine measurement" data
c. describing groups in terms of frequency counts
d testing hypotheses regarding frequency counts

___ 25. The Q Techniques and conventional factor analysis are both techniques for

a. analyzing profiles of students
b. clustering "like things" together
c. comparing large numbers of groups
d. evaluating instructional television

___ 26. The Q sort and the method of paired comparison are both methods which could be used for

a. assigning "priority values" to educational goals
b. measuring problem solving in students
c. designing a feedback loop for instruction
d. testing hypotheses

___ 27. ihe process of generalizing from sample data to population conditions while at the same time specifying the investigator's confidence in drawing correct conclusions is known as

a. summative evaluation
b. interaction analysis
c. statistical inference
d. taking a calculated risk

___ 28. Which of the following is usually not considered a major area of specialization for the educational research methodologist?

a. measurement, testing, instrumentation
b. research design, experimental controls
c. statistical description and inference
d. cost-benefit analysis, program evaluation

___ 29. "In a statistics-book table of Chi square values, the entries in the .05 column indicate the boundary point between the 95% most likely Chi square values to be obtained from sample data and the 5% least likely Chi square values to be obtained from sample data"

The previous statement is true only if the samples are randomly drawn from a population where

a. the "null hypothesis" is true
b. the "null hypothesis" is false
c. all variables are interrelated
d. no subgroups (samples) have any meaning

30. It is uaually not practical to use the method of paired comparisons unless the number of stimulus objects (things to be scaled) is

a. one
b. two
c. four to twelve
d. twenty to one hundred
e. at least two hundred

31. When using a rating scale, the observer

a. measures behavior by questioning
b. measures behavior by recording behavioral events
c. measures behavior by noting degrees of behavior
d. measures behavior by short time samples

Match each entry on the right with one of the three entries on the left by putting a letter in the blank.

int of View on Evaluation

Experimental research

B. Counseling-psychometric

Accreditation study

Emphasis

____ Self study, motivate self-correction

____ Visitation by group of peers

____ Control groups, control variables

____ Correlation among student talents

____ The differences among individual students

____ The traditional subject-matter disciplines

____ Prediction of later student success

____ Comparison of educational "treatments"

____ Norm groups, percentile scores

Writings

____ Campbell and Stanley in the Gage Handbook

____ Thurstone on Test Theory

____ "National Study's" Evaluative Criteria

____ Tyler on the Eight Year

INFORMATION QUIZ ITEM KEY

A.) True - False

1.	T
2.	F
3.	F
4.	F
5.	F
6.	F
7.	F
8.	F
9.	T
10.	T
11.	T - T - T
12.	F
13.	T
14.	F
15.	F
16.	T
17.	T
18.	F
19.	F
20.	T
21.	T

B.) MULTIPLE CHOICE

22.	c
23.	d
24.	d
25.	b
26.	a
27.	c
28.	d
29.	a
30.	c
31.	c
32.	c
	c
	a
	b
	b
	c
	b
	a
	b
	a
	b
	c
	b

Evaluation Institute
Urbana, Illinois
July 29 - August 9, 1968

Participant Interview Schedule*
Part I
(1st half)

Date Administered ____________________

Name of interviewer ____________________

Introduction

1. Identify yourself if it is necessary.

2. Purpose: The reason that I have asked to talk with you has to do with your general reaction to the institute so far. The other interviewers and I are gathering this type of information so that the staff can better organize next week's activities as well as evaluate the overall training experience. While some things cannot be changed in this institute, I'm sure that all of your comments will be useful for designing future training programs of this type.

3. Anonymity: Your name will not be placed on this interview form.

4. Begin: Do you have any questions before we begin?

Institute Design

1. What has been the most beneficial to you in the institute so far?

____

____

____

Could you indicate why this is so? ____________________

____

* *EXPLORE EACH ITEM AS FULLY AS POSSIBLE BY ASKING SUCH QUESTIONS AS, "IS THERE ANYTHING ELSE?", "ANY OTHER IDEAS YOU WANT TO MENTION?", ETC.*

2. Is there anything you would like to see happen more often? Yes ___ No ___

IF YES AND NO ELABORATION - What would that be? ______________________

____

3. In terms of the amount of time spent for activities such as lectures, structured groups, work sessions, video viewing, would you like to see the proportion of time alloted for these activities changed in any way? Yes ___ No ___

IF YES - In what way ______________________________

____

Lectures

4. What is your general impression of the lectures so far?

COMMENTS

Positive	Negative
____________	____________
____________	____________
____________	____________

5. Do the lectures seem relevant to the other institute activities in which you are involved? Yes ___ No ___

IF YES - In what ways do the lectures seem relevant. ______________

____

____

IF NO - What could make the lectures more relevant. ______________

____

____

6. Are there any aspects of the lectures which make them confusing or difficult to understand? Yes ___ No ___

IF YES - What aspects ______________________________

____

What could members of the staff do to improve this situation? ______________________________

____

IF NO - Are there any other comments you would like to make about the lectures? ______________________________

Video Tapes

7. What is your general impression of the video-tapes you have seen?

COMMENTS

Positive	Negative
______________	______________
______________	______________
______________	______________

8. What would be your major criticism of the video-tapes?

____

____

____

____

CONSIDER CATEGORIES BELOW FOR CLASSIFYING STATEMENTS

AWARENESS

PHYSICAL QUALITY

CONTENT QUALITY

UNDERSTANDABILITY

PRACTICALITY

Materials

9. Are the materials, such as the books, papers, evaluation plans, and (statistical exercises) of any help to you? Yes ___ No ___

IF YES - Which of these materials seem to be the most helpful to you?

____

____

How were they helpful ____________________________

____

10. What materials seem to be of little or no help to you? ____________

____

IF MATERIALS ARE INDICATED - Why does this seem to be the case? ______

____

11. What kinds of materials should have been provided which were not made available?

____

____

RECORD "WHY" IF SPECIFIED ______________________

____

Transferability

12. You mentioned that __________, __________ and __________ were helpful to you (or you liked them). Of these and other activities that you mentioned, do you believe they are presented in such a way that they will be helpful to you in your own situation back home? Yes ___ No ___

IF YES - Which ones will be helpful? ____________________

____

Why ____

13. Are there any (other) things occurring in this institute that you will find useful back home? Yes ___ No ___

IF YES - What ______________________________

14. Are there some parts of the institute that you won't be able to use in your own situation back home? Yes ___ No ___

IF YES - Which parts - *IF NOT ELABORATED* ______________________________

____

RECORD "WHY" IF SPECIFIED ______________________________

____

Summary

15. Is there anything else the institute staff should know, so they might impro this experience for you? Yes ___ No ___

IF YES - What would that be? ______________________________

____

16. If you were going to conduct an evaluation institute similar to this one, what changes might you make (other than what you have already indicated)?

____

____

GENERALLY REVIEW ALL OF THE RESPONSES CHECKING FOR CORRECTNESS OF INFORMATION AND ANY FORGOTTEN IMPRESSIONS.

"As I mentioned at the beginning of our talk, this information will be very helpful to the staff in making decisions about next week's activities as well as the designing of future training programs. Thank you for your time."

Evaluation Institute
Urbana, Illinois
July 29-August 9, 1968

Participant Interview Schedule*
Pa~t II
(2nd half)

Date Administered ____________________________

Name of interviewer ____________________________

Introduction

1. Identify yourself if it is necessary.

2. Purpose: The reason that I have asked to talk with you has to do with your general reaction to the institute so far. The other interviewers and I are gathering this type of information so that the staff can better evaluate the overall training experience. While some things cannot be changed in this institute, I'm sure that all of your comments will be useful for designing future training programs of this type.

3. Anonymity: Your name will not be placed on this interview form.

4. Begin: Do you have any questions before we begin?

Institute Design

1. What has been the most beneficial to you in the institute?

____

____

____

Could you indicate why this is so? ____________________________

____

**EXPLORE EACH ITEM AS FULLY AS POSSIBLE BY ASKING SUCH QUESTIONS AS, "IS THERE ANYTHING ELSE?", "ANY OTHER IDEAS YOU WANT TO MENTION?", ETC.*

2. Is there anything you would like to have seen happen more often? Yes ___ No ___

IF YES AND NO ELABORATION - What would that be? ______________________

____

Is there any particular reason why you would like to have seen this happen more often? Yes ___ No ___

____

3. In terms of the amount of time spent for activities such as lectures, structured groups, work sessions, video viewing, would you have liked to see the proportion of time alotted to the activities changed in any way? Yes ___ No ___

IF YES - In what way ______________________________

____

<u>Lectures</u>

4. What was your general impression of the lectures?

COMMENTS

Positive	Negative
________________	________________
________________	________________
________________	________________

5. Did the lectures seem relevant to the other institute activities in which you were involved? Yes ___ No ___

IF YES - Did the lectures seem relevant? ____________________

____

____

IF NO - What would have made the lectures more relevant? ____________

____

____

6. Were there any aspects of the lectures which made them confusing or difficult to understand? Yes ___ No ___

IF YES - What aspects ______________________________

What could members of the staff have done to improve this situation?

____

IF NO - Are there any other comments you would like to make about the lectures?

____

Video Tapes

7. What was your general impression of the video-tapes you have seen?

COMMENTS

Positive	Negative
______________	______________
______________	______________
______________	______________

8. What would be your major criticism of the video tapes?

____

____

____

____

CONSIDER BELOW CATEGORIES FOR CLASSIFYING STATEMENTS.

AWARENESS

PHYSICAL QUALITY

CONTENT QUALITY

UNDERSTANDABILITY

PRACTICALITY

Materials

9. Were the materials such as books,, papers, evaluation plans, and (statistical exercises) of any help to you? Yes ___ No ___

   IF YES Which of these materials seem to have been the most helpful to you?

   ___

   ___

   How were they helpful? ___

   ___

10. What materials seemed to be of little or no help to you? ___

    ___

    IF MATERIALS INDICATED - Why does this seem to be the case? ___

    ___

11. What kinds of materials should have been provided which were not made available?

    ___

    RECORD "WHY" IF SPECIFIED ___

    ___

Transferability

12. You mentioned that ___, ___ and ___ were helpful to you, or you liked them. Of these and others that you mentioned, do you believe they were presented in such a way that they will be helpful to you in your own situation back home? Yes ___ No ___

    IF YES - Which ones will be helpful ___

    In what way? ___

13. Were there any (other) activities occurring in this institute that you will find useful back home? Yes ___ No ___

IF YES - What ______________________________

____

14. Are there some parts of the institute that you won't be able to use in your own situation back home? Yes ___ No ___

IF YES - Which parts - *IF NOT ELABORATED* ______________________________

____

RECORD "WHY" IF SPECIFIED ______________________________

____

<u>Summary</u>

15. Is there anything else the institute staff should have known, so they might have improved this experience for you? Yes ___ No ___

IF YES - What ______________________________

____

16. If you were going to conduct an evaluation institute similar to this one, what changes might you make (other than what you have already indicated)?

____

____

(GENERALLY REVIEW ALL OF THE RESPONSES CHECKING FOR CORRECTNESS OF INFORMATION AND ANY FORGOTTEN IMPRESSIONS)

"As I mentioned at the beginning of our talk, this information will be very helpful to the staff in designing future training programs. Thank you for your time."

OBSERVATION SCHEDULE

Speaker ______________________ Date __________ Lecture ____ Tape ____

Scheduled Starting time ____ Actual Start ____ Difference ____

Scheduled Finishing Time ____ Actual Finish ____ Difference

Staff in Attendance: House ____ Stake ____ Denny ____ Hastings ____ Sjogren____

Number of Participants in Attendance: ____

A. Observer's rating of the speaker's communication with the participants:

1. speaker encourages questions ____ discourages questions ____

   comment ______________________________________

2. total number of questions asked ______________________

3. speaker sensitive to audience reaction ____ insensitive ____

   comment ______________________________________

B. Rating of participants' questions and reactions:

4. questions relevant ____ questions not relevant ____

   comment ______________________________________

5. questions insightful ____ questions not insightful ____

   comment ______________________________________

6. participants comfortable ____ participants not comfortable ____

   comment ______________________________________

7. participants bored ____ interested ____ enthusiastic ____

   comment ______________________________________

C. Participants' attitudes toward instructional techniques:

8. materials distributed ____ materials not distributed____

9. materials relevant ____ materials not relevant ____

   comment ______________________________________

10. audio-visual equipment used ____ not used ____

11. equipment produced an effective presentation ____ not effective ____

   comment ______________________________________

D. Participants' attitude towards presentation:

12. lecture __ tape __...well prepared __; adequate __; not well prepared

comment ____

13. lecture __ tape __...presentation dull __; adequate __; interesting __

comment ____

14. lecture __ tape __...presentation disjointed __; coherent __

comment ____

15. lecture __ tape __...level of material difficult __; moderate__; easy __

comment ____

16. lecture __ tape __...following discussion shallow __; moderate __; deep __

comment ____

17. lecture __ tape __...relevant to stated objectives __; irrelevant __

comment ____

18. lecture __ tape __...relevant to participants' needs __; irrelevant __

comment ____

GENERAL COMMENTS:

____

____

____

____

OBSERVATION SCHEDULE

Date ______________________ Time Session ______________________

Group Work Session ______________________ Individual Work Session ______________________

	generally yes	inconclusive	generally no
1. Did the participants feel that their time could have been better spent in another activity? ______________________ ______________________	________	________	________
2. Did the participants feel that they were sufficiently involved with the expected task? ______________________ ______________________	________	________	________
3. Did the participants attempt to accomplish their assigned task or to work on their evaluation plans? ______________________ ______________________	________	________	________
4. Did the participants believe they actually accomplished something during this time spot? ______________________ ______________________	________	________	________
5. Did the participants feel they needed more structure for this time? ______________________ ______________________	________	________	________
6. Did the participants feel they needed more guidance or help from the staff for this time spot? ______________________ ______________________	________	________	________

PARTICIPANT OPINIONAIRE

Evaluation Workshop
University of Illinois
Urbana, Illinois

Now that this Workshop is drawing to a close, we are certain that you have some reactions as to what parts have been most valuable to you and what parts might have been different. This form is designed to make it easy for you to pass these reactions along to the workshop planners. It is important that every participant complete and return the opinionaire so that the reactions of the total group will be reflected.

The questions are designed to make it easier for you to express your reactions. If they do not provide sufficient opportunity, please write your comments in your own words. You do not need to indicate your name.

1. Did you have enough information about this workshop before you arrived?

   Yes 1 ( )
   No 2 ( )

2. (If no) What else would you like to have known about?

   ____

   ____

3. There are many parts of a Workshop experience that can either contribute to your satisfaction or detract from it. For each of the following, would you let us know how satisfied you've been?

   a. meals
   Really outstanding 1 ( )
   Very satisfactory 2 ( )
   Just acceptable 3 ( )
   Need improvement 4 ( )

   b. hotel rooms
   Really outstanding 1 ( )
   Very satisfactory 2 ( )
   Just acceptable 3 ( )
   Need improvement 4 ( )

   c. meeting rooms
   Really outstanding 1 ( )
   Very satisfactory 2 ( )
   Just acceptable 3 ( )
   Need improvement 4 ( )

   d. other facilities or services
   Really outstanding 1 ( )
   Very satisfactory 2 ( )
   Just acceptable 3 ( )
   Need improvement 4 ( )

e. facilities for working	
Really outstanding	1 ( )
Very satisfactory	2 ( )
Just acceptable	3 ( )
Need improvement	4 ( )
f. opportunity for discussion	
Really outstanding	1 ( )
Very satisfactory	2 ( )
Just acceptable	3 ( )
Need improvement	4 ( )
g. presentations in general	
Really outstanding	1 ( )
Very satisfactory	2 ( )
Just acceptable	3 ( )
Need improvement	4 ( )

(If you have checked "need improvement" for any of the foregoing, please note below any suggestions you may have.)

4. Would you describe the one or two most valuable ideas that you received from attending the Workshop?

5. As far as you're concerned, what would have most improved the Workshop?

6. Which one of these phrases best states how related this workshop was to your interests and background?

a.	It was over my head	1 ( )
b.	I understood almost everything but the conference missed my main interests	2 ( )
c.	It dealt with my main interests in an understandable and interesting way	3 ( )
d.	It was too basic, few if any new ideas	4 ( )

7. Which one of the following statements comes closest to stating your general reaction to the total Workshop?

The most valuable educational experience of my life 1 ( )
An outstanding program, I received much from it 2 ( )
Many parts were valuable, others not very 3 ( )
I gained something from attending but less than I expected 4 ( )
It was almost a complete waste of time 5 ( )

________________________(other) 6 ( )

8. After this Workshop is over, is there anything related to the Workshop topics that you would like to know more about or to study further?

Yes 1 ( )
No 2 ( )

9. (If yes) What specifically would you like to study?

10. (If yes) How would you like to do so?

Study on my own 1 ( )
Attend a class that meets weekly 2 ( )
Attend another Workshop 3 ( )
Take a course by correspondence 4 ( )
In a local study group 5 ( )

________________________(other) 6 ( )

If you have further comments on the Workshop, please write them in your own words.

Evaluation Institute
Urbana, Illinois
July 29-August 9, 1968

Participant Critique Form

Directions: Please respond with a word, a phrase, or one or more sentences to as many of the following questions as you can. Your frank and honest evaluation can only benefit everyone concerned. Do not identify yourself by name unless you prefer to do so.

Environment and Facilities

1. a. To what extent did the relative unavailability of books and journals interfere with your attempts to master the content of this session?

b. To what extent did reproduced materials given to you by the staff improve matters?

2. a. Did you feel that you lacked a "place to work," either alone or in small groups?

b. If you had a room at the Union, was it satisfactory?

c. If you did not have a room at the Union, did your staying elsewhere make the Institute any more or less worthwhile to you?

3. a. Which features of the meeting rooms were inadequate or not conducive to learning?

b. Which features were especially facilitative in the same regard?

Scheduling and Organization

4. a. Was two weeks too long a period to leave your work at home for the purpose of attending this session?

b. Was two weeks too short a period in which to learn much of the content of this session?

5. a. Were you allowed enough time in which to pursue activities of your own choosing?

b. Would you have preferred not to meet in the evening after dinner?

c. Would fewer meetings per day have been preferable?

d. Would you have preferred more meetings per day than there actually were?

6. a. Were the individual lectures too long to sit and listen or take notes?

b. Were the lectures scheduled in an appropriate sequence?

7. ... Did you have sufficient opportunities to interact with other participants?

8. a. Were the instructors too inaccessible or unapproachable so that you did not get the individual attention that you desired?

b. Would it have been advisable to have had a few highly-trained graduate student assistants present from whom you could have obtained help on individual problems?

c. Were the staff members helpful in any way?

9. a. Did the attempts to evaluate your progress and reactions during the session (and at this moment) interfere with your work here?

b. Do you begrudge the time you have spent here answering such questions as these on this critique?

10. In general, was the Institute well organized?

Content and Presentation

11. a. Did the content of the lectures and readings presuppose far more previous training (in math and statistics) than you had?

b. Should less training in these areas or more have been presupposed?

12. To what extent was the content of the lectures and readings relevant to what you hoped to accomplish during the session?

13. Do not be reluctant to single out a staff member for praise or censure.

a. Were the lecturers stimulating and interesting?

b. Were the lecturers competent to speak on the subject assigned them?

c. Were the lecturers well prepared?

14. Were you disappointed in any way with the group of participants?

Answer each of the following only by checking the more appropriate blank:

15. If you had it to do over again would you apply for this Institute which you have just completed? Yes ______ No ______

16. If an Institute such as this is held again would you recommend to others like you that they attend? Yes ______ No ______

17. Do you **anticipate** maintaining some sort of contact with at least one member of of the Institute staff? Yes ______ No ______

18. Do you feel that your understanding of evaluation has been considerably enriched in these two weeks? Yes ______ No ______

19. Is it likely that you will consult in evaluation with someone else attending this institute? Yes ______ No ______

20. Would you say that because of this Institute you are more able to state a given evaluation problem in operational form so that it is, if it can be, amenable to solution? Yes ______ No ______

21. Do you feel that the staff should feel that it has accomplished its objectives during this two week Institute? Yes ______ No ______

Use the remaining space, if you wish, to give us your ideas on what was wrong with this session, or what was particularly commendable in it, or how it could have been done better. Try particularly to mention items which were not dealt with in the questions on the preceding pages.