INFORMATION TO USERS

This reproduction was made from a copy of a document sent to us for microfilming. While the most advanced technology has been used to photograph and reproduce this document, the quality of the reproduction is heavily dependent upon the quality of the material submitted.

The following explanation of techniques is provided to help clarify markings or notations which may appear on this reproduction.

1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting through an image and duplicating adjacent pages to assure complete continuity.

2. When an image on the film is obliterated with a round black mark, it is an indication of either blurred copy because of movement during exposure, duplicate copy, or copyrighted materials that should not have been filmed. For blurred pages, a good image of the page can be found in the adjacent frame. If copyrighted materials were deleted, a target note will appear listing the pages in the adjacent frame.

3. When a map, drawing or chart, etc., is part of the material being photographed, a definite method of "sectioning" the material has been followed. It is customary to begin filming at the upper left hand corner of a large sheet and to continue from left to right in equal sections with small overlaps. If necessary, sectioning is continued again—beginning below the first row and continuing on until complete.

4. For illustrations that cannot be satisfactorily reproduced by xerographic means, photographic prints can be purchased at additional cost and inserted into your xerographic copy. These prints are available upon request from the Dissertations Customer Services Department.

5. Some pages in any document may have indistinct print. In all cases the best available copy has been filmed.

University Microfilms International
300 N. Zeeb Road
Ann Arbor, MI 48106

Ames, Judith Silverman

AN APPLICATION OF DECISION ANALYSIS TO VARIABLE SELECTION IN PROGRAM EVALUATION

City University of New York PH.D. 1982

AN APPLICATION OF DECISION ANALYSIS TO VARIABLE SELECTION IN PROGRAM EVALUATION

by

JUDITH SILVERMAN AMES

A dissertation submitted to the Graduate Faculty in Educational Psychology in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York.

1982

This manuscript has been read and accepted for the Graduate Faculty in Educational Psychology in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

6/9/82
date

Chairman of Examining Committee

6/9/82
date

Executive Officer

Dr. Alan Gross

Dr. David Rindskopf

Dr. Max Weiner
Supervisory Committee

The City University of New York

Dedicated

to my parents

Emanuel and Florence E. Silverman

for their devotion, generosity and bountiful love

Table of Contents

List of Tables

List of Figures

Page

INTRODUCTION

A methodological step that is basic to research activities in all fields of science is the selection of the variables to be studied. The astronomer may decide to study the age of asteroids; the educational psychologist may examine the effect of anxiety on the learning of children at varying stages of life; the botanist may study plant tissue change under several climatic conditions. A phenomenon is observed; its variation is followed; numbers may be attached to the differences noted. If variables are not selected, there is no study; there is no focus of investigation; no data can be collected. Variable selection is the sine qua non of all research.

A variable typically is selected because it is of interest to the scientist: it is selected because the scientist has experience in the specific area or because there is a particular need for the study of this variable.

In addition to intellectual curiosity or a practical need, previous research findings may inexorably lead to the selection of particular variables. Some scientists may have instinctive hunches that cause them to select certain variables; some may select variables in a process of elimination of other possible variables. The reasons for selection of variables are, thus, basically: scientific interest, practical need, prior evidence.

In addition to the factors already mentioned, there are statistical procedures that have been used in the selection of variables. Multiple regression models, including stepwise and backward and forward regression methods, have been used to determine which independent variable among the several possible variables contributes most to the variation in the dependent variables. This procedure is helpful when there are many variables of interest but an inability to study all of them for reasons of time, money or evidentiary complications. When multiple regression methods are used, one is able to narrow the focus of the study to those variables that have the greatest effect on the variance of the dependent variables.

Thus, the "justification" for selection of variables can be either statistical (multiple regression models) or it can be what we shall call intellectual (interest, evidence, need). Since these two bases for choosing variables are common to all the sciences, they are, therefore, the bases of selection in educational psychology.

The aim of this dissertation is to suggest a third possible procedure to select variables in scientific studies and to determine the quality and usefulness of this new basis of selection. Specifically, the use of a model based on decision analysis for variable selection in a particular area of educational psychology, program evaluation, will be examined. Although the present methods of variable

selection may be adequate in some fields of educational psychology, in the area of program evaluation these methods are extremely deficient.

PROGRAM EVALUATION

To examine the present methods of variable selection in program evaluation, it is necessary, first, to describe what program evaluation is. Program evaluation is a branch of educational evaluation, which itself is a field that has not been well defined. In 1954, authors of a text on educational measurement considered evaluation in education to be a new concept: "The term has been used to include appraisal of the school program, curriculum, and instructional materials, appraisal of the teacher, and appraisal of the school child. Its methods run the gamut from observation and testing to elaborate research techniques." (Greene et al., 1954, p. 218) Today, almost thirty years later, our conceptualization of the term educational evaluation is no more refined or delineated. It has been defined in broad terms as "an operation in which the quality of an educational enterprise is judged." (Popham, 1972, p. 1) Educational evaluation still is considered to include a multitude of objectives and goals; its methods still are numerous and varied.

Educational evaluation may refer to as narrow a process as assessing the performance of a single child or as vast an operation as a nationwide assessment involving school districts throughout the country. The well-known work by Bloom et al. (1971) treats evaluation in terms of a single classroom; the book concentrates on the development of instructional objectives which are based on a Table of Specifi-

cation of the requisite skills and the construction of examinations to assess the achievement of these skills by individual students. In contrast, the so-called Coleman report evaluated the equality or inequality of educational opportunity with a focus not based in the classroom but, rather, the entire country: background and school-related variables were examined nationally.

As a branch of educational evaluation, program evaluations are considered to be evaluations that deal with the assessment of how well a particular educational methodology or program has achieved its stated goals. A summative program evaluation is an evaluation performed at the end of a program and when "one wishes to reach a decision as to the adoption of a particular treatment or, perhaps, regarding the continued use of that treatment." (Popham, 1972, p.4) Formative program evaluation judges the worth of educational processes so as to improve them as they are developing.

The need for program evaluations has increased dramatically over the years. These evaluations are not limited to classrooms or schools but to states and nations, and have been required because of the growing expenditures of money for support of educational systems whether for supplementary or innovative needs. With the increased outlay of funds and the concomitant requests for funds, there must be ways to assess whether or not the monies have been or will be spent wisely; this involves evaluating the efficacy of specific programs or achievement by students, etc. In other words,

program evaluation in most cases is not simply appraisal for appraisal's sake; it is an appraisal with a purpose. The purpose common to most evaluations is that of providing material which will form the basis of making decisions. "An evaluation is a process by which relevant data are collected and transformed into information for decision making." (Cooley and Lohnes, 1976, p. 2)

The kinds of decisions based on program evaluations proliferate. Decisions may range from altering teacher training to changing a method of teaching reading to funding science programs for women. The decisions may be required by a principal or a school board or a federal agency. They are decisions that may affect a few or thousands; they are decisions that may be costly in terms of money or time.

Theoretical writings in program evaluation are scanty and one must rely on the theories in texts for the field of educational evaluation. However, even in the more general area of educational evaluation there is not a solid body of techniques and theory, perhaps due to the diversity of decisions and of foci in connection with evaluations. In fact, there are books ostensibly devoted to evaluation in that their titles so indicate, (Nunnally, 1972; Thorndike and Hagen, 1961) but the emphasis in many of these books is on measurement topics and evaluation is dealt with almost peripherally. Thorndike and Hagen (1961, p. 27) distinguish between measurement topics and evaluation in a footnote, noting that evaluation "is closely related to

measurement. It is in some respects more inclusive, including informal and intuitive judgments of pupil progress. It also includes more definitely the aspect of valuing -- of saying what is desirable and good." Others make the distinction more clearly: "Measurement refers to quantifying an entity according to a standard scale, whereas evaluation refers to making a judgment about the quality of an entity in reference to a purpose. Measurement is the process of providing data; evaluation is the process of holding data up to prescribed criteria." (Chase, 1974, p. 266)

In addition to the lack of general theory about program evaluation, there is an absence of specific information regarding how actually to perform a program evaluation. In fact, the design of each evaluation is ascertained by the evaluators in whatever way they feel is best; there are no procedures that they can follow, no guidelines, except those that they themselves create.

This dissertation will try to bring order to one aspect of the process of program evaluation, the selection of the variables to be studied. In many areas of evaluation, variable selection among them, the lack of sturdy techniques has opened the entire field to criticism. With the ever-increasing need for program evaluations and the vagueness in its theoretical bases, a troublesome situation exists. Much work needs to be done to give a clearer picture of the procedures entailed in program evaluation.

VARIABLE SELECTION IN PROGRAM EVALUATION

As stated in the Introduction, all scientific studies require that the investigator select the variables to be studied; so, too, in evaluation studies. In program evaluations, one must select measures that will yield data to aid in the decision about the value of the program. The crucial problem in program evaluation design is the selection of these variables, i.e. the selection of which measures among many to use in the evaluation. "Looked at purely from the standpoint of measurement theory, there is no end to the number and kinds of tests that should be used. Potentially, any test administered at any point in time can provide valuable information to help in making educational decisions." (Nunnally, 1972, p. 508) However, as Nunnally goes on to say, in practice one must set limits; one cannot use all available tests because of restrictions in time and money as well as in interpretation.

The variables in program evaluation should be selected based on whether they are the most meaningful and will provide the most relevant information. In order to make statements about a program the information gathered must have some limits; the "very conscious and serious delimiting" of evaluation data is necessary, for any study "gains meaning as it defines its content and sets up delimited areas of study." (Lindvall & Cox, 1970, p. 56)

Most evaluators advocate the use of more than one

criterion measure in the evaluation of program effectiveness. (Wolf, 1974, p. 209) The use of standardized tests is recommended along with teacher-made measures, questionnaires, self-rating scales, observational data, records, etc. "By gathering more than one indication of the learner's status, the educational decision maker is advantaged in that he can weigh the merits of several indices which...will in combination yield a better picture than any single criterion." (Popham, 1972, p. 41) Some evaluators suggest the use of informal techniques of assessment, (Thorndike and Hagen, 1961, p. 451) but Popham (1972, p. 3) cautions against the use of process variables such as time devoted to specific activities; instead, the central concern should be with "learner performance data."

Whatever tests are selected for program evaluations, whether they are standardized measures or informal interview data, they must be relevant to the task of the evaluator, i.e. to provide data that aid the actions of those who request the evaluation. "Tests are useful only if they help in making decisions." (Nunnally, 1972, p. 4)

How, then, does one determine which measures (variables) are most useful for the decision maker and, therefore, are to be included in a program evaluation? Whatever measures are selected will have a biasing effect on the evaluation insofar as the kinds of facts which are looked at and the types of conclusions which these facts can generate. (Cooley and Lohnes, 1976, p. 16)

The advice given by theoreticians regarding the selection of measures for an evaluation varies from suggesting that the testing program be planned well in advance and that tests not be changed from year to year so as not to be disruptive (Nunnally, 1972, p. 508) to the recognition that it is not feasible to have a fixed, set evaluation that will last unchanged throughout an evaluation study: "The initial design should provide a framework within which greater detail will be added, decisions will be made about specific implementation procedures, and some parts of the design will necessarily be altered as more and more is learned about the program being evaluated." (Coulson, 1978, p. 10) However, Coulson's and most others' analyses of how an evaluation design should function do not enumerate the actual mechanisms or procedures which could carry out their suggestions in a systematic way.

There are some procedures for selection of measures but they do not allow for alterations without loss of information as the evaluation study progresses. The usual procedure for selecting measures is to list the skills which the school intended to develop and select or create instrments that have items that best assess these skills. (Chase, 1974; Popham, 1972; Bloom et al., 1971)

Scriven (1974, p. 50) points out that evaluation methodology does not provide guidance for choosing pertinent variables from among the numerous available ones. He says that evaluators select variables by "relying on the huge

pool of background knowledge...about what treatments tend to have what effects."

When measures are selected, some thought is given to their reliability and validity; in program evaluation, the decision to use a measure is based in part on these considerations, but should not be made exclusively on the basis of these two characteristics: "The value of a test depends on many qualities in addition to its accuracy. Especially to be considered are the relevance of the measurement to the particular decision being made and the loss resulting from an erroneous decision." (Cronbach and Gleser, 1965, p. 1) It is this aspect of variable selection that is absent from the existing recommendations in program evaluation. As noted, the methodology offers few guidelines for variable selection, except for emphasizing the evaluators' familiarity with and prior knowledge of the relationship of measures to outcomes.

Thus, it can be stated that most theoreticians in evaluation do not address the problem of which measures should be utilized in program evaluation. Instead, it is advocated that evaluators rely on experience which will dictate what measures are to be used, and that the variables should be those that would reflect the influence of the program's methods. Although those who are concerned with methodology do not mention this, often in addition to prior evidence, the evaluator may select variables because they are the ones required by the funding agency or the school

district.

There is obviously a need for a new method of variable selection; a method using decision analysis is proposed herein. The prior discussion of how variables presently are selected in program evaluation indicates that there are no directions for the practitioner who wishes to make variable selection on a basis other than subjectivity or past experience. In the field of evaluation, even more than in other behavioral sciences or in research, in general, the absence of a systematic method in variable selection is evident; the evaluator is free to select any variable of interest, but this freedom may lead to the selection of variables that are of little utility in assessing the worth of a program. Unlike much other scientific study, "evaluative research is not designed to support or undermine particular theoretical positions, but rather to contribute to recommendations for action." (Cooley & Lohnes, 1976, p. 2) It is this action-oriented nature of program evaluation that requires better variable selection procedures, so that the usefulness of the evaluation will be enhanced. Using decision analysis techniques, there are ways to provide answers to the possible usefulness of certain variables to the evaluation process.

In a survey of evaluation theorists and practitioners, almost all of those questioned emphasized that evaluation plans should be specified but that they should also be flexible; they noted the danger of focusing on certain variables

which could, in the end, prove to be unimportant: One must be "willing to add additional dimensions...as they become relevant" and have "the courage to acknowledge that data gathered with great care and at great effort is uninformative." (Stake, 1976, pp. 38-39) The selection model proposed herein has the property of being able to aid in the formulation of flexible evaluation strategies; such a model could give advance warning that certain contemplated variables or measures will not be informative.

Evaluators must limit the number of variables that they include in their evaluation plans; some of the reasons for this are time constraints, cost and the possibility of recalcitrance among those being tested. According to Evans (1974, p. 11), in the evaluation of federally supported programs, "the people and institutions who are the objects of these studies have come under an increasing data collection burden." Cronbach and Gleser (1965, pp. 69-70) advocate sequential plans of testing in personnel selection "because it costs something to gather information. Obviously, if observations cost nothing, it would always be better to administer the full series of tests, whatever its length, to every person." Obviously, Cronbach & Gleser did not also consider that perhaps the use of a full series of tests might be contraindicated because of test-weariness and time limitations. The model proposed in this dissertation allows evaluators to take note of limitations of time, money and testee exhaustion.

There are many evaluation models, but none are blueprints for how to design an evaluation. Most evaluators are aware of the range of options available to them in designing an evaluation; what is lacking is a method to aid in choosing among the options. As Airasian notes, the evaluation models "suggest what decisions need to be made but not how they should be made." (1974, p. 148)

E. R. House (1980, pp. 119-120) outlines the possibilities one evaluator faces in the design of an evaluation:

> "He may administer a standardized achievement test and compare the scores to those of students who do not have such a program. He may create a special test, perhaps based on objectives, to assess certain areas of academic 'deficiency'. He may measure student attitudes, opinions, or self-concepts...he may solicit teacher opinions, attitudes, and judgments about the program. He may observe how teachers and students behave in the classroom, in the halls, in the streets. He may record and analyze what teachers say about students, how they grade them, what standards they set. He may ask parents about the program, about the teachers, about the schools. He may examine parent participation in school activities or school efforts to involve parents. He may solicit the opinion of employers about the content or results of the program..."

House (1980, p. 120) then asks, "What does the evaluator do?", and responds that the evaluator uses intuition and what he thinks is "just" when he designs the evaluation. This dissertation proposes that the answer to House's question, "What does the evaluator do?" is that he can use a decision analysis model for the selection of variables for his program evaluation.

THE VARIABLE SELECTION MODEL

Evaluators of educational programs must provide information that will help others make decisions about the merits of particular programs. Often, the information provided by the evaluators leads to crucial decisions with respect to funding programs, hiring faculty, adapting curriculum materials, etc. The evaluators themselves also have crucial decisions to make, decisions that must be made prior to and during the evaluation. The evaluators must decide how to conduct the evaluation, what measures are to be included, when data are to be collected. There is usually a wide range of options in the design of an evaluation: many valid tests are available; there are many days in the year when testing could be done; there are many pertinent behaviors that could be measured.

Decision theory can offer an orderly method of choosing among the evaluation design options. It could allow evaluators to select the most useful set of measures and could eliminate redundant, useless, uninformative testing.

According to Worthen and Sanders (1973, p. 122fn.):

> "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."

The model proposed herein seeks to rectify part of this situation, i.e. it seeks to offer a methodology that is objective and that will aid in the decision of which

variables to study in an evaluation.

In this section a model for variable selection will be outlined that will give evaluators a systematic method for choosing the set of variables or measures that can be most useful to them in writing their evaluation report and in assessing a program's worth. The model is an outgrowth of Multi-attribute utility theory (MAUT) as enunciated by Edwards, Guttentag and Snapper (1975) and Keeney and Raiffa (1976). MAUT provides methods that aid in the selection of a single most useful entity; the proposed model adapts MAUT models so that a composite decision, or a decision among sets of entities, can be made.

Typically this model for variable selection will be used when evaluators are considering the use of a large number of measures, all of which they would like to use but due to considerations of time, cost, differences in validity and reliability, etc. some variables must be eliminated. It has been found that when there are large numbers of variables varying on many attributes, selection decisions made intuitively (i.e. without the assistance of a model) cannot deal with more than 5 or 6 attribute dimensions. (Slovic & Lichtenstein, 1971; Kozielecki, 1970) This model helps decision makers choose the set of variables that are most useful by breaking down the decision process into manageable steps.

The model provides a systematic procedure for variable selection; all of the bases on which variables are selected

can be incorporated into the model. The selection procedure is open to scrutiny; the criteria of selection can be analyzed and criticized by others. The model gives evaluators a practical and orderly solution to the problem of choosing among many variables; the evaluators' judgments and their assessment of the worth of each variable are not ignored, but form an integral part of the model.

A situation such as the one described below would be representative of the occasions when the variable selection model would be used.

Let us say that a team of evaluators must plan a measurement strategy to evaluate a city-wide enrichment program for junior high school students. It is a program geared for those junior high school students whose lack of basic skills may assure their failure in high school. The program has as its primary goal the improvement of students' skills to the point at which success will be noted in academic areas as well as in self-concept.

After a year of operation, the program is to be evaluated. In this summative evaluation, the evaluators must decide whether the goals of the program have been achieved. They must make this decision on the basis of various measures or variables. Because the evaluators' recommendations or decisions about the program have a far-reaching effect, possibly leading to discontinuing the funding of the program, they wish to be as sure of their conclusions about the program as they can be. In order to assure themselves

that they are making the "correct" decisions about the program, they will select measures that they believe will provide an accurate assessment of the program.

In this example the evaluation team proposes 25 measures that would allow them to best assess the program's efficacy. The measures, $m_1, m_2, \ldots m_{25}$, include:

A) Standardized measures -- achievement tests in reading and mathematics and other academic skills, standardized self-inventories regarding motivation, attitudes, self-concept, etc.

B) Teacher-made or informal measures -- questionnaires, interviews with students and teachers, self-evaluation reports, teacher-made achievement tests, teacher-made attitude scales, etc.

C) Records -- Grade Point Average, anecdotal information of critical incidents, attendance records, dropouts, extracurricular activities, etc.

By using all 25 measures the evaluation team feels they would have a complete picture of the program's achievements. However, the evaluation team recognizes that due to the exorbitant costs involved in utilizing 25 measures, as well as the huge blocks of time needed to compile the data, they will have to eliminate some of the measures. The evaluators want to select those measures that will provide the most information yet not exceed the budget, for example, nor the time constraints to which they must adhere.

The proposed model for variable selection has as its conceptual basis the utility of the measures for an evaluation. Since evaluations are conducted for practical purposes rather than for research, experimental or intrinsic interest in the results, it was felt that a model that em-

phasizes the usefulness of evaluation was most appropriate. This approach allows the viewpoints and needs of the program's participants to be considered in conjunction with objective evidence such as test scores, national norms, etc.

In proposing a model based on utility, some information is lost. When every aspect of a program is not examined, there are gaps of knowledge. There may be some outcomes of the program that were not anticipated and, therefore, may be overlooked. However, since most evaluation teams must place limits on what they assess, a model that selects what aspects to assess based on how useful the aspects will be to the eventual evaluation and decision making process will include the most essential information to permit the evaluators to perform most fairly.

The model recognizes that measures overlap and produce redundant information; it allows the intuitions of program participants to be expressed. The variables to be used are examined in light of the characteristics -- cost, reliability, etc. -- deemed to be important and useful to the evaluators. What is sacrificed in this approach is an explanation of why a program works; the focus of the model is on whether a program works.

The model is flexible; it is specific to the requirements of those involved. They may emphasize whatever they wish; if reliability is more important to them than validity, the measures that are selected with the model will reflect this. If one type of validity has greater impor-

tance than another, this, too, will be incorporated into the selection model.

A step-by-step description of the variable selection model follows:

Step I -- Define the variables. Each measure is described and labeled (m_i). Thus, m_2 might be an attitude scale, m_{10} might be a standardized reading achievement test.

Step II -- Enumeration of decision alternatives. The sets of variables among which the evaluators must choose are listed. Let us say that there are 15 variables and the evaluators wish to use between five and ten variables. In such a case, they will list the combinations of five to ten variables from which they must choose. (The choice to limit the number of variables in a set to specific ranges might arise if evaluators wished or were required to use certain tests or a specified number of tests from various categories of measures.) When the combinations of variables are innumerable, a computer is used to assist in listing the decision alternatives.

Step III -- Identify the dimensions of value. Dimensions are those attributes or aspects of a measure or variable the evaluators consider to be important in the decision of whether or not to use or discard a variable. In the case of evaluating an educational program, the evaluators might designate the following dimensions: D_1 -- validity, D_2 -- reliability, D_3 -- cost, D_4 -- administration time.

It is more than likely that with a large-scale evaluation with measures covering a broad range of abilities there will be more than a single dimension of validity. In such a case, the evaluation team can define several dimensions of validity: D_1 -- validity with respect to reading ability, D_2 -- validity with respect to prediction of GPA, etc.

Step IV -- Assignment of ratio-preserving weights. A weight is assigned to each dimension. Here, the decision makers ascertain how much more important dimension x is than dimension y. The evaluators might approach the assignment of weights to the dimensions in the following ways: The dimensions are ranked according to relative importance; then the evaluators ask themselves how much more important it is to use a measure having dimension x than dimension z (the least important dimension). How much more important is it to use a measure having dimension y than dimension z? The decision makers can juggle the weights they give in response until they are satisfied that their judgments are adequately represented.

For ease in calculation, the weights are rescaled using the common MAUT formulation:

$$d_x = (w_x/(w_x+w_y)) \times 100$$

$$d_y = (w_y/(w_x+w_y)) \times 100$$

(w = original dimension weight; d = rescaled dimension weight; x, y = individual dimensions.)

Step V -- Determination of the composite entities for each set of variables, i.e. the composite cost, the

composite reliability, etc. for each set of variables.

The composite entities are determined in distinct ways:

a) For entities like cost or time, the composite cost or composite time are additive. Thus, the cost of using variables m_3, m_6, m_8 will be the sum of the individual costs. $Cost(m_3,m_6,m_8) = Cost\ (m_3) + Cost\ (m_6) + Cost\ (m_8)$.

b) Derivation of composite reliability. Several methods for estimating the reliability of a set of tests i.e. composite reliability, have been put forth, (Cronbach, 1951, Novick & Lewis, 1967, Tryon, 1957, etc.) Most of these methods are based on the assumption that each of the tests that make up the set of tests is parallel to the others.

One formulation of composite reliability that does not assume that the measures in the composite are parallel is presented by Nunnally (1967, 229) and is also cited by Lord & Novick (1968, 100) and Tryon (1957).

In this formulation the composite reliability is given as the ratio of the composite true score variance to the composite observed score variance:

(1) $$r_{oo} = \sigma_{to}^2 / \sigma_o^2$$

where σ_{to}^2 = variance of the true scores of the composite and σ_o^2 = variance of the observed scores of the composite. This formulation is an extension of the basic definition of reliability for a single measure:

(2) $$r_{ii} = \sigma_{ti}^2 / \sigma_{oi}^2$$

where r_{ii} = reliability of a single measure

$\sigma^2_{t_i}$ = variance of true score of the measure

$\sigma^2_{o_i}$ = variance of observed score of the measure

The variance of obtained scores of the composite (σ^2_o) is equal to the sum of all of the elements in the covariance matrix of the measures comprising the composite. The composite true score variance, $\sigma^2_{t_o}$, equals the sum of the elements for the true score covariance matrix. The true score covariance matrix is identical to the covariance matrix for observed scores except for the diagonal elements. Thus, $\sigma^2_{t_o}$ will equal the covariance matrix of observed scores minus its diagonal elements plus the diagonal elements of the true score covariance matrix, i.e.:

$$\sigma^2_{t_o} = C_o - D_o + D_t \quad (3)$$

where C_o = sum of all elements of the covariance matrix for observed scores
D_o = sum of diagonal elements from the covariance matrix for observed scores
D_t = sum of diagonal elements, from the covariance matrix for true scores

In the covariance matrix for observed scores, the diagonal elements are the variances of the individual measures, so

$$D_o = \Sigma \sigma^2_{o_i} \quad (4)$$

In the covariance matrix for true scores, the diagonal elements are the variances of true scores of the individual measures, so

$$D_t = \Sigma \sigma^2_{t_i} \quad (5)$$

However, in equation (2)

$$\sigma^2_{t_i} = r_{ii} \sigma^2_{o_i}$$

therefore

(6) $$D_t = \sum r_{ii}\sigma_{oi}^2$$

Substituting the quantities from equations (4) and (6) in equation (3);

(7) $$\sigma_{to}^2 = C_o - \sum\sigma_{oi}^2 + \sum r_{ii}\sigma_{oi}^2$$

Furthermore, substituting the value for σ_{to}^2 from equation (7) in equation (1):

(8) $$r_{oo} = \frac{C_o - \sum\sigma_{oi}^2 + \sum r_{ii}\sigma_{oi}^2}{\sigma_o^2}$$

Since $\sigma_o^2 = C_o$,

$$r_{oo} = 1 - \frac{\sum\sigma_{oi}^2 - \sum r_{ii}\sigma_{oi}^2}{\sigma_o^2}$$

c) Determination of composite validity. In a previous step it was noted that there may be several dimensions of validity; the evaluators may want to select tests that are valid along more than one dimension. They may desire tests that correlate with several criterion variables. Accordingly, for each of the dimensions of validity, the composite validity (or the validity of groups of measures with regard to each dimension of validity) must be ascertained. If test validity is considered to be the extent to which the test measures a particular criterion, then the validity of a set of measures (composite validity) will be the extent to which the set measures or predicts the criterion or variable of interest.

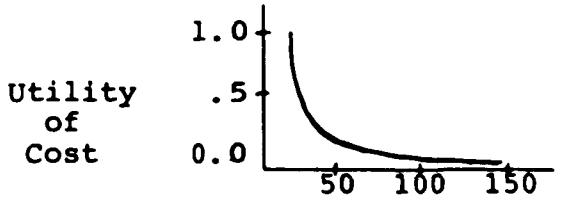
For each dimension of validity a criterion variable is defined, e.g. if it is desirable to have measures that are good predictors of academic success, the criterion variable

for this dimension of validity could be GPA. Once a criterion variable is established, the degree of association between the criterion and each set of measures is calculated; this degree of association is represented by the multiple correlation between the criterion and sets of measures. The multiple correlation is considered to be the composite validity coefficient. Thus, if the criterion is GPA and the validity of measures 2,4,5 with respect to GPA is to be determined, one would calculate the multiple correlation of measures 2,4,5 with GPA, i.e. $R_{GPA(m_2m_4m_5)}$.

The correlations of the individual measures with the criterion variables and with each other, which are needed for the calculation of multiple correlations, could come from several sources: 1) data provided by test manufacturers, 2) data taken from prior administrations of the measures, 3) data taken from pilot testing the measures.

Step VI -- Construction of utility functions of sets of variables on single dimensions. For each dimension a utility function is defined. The utility function expresses the relationship between the numerical values of the measures on each dimension and the intrinsic worth of these numerical values to the evaluator. Thus, for the dimension of "cost" the utility function describes the intrinsic value associated with a given cost. The following could be a utility function for cost, in which as cost rises its utility drops and when the cost reaches a certain point

($100) the utility is so small that at this cost and higher the utility of the measures is virtually nil.



To define the utility functions for the dimensions several approaches can be used. Fishburn (1967) cites 24 methods of estimating utility functions. Some aspects of the various approaches include the use of probabilities, preference judgments, indifference judgments. One method that would allow evaluators to express judgments about the utility of measures with respect to the dimensions is the standard gamble technique. This method can handle dichotomous or continuous data and operates in the following way:

I. The outcomes of measurement are ranked in order of desirability: A (most desirable outcome), B (some intermediate outcome), C (least desirable outcome); in symbols, $A > B > C$. Suppose, for example, the dimension were one of reliability. One would designate as A the reliability value of 1.00, the highest possible value for reliability; C would be a reliability of 0.0, the lowest possible value.

II. A choice is set up between the following options:

a) outcome B for sure or

b) either A or C, with Prob A = p and Prob C = 1-p.

In other words, one must choose between the intermediate value (B) for certain or the possibility of the best outcome

(A) with a specified probability (p) and the worst outcome (C) with a specified probability (1-p).

III. A selection is made between choice a and choice b as p is allowed to vary. Thus, if p=1, the choice would be between outcome B for sure vs. outcome A for sure, and certainly the choice made would be outcome A (i.e. choice b). If p=0, the choice is between outcome B for sure vs. outcome C for sure, and the choice made would be outcome B (i.e. choice a).

IV. As p decreases from 1 to 0, the choice will change from choice b to choice a. There will be a point at which there is indifference to choices a and b. At this point, the utility of B is expressed as:

$$u(B) = p \cdot u(A) + (1-p) \cdot u(C)$$

where u(A) = utility of outcome A, u(B) = utility of outcome B, etc.

V. Since one can linearly transform a utility function, one can assign utilities to the most desirable and least desirable outcomes as follows: u(A) = 1.0, u(C) = 0.0. Thus, at the point of indifference

$$u(B) = p(1) + (1-p)(0) = p.$$

In other words, the probability value which yields indifference is the utility of B.

Step VII -- Calculation of overall utility for each set of measures, $U_{(m_a m_b \ldots m_k)}$, i.e. the utility of a set with respect to all dimensions. The formulation for overall utility follows the linear pattern advocated by most of

those involved in utility theory. According to Dawes and Corrigan (1974, p. 102): "...the linear model catches the essence of a judge's expertise and at the same time eliminates unreliability..." The formulation for utility of a set of variables across all dimensions uses the dimension weights (d_j) from step IV and the utility (u_j) from step VI:

$$U_{(m_a m_b \ldots m_k)} = \sum_j d_j u_j (m_a m_b \ldots m_k)$$

where j=dimension, d=dimension weight, u=utility on a particular dimension, $m_a m_b \ldots m_k$= variable set.

Step VIII -- Decision making. When each $U_{(m_a m_b \ldots m_k)}$ is determined, the set with the largest U , i.e. the set with the largest utility value across dimensions, will be selected by the evaluators as the best set of variables for their evaluation study.

DECISION ANALYSIS AND VARIABLE SELECTION

A model for variable selection based on decision analysis is being proposed. The type of decision analysis that the model relies on is an adaptation of the MAUT decision model. The model proposed here is designed to choose among sets of independent and non-independent variables. MAUT models have been used previously in several fields including evaluation, Edwards, Guttentag and Snapper (1975), but they have always been limited to decisions about the single most useful entity; additionally, the decision alternatives have been independent of each other. MAUT models have been used in evaluation to decide what program should be funded, what program meets the community's needs, which curriculum is best for the school, etc.

In a previous section it was noted that in program evaluation as in all branches of educational psychology and other behavioral sciences, variable selection is most frequently accomplished by relying on evidence that a relationship exists between the variable and the expected outcome. Typically, evaluators decide to use certain variables (test data, observations, etc.) because they feel that these data will indicate whether or not the program has achieved its goals; in these cases, variable selection can be based on intuition, past experience, the recommendation of others, the preferences of administrators, etc. What these methods lack is the ability to project how useful the data will be,

or which variables should be used in combination with others.

Program evaluators using the decision analysis model would have before them the set of variables among all the alternatives that would be most useful to them in evaluating the success of an educational program. Instead of haphazard decisions regarding the choice of a set of variables, the decision analysis model permits a systematic procedure which breaks down a decision with an unmanageable number of factors into discrete, easily handled steps. The model forces decision makers to examine what aspects (dimensions) of the variables make them useful. The model allows the decision makers to consider all sets of variables on the basis of all the stated dimensions. Without the model, a decision among many sets of variables with several dimensions could not be handled without ignoring certain dimensions or certain sets of variables since there are too many factors to be considered simultaneously. The model makes a cumbersome decision process an orderly one.

Program evaluators are often confronted with many appropriate tests; some tests are more difficult to administer than others; some are more costly; some have been used more often thus yielding valuable prior data; some are more reliable than others; some are more valid than others. Unless there is one ideal test, a test that is the least difficult to administer, the least costly, the most frequently used, the most reliable and the most valid in comparison to all others of its type, then the decision of which test(s)

to use becomes complicated and in all likelihood it would be made arbitrarily. MAUT decision analysis could remove the arbitrary nature from these decisions and provide a method of variable selection that can consider the cost, reliability and whatever other pertinent details the evaluators put forth.

Although one of the major roles of evaluation is "treatment adequacy assessment" (Popham, 1972, p. 5), there is presently no methodology for planning an evaluation so that the data provided will allow the evaluation's decision regarding the adequacy of a program to be optimal. Because the future of a program depends so much on the findings of the evaluation, it is imperative that the evaluators' assessment of the program be as fair and accurate as possible. The evaluators must decide which kinds of data would allow their recommendations to be as accurate a representation of the program as is possible.

Most methods do not take in consideration the question of how useful certain design strategies might be. "Methods for incorporating concerns about the usefulness of an evaluation into design and analysis strategies are yet to be developed." (Rogosa, 1978, p. 81) The proposed model might be such a method, for it would allow the evaluators to use a utility table in conjunction with other data; the utility table could express the usefulness of each evaluation strategy.

In summary, program evaluation lacks concrete methods

for how to select variables. Program evaluation needs guidelines in order to select variables that will provide the most useful information to the evaluators at the least cost and with the best use of time. Variable selection should be orderly, reasoned and coherent since all relevant tests cannot be administered. Variable selection should be flexible since certain data may indicate that changes should be made in the initial evaluation strategy. Decision analysis would permit evaluators to select variables in a logical, cost-effective way, and it would incorporate previous findings and present needs and judgments into the evaluation plan.

Another factor that would recommend the use of an MAUT-based decision analysis model in the selection of variables is the previous use of MAUT methods in many practical decision making situations. A description of a few of these situations follows on the next pages.

PRIOR USE OF MAUT MODELS IN REAL LIFE SETTINGS

There have been many studies that have used the MAUT paradigm in real life settings. These settings are extremely varied and diverse, covering areas from business investment to job preference.

The number of studies that apply MAUT grew dramatically in the 1970's soon after the work of Raiffa (1968) and later Keeney and Raiffa (1976) became known.

In the fields of psychology and education, however, there have been few applications of the MAUT approach. Some proponents of MAUT have applied the methodology at the Office of Child Development (OCD) as described by Guttentag, (1973) and Guttentag & Snapper, (1974). OCD wished to make plans for future fiscal years. They wanted to evaluate strategies and programs that were to be recommended for the future. A small group at OCD spent 2 days discussing what appropriate dimensions of value would be; they formulated a list of 14 dimensions including: promote child health, develop leadership capacities, promote self-respect. Each member of the group ranked the dimensions and assigned weights individually to the dimensions. After this process, they decided to eliminate one dimension. Whenever there was disagreement about the weights that each dimension should have, a median value was used to reflect the group consensus. Then, from a wide range of sources came a list of the strategies and the programs they were to consider;

there were more than one hundred recommendations. Thereafter, each recommendation was indexed for each value dimension. The group was told to estimate the degree to which a recommendation reached or contributed to a value on a scale of 1 to 1,000. Three people judged each value dimension, and each person was assigned three value dimensions. Thus, 3 scores were assigned to each recommendation on each value dimension. Additive utilities were calculated with $U_i = \Sigma w_j u_{ij}$ when w_j = weight of the value dimension, u_{ij} = utility of individual recommendation on each dimension. When this process was completed, it was felt that changes should be made in both the value dimensions and the recommendations. Therefore, each of these categories was refined and generalized and the MAUT steps were reapplied.

A rural mental health center used MAUT to aid its decision making in setting priorities and goals for the center. (Gibson, 1976) First, the committee's goals were identified; then, the purpose for constructing a decision making model was stated. The entities or, in this case, actions to be taken were identified by asking the staff to rely on their perceptions, areas of interest, needs, etc. The dimensions of value were identified by a committee of five who reduced a list of 13 dimensions to 9 dimensions which were then ranked in importance by each committee member. Importance preserving ratios were assigned to the dimensions by each member; the median value of the members' assignments was used to identify the weight for each dimen-

sion. A 1 to 7 scale was created for each value dimension, and committee members estimated the plausibility for each entity (action) on the scale. The utilities were calculated by multiplying the plausibility ratings by the value dimension weight and summing across dimensions for each entity. It was noticed that in the decision making process the use of several dimensions rather than one led to less variability and less error. Even when weights were not applied to the dimensions, the ranking of the entities was similar. Among the benefits of using a MAUT-based model were: a sense of satisfaction due to the use of a "rational, systematic methodology"; the establishment of priorities was accomplished in a shorter time; the final priority ranking was strongly supported and accepted by the staff members; the use of the model affected other decision making at the mental health center.

In Keeney and Raiffa (1976) reference is made to several studies that have described the use of MAUT as a basis for decision-making. Among the areas studied are:

1. A hospital's policy regarding investment in equipment was studied by Bodily (1974). The decisions to be made were a) whether to invest in expensive blood freezing equipment and b) what proportions of frozen or non-frozen blood should a blood bank have. Objectives for blood banks were obtained from members of blood banks. Six attributes were listed, including cost, delay, wastage, purity, age, ability to meet special needs. Some attributes were aggre-

gated or eliminated, so that there remained three attributes: X_4=purity, X_5=age and a compound attribute, Y, where $Y=X_1+dX_2+wX_3$, with X_1=cost, X_2=delay, X_3=wastage and d=cost per unit of blood delayed and w=cost per unit of wastage. Subsequently, X_5, age, was eliminated, so the utility function became $u(Y, X_4)$ which was assessed by two decision makers. For both decision makers Y and X_4 were mutually utility independent; thus, they applied the theorem that alternative X^A is more or at least as preferable as X^B if $u(X^A) \geq u(X^B)$. To aggregate and exclude attributes, they were arranged hierarchically and the unimportant ones were discarded. Then, value functions over attributes were assigned in order to further reduce the number of attributes; then, quantitative decisions led to the elimination of other attributes.

2. The selection of a job from among several alternatives. (Teweles, 1972) The desirability of a profession was determined from several attributes or objectives: job satisfaction, wealth, security, family considerations, independence, self-esteem, prestige. An index from 0-100 was used to show to what extent specific professions reached the objective. The author evaluated five job alternatives in terms of the objectives, using an additive utility function. Available data plus personal judgment allowed him to assess the probabilities with regard to the job alternatives meeting the objectives. The author concluded that

he felt more able to make the proper decision, and in addition that through the analysis he learned more about his own goals and about the details of each job alternative.

3. Yntema & Klem (1965) studied the safety of landing an airplane under various conditions. Twenty Air Force pilots were the decision makers; they had all landed aircraft in diverse situations. The decision makers assessed the utility functions over the attributes a) ceiling, b) visibility, and c) remaining fuel. To find the utility function, the additive utility function of three attributes was used:

$u(x_1x_2x_3) = k_1u_1(x_1) + k_2u_2(x_2) + k_3u_3(x_3)$ where k_i = weights, u_i = utility, x_i = attributes. Furthermore, each pilot was given 40 pairs of consequences and had to pick the pair that was preferable. The authors felt that the utility functions and the preference choices were comparable.

VALIDATION OF THE MODEL

The validation of the proposed variable selection model was accomplished by comparing the judgments made with the model to judgments made without the model. It was theorized that if the judgments made with the assistance of the model were similar to the judgments made without the model, this would be tantamount to establishing the model as being able to adequately represent the "true" behavior of decision makers, thus validating the model.

According to vonWinterfeldt & Fischer (1975, p. 50):

> "...a rigorous axiomatic test of [MAUT] models is impossible in complex real situations because it would require judgments which the decision maker is unable to make, for example, ordering complex alternatives consistently. It was just this inability which led to the application of MAUT as a decision aid."

Some behavioral scientists have stated that validation is not possible and that MAUT-based models should simply be accepted as valid. (Shepard, 1964; Hoepfl & Huber, 1970) Of those who have validated MAUT models, the most common method is one of convergent validity, whereby the MAUT model is compared to one or more other models, as in Huber et al. (1971) where subjects' intuitive judgments regarding hypothetical jobs were correlated with their ratings of the same jobs using a MAUT-type rating model. Similarly, Yntema & Klem (1965) had pilots rate the adequacy of landing sites described by three attributes. The subjects first made intuitive, wholistic judgments and then used the de-

composition processes of an MAUT model to rate the landing sites. In these two studies and others a high positive correlation was found between the wholistic judgments and those made with the model, and, therefore, the MAUT models were found to be valid predictors of intuitive selections.

In the studies cited above, the MAUT models dealt with the rating of a single item--a job, a landing site--on a number of attributes. In the present situation, the model is based on MAUT but deals with assessing sets, rather than single entities. However, just as the method of correlating selections made with and without the models was utilized in validating MAUT models in a single item context, so, by analogy, can this method be used in validating a model dealing with sets of items.

In this validation study, the subjects are asked to rank sets of variables, each variable being described in terms of 3, 5 or 7 dimensions. This ranking will be designated as the "intuitive" ranking, in that the ranking is accomplished without any formal, systematic model. In addition, the subjects rank the sets of variables using the procedures outlined by the model. The subjects supply the weights for the 3 or 5 or 7 dimensions and construct utility functions for the variable sets on 3 or 5 or 7 dimensions. These weights and utility functions are inserted into appropriate places in the model's procedures, and the outcomes are the utilities of each set which are then ranked. This ranking is designated as the "model" ranking. To say that the

model is valid, it is hypothesized that there is a significant positive correlation between the intuitive ranking and the model ranking and, thus, between the judgments made with the model and without it. In other words, the null hypothesis is that the intuitive ranking and the model ranking are independent, as opposed to the alternative hypothesis that they have a positive rank-order correlation. (H_0: ρ=0). For those subjects making judgments about sets of variables described by only 3 dimensions, it would be expected that the hypothesis would be rejected. The subjects dealing with variables having 5 or 7 dimensions would be expected to suffer from cognitive overload in making the intuitive ranking judgments. This would make the intuitive rankings less useful and less representative of their "true" judgments had their cognitive capabilities not been exceeded. Because of the reduced efficiency of these subjects in regard to the intuitive rankings, the correlation of the model ranking and the intuitive ranking would be expected to be much lower than the correlation for the subjects dealing with 3 dimensions. Thus, while it is reasonable to expect that the hypothesis $\rho = 0$ would be rejected for those making judgments based on 3 dimensions, it is also to be expected that as the dimensions increase from 3 to 5 or 7 dimensions, the correlation of judgments made with and without the model will diminish.

Method

The subjects were 76 graduate students at two branches of the City University of New York who were taking basic courses in psychology and statistics. All of the subjects had familiarity with the facts of elementary statistics and psychometrics and knew terms such as reliability, validity and correlation. The subjects were divided into three subgroups: Subgroup 1, n=25, subgroup 2, n=26, subgroup 3, n=25.

The validation study collected two kinds of information:

1) information about ranks of sets of measures
2) information about the relative importance given to certain attributes (dimensions) and the relative value of specific levels of these attributes.

To gather the first kind of information, all subjects were presented with a description of a hypothetical educational program for bilingual children for which an evaluation is to be designed. The subjects were told that the hypothetical program was in a local Junior High School, that it was an enrichment program in which English language skills were emphasized. The subjects were informed that there were 10 measures available that could each provide useful information to aid the evaluator.

Each subgroup was given a chart describing the 10 variables proposed for the evaluation. Subgroup 1 subjects

received a chart that described the 10 variables along 3 dimensions; subgroup 2 received a chart describing the variables along 5 dimensions; subgroup 3 received a chart describing the variables along 7 dimensions. (See Table 1 for a listing of the dimensions for each subgroup.) The seven dimensions were 1) Administration time, which was described as the number of minutes needed to administer the measure; 2) Cost of the measure, in terms of dollars; 3) Correlation of the measure with teacher-made classroom measures, expressed as a correlation coefficient; 4) Reliability, expressed as a reliability coefficient; 5) Minority non-bias rating, in terms of a scalar value from 0--10, with 10 representing the point of least bias; 6) Ease of administration, represented on a scale of 0--10, with 10 equivalent to the point of easiest administration; 7) Correlation with Grade Point Average, expressed as a correlation coefficient. (See Appendix A for a sample chart.)

Based on the information about the variables provided in the charts, each subject was asked to select the best set of three measures to be used for the evaluation and the second best set of three measures. In the selection of the second best set of three measures, subjects were told that a particular measure could be part of both the first and the second best sets. Thus, one could select measures 1,4,10 as the best set and measures 1,4,9 as the second best set.

Each subject, therefore, selected 2 sets of three mea-

Table 1

Dimensions of the Variables for Subgroups

Dimension	Subgroup 1	2	3
1. Administration Time	*	*	*
2. Cost	*	*	*
3. Correlation with classroom measures	*	*	*
4. Reliability		*	*
5. Non-bias rating		*	*
6. Ease of administration rating			*
7. Correlation with GPA			*

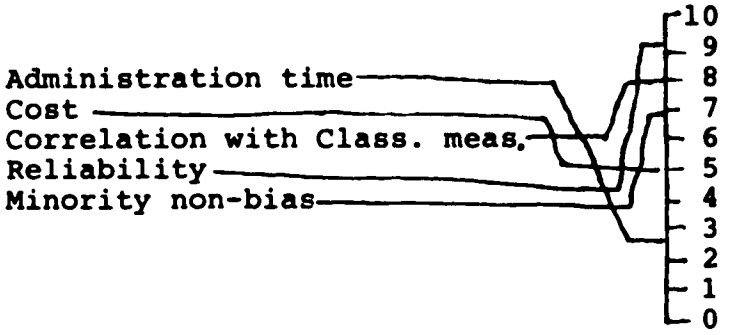
Note: An asterisk indicates the inclusion of the dimension in the description of the variables.

sures. The selection designated by the subject as the best set was given a value of 10 points; the second best set was given a value of 5 points. Within each subgroup, the accumulated points for each set were summed, and the sets were ranked according to descending point value. This is the "intuitive ranking" of the sets, the ranking made without benefit of the model, and it is based on how frequently a set was chosen by the subjects. Thus, if in subgroup 1 the set 1,4,10 was the first choice of 3 subjects and the second choice of 3 other subjects, it would have a total point value of 45 (30 + 15) in that subgroup and would be placed in the intuitive ranking below sets receiving 50 or more points.

The second type of information obtained in this study were a) the weights given to dimensions and b) the relative value of different levels of these dimensions. a) Subjects in subgroup 1 were given a list of the three dimensions that had described the 10 measures in the initial stage of the study; subjects in subgroup 2 received a list of 5 dimensions, and subgroup 3 received a list of 7 dimensions. The subjects denoted on a scale of 0--10 the relative importance, or weights, of each of the 3 or 5 or 7 dimensions in the selection of measures. A graphic scaling method used by Eckenrode (1965) was used; each subject was asked to draw a line from the dimension name to the point on the scale indicating the relative weight for each dimension (See Figure 1 for an example of this procedure.)

Figure 1

Example of a Completed Weighting of Dimensions for a Subject in Subgroup 2



b) Subjects were asked to indicate on a scale of 0-10 the relative value of different levels of the dimensions. The method used was again the graphic scaling approach that is illustrated in Figure 1. For each dimension, four levels were to be rated. For example, for the dimension of cost, subjects were asked to indicate their ratings of a set of measures that cost 1) $3.00, 2) $6.00, 3) $10.00, 4) $18.00. The subjects of subgroup 1 had to rate four levels of each of the three dimensions; subgroup 2 rated four levels of five dimensions; subgroup 3 rated four levels of seven dimensions. In this second stage, therefore, subjects of subgroup 1 gave 3 dimension weights and 12 preference ratings of the levels of the dimensions; subgroup 2 gave 5 dimensions weights and 20 preference ratings; subgroup 3 gave 7 dimension weights and 28 preference ratings.

In each subgroup, the dimension weights of each subject were averaged, yielding the mean dimension weights (w_j) for each subgroup. In addition, within each subgroup, the subjects' ratings of each of the 4 levels of the dimensions were averaged, yielding mean preference ratings (y_{jk}) for each of the levels of the dimensions. These mean preference ratings were used as the utility functions of the dimensions for each subgroup; using preference ratings as utility functions is a common method of constructing utility functions. Thus, if the mean preference rating of dimension X at level Y were equal to Z, any set with the value of Y

on dimension X would have a utility of Z on that dimension. If a set had a value on a dimension falling between levels Y and T, then the utility value would be found by interpolating the mean preference ratings of levels Y and T.

These utility functions together with the dimension weights were used to calculate the mean utility value of each set of three measures that had been chosen in the initial stage of the study. The mean utility of each set, within the subgroups, was determined by the formula described by the model:

$U_{set} = \sum_j w_j u_{j(m_a m_b m_c)}$ where j=dimension, w=dimension weight, and $u_{j(m_a m_b m_c)}$ = utility of set $m_a m_b m_c$ on dimension j.

When each set's utility value was determined, the sets were ranked, with the set having the largest utility value being ranked first. This ranking is termed the "model ranking". Thus, in the end, there was a model ranking of the sets and an intuitive ranking of the sets for the three subgroups. These two rankings for each subgroup were correlated using the rank-order correlation:

$R = 1 - (6 \Sigma d^2 / [N(N^2-1])$ where d=difference in rank and N = number of items ranked.

Results

Intuitive Ranking:

The intuitive ranking for each subgroup was the ranking based on the frequency of selection of a particular set. The set chosen most often by the subjects was the highest ranking set, and so forth. In subgroup 1, there were 18 sets that were selected and ranked; in subgroup 2 there were 29 sets, and in subgroup 3 there were 26 sets that were selected. The set ranked as number one in subgroup 1 was chosen by 14 subjects; the number one ranking set in subgroup 2 was chosen by 6 subjects, and the number one ranking set in subgroup 3 was chosen by 6 subjects. (See Table 2.)

Model Ranking:

In this ranking, the sets that had been selected in the intuitive ranking were now ordered according to their mean utility value, which was derived from the mean dimension weights (w_j) and the mean ratings of the levels of the dimensions (y_{jk}).

The mean dimension weights for each subgroup were compiled from the dimension weights designated by the subjects. (See Table 3 for a summary of the mean weights for each subgroup.) There were fairly high levels of agreement among the subjects with respect to the dimension weights. For example, in subgroup 1, 23 of the 25 subjects gave dimension 3 (correlation with classroom tests) the greatest weight. In this subgroup, 12 subjects gave the second highest weight

Table 2

Frequency Data -- Intuitive Ranking

Group	No. of sets selected	No. of subjects selecting first-ranking set
Subgroup 1 (n=25)	18	14
Subgroup 2 (n=26)	29	6
Subgroup 3 (n=25)	26	6

Table 3

Mean Dimension Weights

Dimension	Subgroup 1	2	3
1. Administration Time	6.37(2.6)	5.4(2.2)	4.8(2.4)
2. Cost	6.6(1.8)	5.9(2.4)	5.1(2.7)
3. Correl. with class tests	9.4(1.1)	8.1(1.5)	7.7(1.8)
4. Reliability		8.8(1.1)	8.6(1.3)
5. Non-bias rating		8.0(2.4)	7.6(2.0)
6. Ease of administration			5.1(1.7)
7. Correlation with GPA			7.2(2.6)

Note: Figures in parentheses are the standard deviations.

to dimension 1, administration time, and 11 subjects gave the second highest weight to dimension 2, cost. In subgroup 2, of the 26 subjects, 17 gave the greatest weights to dimensions 3, 4, and 5 (correlation with classroom tests, reliability, non-bias rating); 16 of these 26 subjects gave dimension 1 (time) the least weight. In subgroup 3, of the 25 subjects, 12 gave dimension 4 (reliability) the greatest weight.

The mean preference ratings in the 3 subgroups were compiled from the ratings assigned by the subjects to each of 4 levels of the dimensions. (All of the mean preference ratings accompanied by their standard deviations can be found in Table 4.) As mentioned previously, the mean preference ratings were used as the utility functions from which the utilities of a set of variables on each dimension were derived.

Once the weights of the dimensions and the utilities for each set on each dimension were found, the mean utility value for each set was calculated, based on the equation,

$$U_{set} = \sum_{j} w_j u_j(m_a m_b \ldots m_k).$$

The sets of each subgroup were ranked according to utility value (model ranking). This ranking was correlated with the ranking made on the basis of the data chart (intuitive ranking). For subgroup 1, which selected sets based on 3 dimensions, R = .68; for subgroup 2, which dealt with 5 dimensions, R = .04; for subgroup 3, dealing with 7

Table 4

Mean Preference Ratings of Dimension Levels

Dimension Level	Subgroup 1	2	3
Administration Time			
60 minutes	8.7(1.8)	9.0(1.5)	9.1(1.3)
90	7.2(1.9)	7.5(1.5)	7.5(2.0)
120	4.6(2.4)	4.3(2.0)	4.9(2.6)
180	2.3(2.4)	2.1(2.3)	2.8(2.9)
Cost			
$3.00	9.0(1.4)	8.8(2.1)	9.0(1.9)
6.00	7.3(1.5)	7.4(1.7)	7.7(2.1)
10.00	3.9(1.9)	4.8(2.1)	5.3(2.5)
18.00	.9(1.5)	1.7(2.0)	2.5(2.5)
Correl. with Class Tests			
.9	9.4(.9)	9.2(.9)	9.4(1.1)
.8	8.1(1.3)	8.1(.8)	8.0(1.3)
.6	5.0(1.7)	5.5(1.5)	4.7(1.8)
.2	.8(1.5)	1.4(1.8)	.9(1.2)
Reliability			
.99		9.6(.7)	9.8(.6)
.90		8.5(.9)	8.9(.6)
.85		6.8(1.3)	7.2(1.3)
.50		2.4(2.1)	2.0(1.7)
Non-bias rating			
10		9.3(1.1)	9.7(.4)
8		6.9(2.1)	7.8(1.8)
5		3.8(2.0)	3.9(2.0)
1		.9(1.6)	.7(1.3)
Ease of Administration			
10			8.7(1.9)
8			8.3(1.8)
6			7.1(1.9)
4			5.3(1.7)
Correlation with GPA			
.9			8.9(1.3)
.8			7.7(1.1)
.6			4.9(1.5)
.2			1.7(1.7)

Note: Figures in parentheses are the standard deviations.

dimensions, R = -.12. It had been hypothesized that there would be a positive correlation between the two ranking procedures when there was no cognitive overload as would be expected with subgroup 1; the correlation of .68 for subgroup 1 was significant at $p < .01$. (See Table 5.) The hypothesis that the two rankings were independent was rejected for sets described by 3 dimensions. It had also been expected that as the number of dimensions increased, the correlation between the two ranking methods would drop. In view of the non-significant rank-order correlations for the 5 dimensions and the 7 dimension rankings, the hypothesis that the rankings are independent in these two situations cannot be rejected.

Table 5

Rank-Order Correlations of Intuitive and Model Rankings

Subgroup	n	No. of Dim.	N	R
1	25	3	18	.68**
2	26	5	29	.04
3	25	7	26	-.12

**$p < .01$

Discussion

The central question of this study concerned whether or not the proposed selection model is valid. To answer this question selections made without benefit of the model, i.e. wholistic judgments, were compared to selections made with the model. The data analysis indicates that there is a significant correlation between the selection of sets of variables made without the model and with the model when the number of dimensions is 3. Thus, the model can mirror subjective judgments and evaluations.

When the number of dimensions increases to 5 and to 7, the correlation drops. It can be argued that in these cases it is not the model that breaks down, since the model avoids overload by allowing subjects to deal with one weighting or rating at a time; instead, it is the intuitive selection that falters due to the subjects' inability to handle so much information simultaneously.

The cognitive overload without the use of the model can be seen in the case of subgroup 2, for example. Without the model, subjects in subgroup 2 were presented with 10 variables that were described in terms of 5 dimensions. To choose a set of three variables, subjects had to compare each set of three with the others, keeping in mind 5 different pieces of information about each variable. In contrast, while using the model, there were 25 components that the subjects evaluated singly. In other words, the subjects either made decisions about combinations of 5-dimen-

sional stimuli (without model--intuitive ranking) or they evaluated 25 single dimensional stimuli (with model--model ranking). With the model, no matter how many dimensions are involved, the stimuli are responded to one at a time, thus preventing cognitive overload. Without the model, as the number of dimensions increases, the amount of information to be evaluated simultaneously increases, inevitably creating limitations in making satisfactory decisions.

Based on the significant correlation between the intuitive and the model rankings when there are 3 dimensions, the variable selection model appears to have validity. The selection model appears to represent adequately the evaluative behavior of the subjects. When there are 5 or 7 dimensions, the correlation drops sharply to non-significant levels. Since the efficiency of the model is not diminished by the addition of dimensions, one could argue that the lack of significantly high correlations may stem from the inefficiency of human judgments without a model.

The cognitive overload that occurs without the model when one chooses among variables described in terms of 5 or 7 dimensions may cause the subjects to make selections based on only a portion of the information provided to them. This notion requires further investigation. With the data on hand, a preliminary examination of this proposition was made.

If the reason for low correlations (in the 5 and 7 dimension subgroups) between the intuitively selected and

the model selected sets is that not all of the dimensions are used in the intuitive selection whereas all of the dimensions are used by the model, then one would expect that there would be a higher correlation when the intuitive selections are correlated with model selections which are derived from fewer than five dimensions. Thus, if subjects who must select sets described by five dimensions do tend to focus on, let us say, only three dimensions and ignore the other two dimensions, it is to be expected that the correlation of the ranking of the intuitive selections with the ranking of the model will increase when the expected utility equations used for the model ranking have only 3 rather than 5 dimensions.

To investigate this hypothesis, the expected utility equations for subgroup 2 and subgroup 3 which were originally based on 5 and 7 dimensions were reduced to equations having 3 dimensions. In other words, a reduced equation for the mean utility across dimensions was used for the model ranking.

In the 5 dimension case, instead of the equation

$$U_{set} = w_1 u_1 (m_a m_b m_c) + \ldots . w_5 u_5 (m_a m_b m_c)$$

the equation was reduced to include the three expressions having the largest dimension weights, in this case w_3, w_4, w_5. Thus, the reduced equation for all sets in subgroup 2 was

$$U_{set} = w_3 u_3 (m_a m_b m_c) + w_4 u_4 (m_a m_b m_c) + w_5 u_5 (m_a m_b m_c)$$

Similarly, in subgroup 3 (7 dimension group), the 7 expres-

sion mean utility equations were reduced to 3 expression equations. In each subgroup, the sets were re-ranked based on the new utility values derived from the reduced equations. These rankings were correlated with the intuitive rankings for each subgroup. The resulting correlations were R=.67 for subgroup 2 and R=.34 for subgroup 3. (See Table 6.)

In both subgroup 2 and subgroup 3 the correlations jumped dramatically from the levels of the original rank-order correlations. In both subgroups where there originally were non-significant correlations between the intuitive and the model rankings, there are significant correlations when the intuitive ranking is correlated with model rankings based on a reduced number of dimensions. This would tend to confirm the hypothesis that there may be some cognitive overload in dealing with more than five dimensions when one does not have a model. Furthermore, it would tend to confirm the hypothesis that in order to cope with the overload, there may be a tendency for subjects to ignore some factors when they select sets with five or seven dimensions.

Table 6

Statistics for Subgroups 2 & 3--Full & Reduced Equations

Statistic	Subgroup	
	2	3
n (Subjects)	26	25
N (Sets)	29	26
Dimensions in full equation	5	7
Dimensions in reduced equation	3	3
R (full equation)	.04	-.12
R (reduced equation)	.67**	.34*

** $p < .01$
* $p < .05$

PRACTICAL USEFULNESS OF THE MODEL

To assess how well it functions in a "real-life" situation, the model was used in connection with an evaluation of a bilingual program in a school district in Brooklyn. The program was its fourth year of operation. Each year approximately 300 students had participated in the program; most of the students were recent immigrants to the United States and spoke either Spanish or French as their first language. The program had intensive English instruction in addition to classes conducted in the native language. (All the teachers and paraprofessionals were bilingual.) Special attention was given to the reinforcement and maintenance of the child's cultural heritage through books, tapes, trips, dramatic presentations, etc.

Two evaluators and one of the program's administrators were the evaluation team (hereinafter referred to as the evaluators). All three had experience in teaching bilingual students and had been working in this program as staff members since the program's inception. Meetings were held on several days in order to proceed through the steps of the model, to obtain the data needed for the model, and to discuss the model's results and the merits of the steps and the structure of the model.

All three evaluators met with the author after they had read a description of the model. Although they were not yet convinced that this model was the ultimate solution

to the problem of selecting variables, they all agreed that the selection of a set of measures to use in an evaluation had always been a disorganized, haphazard process. In the past, they had used what they called "common sense," intuitive judgments or relied on the recommendations of other school districts. None of the evaluators was able to say that the set of measures selected without benefit of a systematic procedure had been the "best" set, only that it was a set of measures that would provide them with some useful information for their evaluation. All of them thought that it was important to seek better methods of variable selection; however, one of the evaluators, after having read the description of the model, felt that it was too "overwhelming" and were it not for having agreed to participate in this study, he would be apprehensive and unwilling to use what appeared to be a "complicated" model. The other evaluators expressed no apprehension nor negative feelings prior to using the model. All of them, however, appeared to feel that, although they did not like the lack of methods to select measures, they were not eager to do much additional work in order to select variables systematically.

The following section will describe in detail how the model was used and will report the reactions of the evaluators to the model. Each of the steps of the model was explained to the evaluators as they went through it.

Step I -- Define the variables. The evaluators had five measures under consideration. All of these measures had

been used in this district at least once and had been considered to be satisfactory measures. Other variables such as teacher reports, curriculum design, etc., were considered to be mandatory features of the evaluation report and, therefore, could not be omitted; thus, in using the model, the only variables to be decided upon were the variables that tested skills such as reading and vocabulary. The five achievement measures under consideration were: CTBS--Comprehensive Tests of Basic Skills, published by CTB/McGraw Hill. It includes testing in word recognition, comprehension, reading, mathematical concepts and applications. St. Martin, which includes reading, spelling, language usage, mathematical computation and concepts. SAT, Primary--The Stanford Achievement Test, published by Psychological Corporation, with subtests in reading skills--comprehension, words plus comprehension--auditory skills--listening comprehension, vocabulary--and mathematics--concepts, computation, applications. TOBE--Tests of Basic Experience, published by CTB/McGraw Hill; the battery consists of tests of concepts that involve no reading. The concepts are from the fields of language, mathematics, science and social studies. CREST--a test developed by the Board of Education of the City of New York to test English language usage.

Thus, the variables were:

m_1: CTBS, m_2: St. Martin's Basic Skills, m_3: SAT, Primary 1, m_4: TOBE, m_5: CREST (NYC, Bd. of Ed.)

Step II -- Enumeration of decision alternatives. There was agreement among the evaluators that all five measures could not be used for several reasons; foremost among these reasons was the over-testing of the students who had city-wide tests to take, as well. The evaluators decided that only two of the measures should be used. One evaluator felt that it would be best not to test at all, but he did agree with the others that they needed objective measures to "prove" the value of the program. They felt, however, that with all the other testing the students would undergo, it would be preferable to have as few additional tests as possible. In the past, they had used two measures. Thus, they wanted to use only two measures on this occasion. The model, however, is designed to select the best combination, or set, of measures from all possible combinations. The selection is based on the utilities and weights given by the decision makers. Therefore, if it is important to these evaluators that only two tests be chosen because of time constraints, this concern should be reflected by the values they insert into the model with respect to time. It would be expected that if the model is truly responsive to the wishes of the decision makers, then a set of two measures will be the set that the model finds to be the best.

The decision alternatives, thus, were all possible combinations of the 5 variables, i.e. sets of 1, sets of 2, ...sets of 4, a set of 5. In all, there were a total of

31 sets, or decision alternatives.

Step III -- Identify the dimensions of value. This step elicited a great deal of discussion among the evaluators. Before they began to enumerate the dimensions, they were asked if they had ever before concretized or stated what qualities they required of a measure. None of them had ever done so. They had chosen a measure because it was "known" to be good or it "sounded like it would be good" or it had been used in the district in the past. Occasionally, in the past, they had observed that a particular measure had high validity or was not expensive. However, none of them had ever specifically analyzed the attributes (dimensions) of the measures they were considering.

In deciding which dimensions to include, first the evaluators formulated a list of possible dimensions and then discussed each one, deciding which of them to keep as dimensions. The dimensions in the initial stage were:

Dim 1 -- Cost
Dim 2 -- Reliability
Dim 3 -- Correlation with Language Aptitude Battery (LAB)
Dim 4 -- Time needed to administer test
Dim 5 -- Non-bias against minorities
Dim 6 -- Correlation with teacher ratings

After this list was developed, each dimension was discussed separately. Dimension 1 (Cost) was discarded. One of the evaluators stated that he had put cost on the list of dimensions as a "reflex" but that in reality in this particular situation, the cost of a measure had no relevance. Funding had been received for testing, and

since there were sufficient monies to cover any of the possible decision alternatives, there was no reason to consider cost as a determinant or whether or not to select a particular measure for the evaluation. The other evaluators agreed with these remarks. Dimension 2 (Reliability) was also discarded. The evaluators stated that while they thought that reliability was vital, in this case all of the measures were highly reliable and were virtually indistinguishable in this respect.

The third dimension (correlation with LAB) was retained. The LAB was used as the measure to place students in the bilingual program as well as to phase the students out of the program. For this reason, the evaluators felt that it was crucial that the selected measures correlate with the LAB.

Dimension 4 (Administration time) was retained. There were great variations in time among the measures (from 25 minutes to 150 minutes), and the evaluators felt it was important to limit the number of hours of testing.

Dimension 5 was also retained. (Dim 5: minority non-bias). The evaluators, none of whom was a member of the minority groups represented in the program, were especially interested in assuring test-fairness. They said that all of the measures were free from bias, but that since a few of the measures had been created especially for minority groups, there were some small differences along this dimension that should be made note of. Dimension 6 (correlation

with teacher evaluations) was retained. The evaluators felt that the best measure of how much a student has progressed is the opinion of the teacher. Thus, how well a standard measure correlates with the teacher's rating was important to these evaluators.

In the discussion about the dimensions, there was virtual unanimity among the members of the evaluation team. When there is not unanimity about the inclusion of a particular dimension, that dimension can be retained and dissenting members can register their disagreement when the weights are given to the dimensions.

Thus, after the discussion, the list of dimensions was reduced to:

D_1: Correlation with LAB

D_2: Time needed for administration

D_3: Minority non-bias rating

D_4: Correlation with teacher evaluations

Step IV -- Assigning of weights to dimensions. Each member of the team separately gave weights to the four dimensions. The mean value of the three evaluators' weights was considered to be the assigned weight of the dimension. (As in Edwards, et al., 1975)

Table 7 summarizes the individual dimensions weights and the averaged dimension weights.

The mean weights were: 6.3 for dimension 1 (Correlation with LAB); 4.7 for dimension 2 (Administration time); 8.3 for dimension 3 (Minority non-bias rating); 7.7 for dimen-

Table 7

Dimension Weights -- Field Study

Dimension	Subject			Mean	S.D.
	S_A	S_{E_1}	S_{E_2}		
1. Correlation with LAB	6	6	7	6.3	.58
2. Time	7	4	3	4.7	2.1
3. Non-bias rating	7	9	9	8.3	1.2
4. Correlation with teacher ratings	7	7	9	7.7	1.2

sion 4 (Correlation with teacher ratings.) The greatest unanimity among the evaluators in assigning weights was found with respect to dimension 1, where the standard deviation among the individual weights was .58; the largest range of weights among the individuals was with respect to dimension 2, administration time, with a standard deviation of 2.1.

All of the evaluators found it to be simple to assign weights, although they felt a little reluctant about quantifying their feelings into an assignment of a specific value. They felt that they could say that, for instance, dimension x was "somewhat more important" than dimension y, but translating "somewhat more important" into a number made them slightly uneasy. However, all of them accomplished the task quickly, and when asked if the numbers they had given as weights adequately represented their subjective judgments, they all replied that they did.

Step V -- Determination of the composite values. The composite values for each set are derived from the values of the individual measures. (See Table 8 for the values of the measures on each dimension.) The composite values are determined according to the methods set forth in the description of the model. In this case, for each of the 31 sets of variables there were two composite correlations: 1) correlation with LAB, 2) correlation with teacher evaluations. According to the model, the proper statistic to determine composite validity is the multiple

Table 8

Values of Individual Measures--Field Study

Measure	Dimension			
	Correlation with LAB	Time	Non-bias rating (0-100)	Correlation with teacher rating
1. CTBS	.24	150 min.	90	.22
2. St. Martins	.20	30 min.	94	.22
3. SAT	.21	120 min.	91	.21
4. TOBE	.19	25 min.	91	.25
5. CREST	.21	60 min.	95	.20

correlation of the members of the set with the criterion variable. The author calculated these composite correlations based on data supplied by the evaluators. To obtain the multiple correlation of each set with the LAB, test scores from previous years were used; similarly, the multiple correlation coefficient for each set with respect to teacher evaluations was calculated by correlating prior test scores with prior teacher ratings. (These ratings were on a scale of 0-10.) Next, the composite administration times were determined by adding the administration times of the members of each set. The last of the composite entities was the composite non-bias rating. The model does not specify a procedure for determining this composite, so a method was improvised. The evaluators were asked to rate the 5 individual measures as to non-bias toward minorities on a scale from 0-100. (See Appendix B.) A rating of 100 represented total absence of bias toward minorities and a rating of 0 represented total bias toward minorities in every aspect of the measure. The ratings for each measure given by each of the evaluators were averaged to obtain the non-bias rating for the individual measure. To ascertain the composite non-bias rating for each set of measures, the non-bias ratings for each measure in the particular set were averaged and the resulting mean value was designated as the composite non-bias rating of the set.

Step VI -- Construction of the utility functions.

The evaluation team found the utility functions of the

sets of measures with respect to the dimensions by using the standard gamble technique as described by the model. Anchor points for each of the utility functions were the highest and lowest possible outcomes on each dimension; in this situation, the outcomes are the composite values for each set. The evaluators were extremely resistant to the standard gamble technique; they said that it was "almost impossible" to conceive of outcomes in the manner prescribed by the technique. One complained, "I can't do this." They felt uneasy, uncomfortable with what they were doing, as well as feeling unsure of their judgments. The resulting utility functions can be found in Table 9. The three evaluators assigned utility functions that satisfied the axiom of transitivity (i.e. if one prefers A_1 to A_2 and A_2 to A_3, then A_1 should be preferred over A_3.) The three evaluators had a high degree of similarity in the utility functions they designated; the standard deviations for all except two of the utility values were less than or equal to 1.0. The mean utility values were used in the ensuing step.

Step VII -- Calculation of the utility values for each variable set. This step requires calculation of the expected mean utility of a set based on the equation:

$$U_{set} = w_{D_1} u_{set \text{ on } D_1} + \ldots + w_{D_N} u_{set \text{ on } D_N}$$

where D=dimension, N=number of dimensions, u=utility, w=weight. Using the dimension weights from Table 7, the w_{D_I} were

Table 9

Utility Functions of the Dimensions--Field Study*

Dimension level	Evaluators S_A	S_{E_1}	S_{E_2}	Mean	S.D.
Correlation with LAB					
.2	2	3	2	2.3	.57
.4	6	7	5	6	1.0
.5	8	7.5	6	7.2	1.0
.6	8.5	8	7.5	8	.5
.75	9	8.5	9	8.8	.29
Time					
50 minutes	9	9.5	9	9.2	.29
90	8	9	8	8.3	.58
120	5	7	7	6.3	1.2
180	2	3	2	2.3	.58
290	0.5	0.1	1	.5	.45
Non-bias rating					
50	2	1	0.5	1.2	.76
80	4	2	2	2.7	1.2
90	9	8	8	8.3	.58
92	9.2	8.5	9	8.9	.36
95	9.8	9	9.5	9.4	.4
Correlation with teacher rating					
.2	1	3	2	2	1.0
.4	5	6	6	5.7	.58
.5	7	7.5	6.5	7	.5
.6	7.5	8	9	8.2	.76
.75	8.3	9	9.5	9	.5

*Note: Anchor points are the highest and lowest levels. The levels included in this table are intermediate values to be used in plotting the utility functions.

obtained. Using the composite values for each set which were found in step V and using the utility values of Table 9, all of the $u_{set\ on\ D_I}$ were obtained for each set. These values were inserted in the equation for the expected mean utility of a set. The utility values for each set on each dimension can be found in Table 10. As can be seen in Table 10, the sets with the greatest utility on dimension 1 are sets $m_1m_2m_3m_4$ and $m_1m_2m_3m_4m_5$. On dimension 2, set m_4 has the greatest value. On dimension 3, set $m_2m_3m_5$ and set $m_2m_4m_5$ have the largest utility values. On dimension 4, set $m_1m_3m_4m_5$ and set $m_1m_2m_3m_4m_5$ have the largest utility values. The fact that different sets have the greatest utilities on each dimension further underlines the difficulty in choosing among variables without a model.

When the dimension weights (Table 7) and the utility values for each set on the dimensions (Table 10) were placed in the equation for expected mean utility of a set, the resulting values were used to order all the sets. The ranking of the sets can be found in Table 11.

Step VIII -- Decision-making.

According to the model, the set with the greatest utility is ranked the highest and should be the set chosen by the decision makers. In this case, the set with the greatest expected utility was set m_4m_5 (TOBE & CREST). Initially, the evaluators had stated their preference for using only two of the measures. Although the procedures of the model deal with all possible combinations, not just

Table 10

Utility Values for Each Set--Field Study

Set	Dimension 1 Correl w. LAB	2 Time	3 Non-bias	4 Correl w. Teach. eval.
m_1	7.9	4.3	8.3	0
m_2	.65	9.9	9.2	0.2
m_3	8.74	6.3	8.6	2
m_4	8.22	10.0	8.6	0
m_5	4.34	9.0	9.4	1.0
m_{12}	7.9	2.3	8.9	0.4
m_{13}	8.8	.95	8.45	3.48
m_{14}	8.74	2.6	8.45	0
m_{15}	7.9	1.85	8.98	1.4
m_{23}	8.74	4.3	8.98	2
m_{24}	8.28	9.1	8.98	0.4
m_{25}	4.34	8.3	9.32	1.2
m_{34}	9.1	4.63	8.6	2.9
m_{35}	8.74	2.3	9.07	2.18
m_{45}	8.28	8.4	9.07	1.0
m_{123}	8.8	0.0	8.78	3.67
m_{124}	8.74	1.9	8.78	0.2
m_{125}	7.9	1.18	9.07	1.6
m_{134}	9.2	0.16	8.48	4.04
m_{135}	8.8	0	8.9	3.85
m_{145}	8.74	1.29	8.9	1.2
m_{234}	9.2	2.63	8.9	3.11
m_{235}	8.74	1.85	9.12	2.18
m_{245}	8.28	6.63	9.12	1.4
m_{345}	9.1	1.9	8.95	3.11
m_{1234}	9.3	0	8.75	4.22
m_{1235}	8.8	0	8.98	3.85
m_{1245}	8.74	.99	8.98	1.6
m_{1345}	9.2	0	8.84	4.41
m_{2345}	9.2	.79	9.04	3.3
m_{12345}	9.3	0	8.93	4.41

Table 11

Ranking of Sets according to Utility Value across Dimensions

Set	Rank	Set	Rank
m_4m_5	1	m_3m_5	17
m_3m_4	2	m_1m_3	18
m_2m_4	3	$m_1m_2m_3$	19
m_3	4	$m_2m_3m_5$	20
m_4	5	m_5	21
$m_2m_4m_5$	6	m_2m_5	22
$m_2m_3m_4$	7	$m_1m_2m_4m_5$	23
$m_1m_2m_3m_4m_5$	8	$m_1m_4m_5$	24
$m_1m_3m_4m_5$	9	m_1m_5	25
m_2m_3	10	$m_1m_2m_5$	26
$m_3m_4m_5$	11	m_1	27
$m_1m_2m_3m_4$	12	$m_1m_2m_4$	28
$m_2m_3m_4m_5$	13	m_1m_2	29
$m_1m_3m_4$	14	m_1m_4	30
$m_1m_2m_3m_5$	15	m_2	31
$m_1m_3m_5$	16		

combinations of two variables, nevertheless a set of two measures, indeed, was the first ranking set. In fact, the sets ranked second and third were also sets consisting of two measures. Thus, it could be said that the weight given to the dimension of Time and the utility function of Time both had an effect on the model's choice, and the model was sensitive to the wishes of the evaluators.

If the utility values for each set on each dimension (Table 10) are examined in order to explain the selection by the model of set m_4m_5, one can see that the set m_4m_5 did not have the greatest utility among the sets on any single dimension. However, this set (m_4m_5) had a utility of 8 or more on three of the four dimensions. Two other sets (m_4 & m_2m_4) also had utilities of 8 or more on three dimensions, but their utilities on the fourth dimension were lower than that of set m_4m_5.

A summary of the evaluators' comments regarding the use of this model in the selection of measures: With respect to the ease or difficulty of the procedures, all of the steps of the model except for Step VI, construction of the utility functions, were relatively easy. The evaluators felt that Step III--identify the dimensions of value--was very helpful, making them aware of what they were seeking in a measure, and that the discussion assisted them in several ways. They better understood their own priorities and learned about the thinking of their colleagues. It must be reiterated here that the author did the actual cal-

culations necessary for Step V--determination of composite entities--and for Step VII--calculation of overall utilities. When the evaluators were shown what had been done in order to carry out these two steps, only one of them stated that he would be willing to do the work involved. The other two claimed that the work of these steps was too tedious for them, and if they were to use the model, they would, perhaps, hire someone for a day to do these calculations. In this particular trial use of the model, approximately 2½ hours were spent in going through the process: One hour was spent in introducing the model to the evaluators, eliciting the decision alternatives, the dimensions, and the dimension weights; at another session, each evaluator spent ½ hour constructing utility functions; at the final session all three evaluators spent almost one hour discussing the results, etc.

With respect to improving the model, the evaluators said that they would prefer constructing the utility functions by a method other than the standard gamble technique.

CONCLUSIONS

Based on the practical use of the model and the validation of the model, it appears that the proposed variable selection model can be used in some settings.

One change in the model that should be considered in order to enhance its usefulness is with respect to the utility functions. Although in the description of the model the standard gamble technique was elucidated, it was also noted that this technique is only one technique among many. In light of the difficulties the evaluators had using this method, other users of the model might want to utilize another technique to construct utility functions. Two possible alternative methods might be: Rating. For each dimension a scale is constructed with the least preferred outcome on the dimension given a scalar value of 0 and the most preferred value given a value of 1.0. The scale is divided into equal intervals with the value .5 representing a preference midway between the least preferred and most preferred points. Several levels of the dimension are then rated along the scale with regard to one's relative preference for this level. If, for example, cost is the dimension, one might set a cost of $1.00 as the most preferred and the cost of $25.00 as the least preferred outcome. The user might feel that an intermediate value such as $15.00 has very little preference and, thus, might assign it a value of .1 on the rating scale; if one feels the

preference value of $15.00, on the other hand, is slightly lower than average, then one might assign it to .4 on the scale. These preference ratings are used as utility values.

Direct Midpoint. In this method, one states the least preferred level, z, and the most preferred level, x, of dimension i. Then one designates which level of the dimension (level y) is midway between x_i and z_i in utility value. Then utility midpoints between x_i & y_i and y_i & z_i are designated, and so on, until enough points are generated to obtain a curve. A variant of this method, as cited by Galanter (1962), is to select a level of low preference, for example receiving a gift worth $1.00, and to set the utility of $1.00 = 1. If $1.00 = z, one would estimate how expensive a gift, x, ($x > z$) would be twice as pleasing. This response, x, becomes the new z and one estimates again which x would be twice as pleasing as z. If $u(x) - u(0) = 2[u(z) - u(0)]$ and if $u(0) = 0$ and the utility of $1.00 = 1, then the utility of $x_1=2$, the utility of $x_2=4$, etc.

A feature of the model whose importance was demonstrated is the model's integration of the views of dissenting members of a decision team. Since each member of the team can give his or her own opinion regarding dimension weights, utilities, etc., it is possible for the judgments of all of the members to be included in the decision making process. In some situations, it might be decided to include a dimension even though only one member considered this dimension to be worthy of inclusion. The other evaluators could ex-

press their disagreement by assigning appropriate dimension weights, etc. In this way, dissenting members may have an impact on decision making and not feel totally overwhelmed by the majority. It should be noted, however, that averaging may not properly represent the weights or the utility functions of the decision makers when there is a divergence in their judgments. In such an adversarial situation, however, the model still is useful in that each member of the decision team publicly reveals the weights and utilities. In this way, the antagonists can see where there is disagreement and can see where their priorities differ.

One aspect of the model that may require ingenuity on the part of the user is the determination of composite entities other than the four composites described in the model --composite cost, composite time, composite reliability, composite validity. If the decision makers list dimensions besides these four, they will have to innovate the method allowing them to determine the composite values. In some cases this may not be difficult. For instance, in the use of the model by the three evaluators, they listed as a dimension, "non-bias against minorities," which was not a dimension that the model had anticipated and described. In this case, the composite non-bias rating was ascertained by calculating the average non-bias rating of the variables of the set. In other situations, there may be dimensions that require greater inventiveness.

A useful area for future MAUT research is the consid-

eration of sampling problems. In a given evaluation situation, a team of evaluators may be selected from some larger group of potential evaluators. For example, in an educational setting, a particular principal, a particular administrator, etc. are chosen from among many principals and many administrators. It would be interesting to study the sampling properties of the MAUT model with respect to the sampling of evaluators and to investigate the reliability of the judgments.

In conclusion, the proposed model can be used to select a single set of variables from among other alternative sets. Although the model has been validated and field tested with data based on program evaluations, there is no reason to reject the possibility of applying this model to other variable selection situations in the behavioral sciences, such as selection of variables in experimental studies. The model may not be useful in all situations, but it has been shown to be of assistance in some situations.

USER'S MANUAL

In order to facilitate the use of the variable selection model, a step-by-step user's manual is provided. Accompanied by illustrations from a hypothetical example, the user is led through the procedures of the model.

Step A

List all the variables and label as $v_1 \ldots . v_n$.

Example:

v_1 = standardized test a
v_2 = standardized test b
v_3 = classroom measure a
v_4 = classroom measure b
v_5 = teacher ratings

Step B

List decision alternatives, i.e. list all combinations (or sets) of the variables listed in step A, i.e. sets of 1 variable, sets of 2,....sets of n-1, sets of n variables.

Example:

v_1	v_1v_2	$v_1v_2v_3$	$v_1v_2v_3v_4$	$v_1v_2v_3v_4v_5$
v_2	v_1v_3	$v_1v_2v_4$	$v_1v_2v_3v_5$	
.	.	.	.	
.	.	.	.	
.	.	.	.	
.	.	.	.	
.	.	.	.	
v_5	v_4v_5	$v_3v_4v_5$	$v_2v_3v_4v_5$	

Step C

List the dimensions of the variables. The dimensions are the attributes or characteristics of the variables that help you decide among them. Label the dimensions $D_1 \ldots . D_N$.

Example:

D_1=Cost, D_2=Time, D_3=Reliability, D_4=Validity.

Step D

For each individual variable from step A, list the value of it on each dimension listed in step C. This information can be obtained from test manufacturers, from pilot testing, from previous results, etc.

Example:

	Dimensions			
Variables	Cost	Time	Reliability	Validity
v_1	$5.00	30min.	.80	.78
v_2	7.50	60	.90	.77
v_3	10.00	60	.80	.89
v_4	2.50	30	.70	.89
v_5	1.00	30	.60	.78

Step E

Using the values listed in step D, determine the composite values for each set listed in step B.

a) Composite Cost. To determine the composite cost for each set, add the costs of the individual variables in the set.

Example:

Cost, v_1v_2 = \$12.50 $v_1v_2v_3$ = \$22.50
v_1v_3 = 15.00 $v_1v_2v_4$ = 15.00

b) Composite Time. To determine the composite time for each set, add the times of the individual variables in the set.

Example:

Time, v_1v_2 = 90 minutes $v_1v_2v_3$ = 150 minutes
v_1v_3 = 90 $v_1v_2v_4$ = 120
etc.

c) Composite reliability. To determine the reliability of each set, the following data are needed:

The reliability of each indidivual variable, r_{ii}
The variance of each variable, s_i^2
The variance of each set, s_c^2

This information is inserted into the equation:

$$\text{Reliability}_{set} = 1 - (\Sigma s_i^2 - \Sigma r_{ii} s_i^2) / s_c^2$$

Example:

For set $v_1v_2v_3$, with r_{ii}=.8, r_{22}=.9, r_{33}=.8
$s_i^2 = 53$, $s_2^2 = 91$, $s_3^2 = 80$
$s_{123}^2 = 541$

$$\text{Reliability}_{123} = 1 - \frac{((53+91+80)-(53(.8)+91(.9)+80(.8)))}{541}$$
$$= .93$$

d) Composite validity. To determine the composite validity, one needs the following data: the correlation of each variable with the criterion, (from step D), the intercorrelation of the individual variables. With this information one calculates for all sets listed in step B

the multiple correlation of the criterion with the variables of the set.

Example:

With r_{12}=.81, r_{13}= .68, r_{23}=.68

r_{1y}=.78, r_{2y}= .77, r_{3y}=.89 (from step D)

Then R_{y12}=.81 R_{y13}=.92 R_{y123}=.93

Step F

Taking the list of dimensions from step C, place weights (w_j) on each of the dimensions. These weights indicate the relative importance to you of the dimension in selecting a set of variables in this circumstance. One method of assigning weights could be to give the least important dimension a weight of one (1) and then to ask yourself how much more important the other dimensions are. If a dimension is 3 times as important as the least important one, it would be given a weight of three (3). If there are N dimensions, there will be weights $w_1, w_2 \ldots w_N$.

Example:

The users felt that D_1 (Cost) and D_2 (Time) were equally important and, at the same time, they were less important than D_3 and D_4. So, D_1 and D_2 were both given weights of 1. D_3 (Reliability) was felt to be 3 times as important as D_1 and D_2, so it was given a weight of 3. D_4 was felt to be twice as important as D_1 and D_2, so it was given a weight of 2. Thus, the dimension weights are:

$w_1=1$, $w_2=1$, $w_3=3$, $w_4=2$.

Step G

Construct the utility functions for each dimension. Several methods for finding utilities are described in the conclusion section of the dissertation.

Example:

Using rating scales, the following utility functions were established.

Cost of set	$3.00	5.00	10.00	21.00	30.00
Util. (Cost)	9.5	9	5	3	.05
Time of set	30min.	60	90	120	180
U (Time)	9	7	4	2	.5
Reliability	.99	.95	.9	.8	.5
U (Reliabil)	9.9	9	8	6	1.5
Validity	.98	.9	.6	.5	.2
U (Validity)	9.9	9	7	6	2

Step H

Find the utility value of each set on each dimension. The utility of set (a,b) on dimension j is determined by taking the utility value from step G that corresponds to the value of set (a,b) on dimension j. If the value of the set on a particular dimension should fall between the listed values, the utility can be determined by interpolation.

Example:

set $_{123}$ with a cost of $22.50 has a utility of 2.5 on D_1
with a time of 150 min. has a utility of 1.1 on D_2
with a reliability of .93 has a util. of 8.6 on D_3
with a validity of .93 has a utility of 9.3 on D_4

Step I

Calculate the utility values of each set across dimensions using the dimension weights (w_j) from step F and the utility values of step H. The utilities of the sets are found with the equation:

$$U_{set} = \sum_j w_j u_{j(set)} \quad \text{or}$$

$$U_{set} = (\text{weight}_{Dim1})(\text{utility of set on } D_1) + \ldots.$$
$$(\text{weight}_{DimN})\ (\text{utility of set on } D_N)$$

Example:

$$\text{Utility of set}_{123} = 1(2.5) + 1(1.1) + 3\ (8.6) + 2(9.3)$$
$$= 48.1$$

Step J Decision making. Choose the set that in step I had the highest utility value; this is the best decision alternative.

Appendix A

Sample Chart -- Validation Study, Subgroup 3

Measure	Dimension 1	2	3	4	5	6	7
#1	30	$6.00	.75	.91	9	9	.75
#2	60	5.00	.70	.90	8	7	.60
#3	15	1.00	.65	.85	8	10	.50
#4	15	3.00	.65	.85	8	10	.50
#5	30	3.00	.70	.90	8	9	.60
#6	30	5.00	.60	.95	7	7	.60
#7	60	4.00	.85	.92	7	7	.80
#8	30	1.00	.70	.90	8	9	.65
#9	60	3.00	.75	.95	10	8	.70
#10	30	2.00	.75	.85	9	9	.75

Note: Dimension 1= Time, in minutes, needed for test administration.
Dimension 2= Cost in dollars.
Dimension 3= Correlation with classroom tests.
Dimension 4= Reliability
Dimension 5= Non-bias rating with 10 equal to least bias.
Dimension 6= Ease of administration, with 10 = easiest to administer.
Dimension 7= Correlation with Grade Point Average.

Appendix B

Minority Non-bias Ratings for Measures -- Field Study

Measure	Evaluator S_A	Evaluator S_{E_1}	Evaluator S_{E_2}	Mean	S.D.
CTBS	90	90	90	90	0.0
St. Martins	95	92	95	94	1.7
SAT	92	90	92	91	1.15
TOBE	90	90	92	91	1.15
CREST	95	95	95	95	0.0

BIBLIOGRAPHY

Airasian, P.W. Designing summative evaluation studies at the local level. In W.J. Popham (Ed.), Evaluation in Education. Berkeley, California: McCutcheon, 1974, 147-199.

Anscombe, F.J. Sequential medical trials. Journal of the American Statistical Association, 1964, 58, 365-383.

Bloom, B.S., Hastings, J.T. & Madaus, G.F. Handbook on Formative and Summative Evaluation of Student Learning. New York: McGraw-Hill, 1971.

Brown, R.V., Kahr, A.S. & Peterson, C. Decision Analysis for the Manager. New York: Holt, Rinehart and Winston, 1974.

Chase, C.I. Measurement for Educational Evaluation. Reading, Massachusetts: Addison-Wesley, 1974.

Cooley, W.W. & Lohnes, P.R. Evaluation Research in Education. New York: Irvington, 1976.

Coulson, J.E. National evaluation of the Emergency Aid Act (ESEA): A review of methodological issues. Journal of Educational Statistics, 1978, 3, 1-60.

Cronbach, L.J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297-334.

Cronbach, L.J. & Gleser, G.C. Psychological Tests and Personnel Decisions. Urbana: University of Illinois Press, 1965.

Dawes, R. & Corrigan, B. Linear models in decision making. Psychological Bulletin, 1974, 81, 97-106.

Eckenrode, R.T. Weighting multiple criteria. Management Science, 1965, 12, 180+.

Edwards, W., Guttentag, M. & Snapper, K. A decision-theoretic approach to evaluation research. In M. Guttentag & E.L. Struening (Eds.), Handbook of Evaluation Research. Beverly Hills: Sage Publications, 1975, Vol. I, 139+.

Evans, J.W. Evaluating educational programs--Are we getting anywhere? Educational Researcher, 1974, 3(8), 7-12.

Fishburn, P.C. Methods for estimating additive utilities. Management Science, 1967, 13, 435-453.

Fitzgibbon, C.T. & Morris, L.L. How to Design a Program Evaluation. Beverly Hills: Sage Publications, 1978.

Galanter, E. The direct measurement of utility and subjective probability. American Journal of Psychology, 1962, 75, 208-220.

Gibson, K.D. Findings. Evaluation, 1976, 3, 27-28.

Greene, H.A., Jorgensen, A.N. & Gerberich, J.R. Measurement and Evaluation in the Secondary School. New York: Longmans, Green, 1954.

Guilford, J.P. Psychometric Methods. New York: McGraw-Hill, 1954.

Guttentag, M. Subjectivity and its use in evaluation research. Evaluation, 1973, 1, 60-65.

Guttentag, M. & Snapper, K. Plans, evaluations and decisions. Evaluation, 1974, 2, 58+.

Heinze, D.C. Statistical Decision Analysis for Management. Columbus, Ohio: Grid, 1973.

Hoepfl, R.T. & Huber, G.P. A study of self-explicated utility models. Behavioral Science, 1970, 15, 408-414.

House, E.R. Evaluation with Validity. Beverly Hills: Sage Publications, 1980.

Huber, G.P. Daneshgar, R. & Ford, D.L. An empirical comparison of five utility models for predicting job preferences. Organizational Behavior and Human Performance, 1971, 6, 267-282.

Isaac, S. Handbook in Research and Evaluation. San Diego: EdITS Publishers, 1971.

Keeney, R.L. Utility functions for multiattribute consequences. Management Science, 1972, 18, 276-287.

Keeney, R.L. & Raiffa, H. Decisions with Multiple Objectives: Preferences and Value Tradeoffs. New York: John Wiley & Sons, 1976.

Kozielecki, J. Psychological characteristics of probabilistic inference. Acta Psychologica, 1970, 34, 480-488.

Lindley, D.V. Making Decisions. London: John Wiley, 1971.

Lindvall, C.M. & Cox, R.C. Evaluation as a Tool in Curriculum Development: The IPI Evaluation Program. Chicago: Rand McNally & Co., 1970.

Lord, F.M. & Novick, M.R. Statistical Theories of Mental Test Scores. Reading, Mass. : Addison-Wesley, 1974.

Metfessel, N.S. & Michael, W.B. A paradigm involving multiple criterion measures for the effectiveness of school programs. Educational and Psychological Measurement, 1967, 27, 931-943.

Novick, M.R. Bayesian Methods in Psychological Testing. Princeton: ETS Research Bulletin, RB69-31, 1969.

Novick, M.R. & Lewis, C. Coefficient alpha and the reliability of composite measurements. Psychometrika, 1967, 32, 1-13.

Nunnally, J.C. Educational Measurement and Evaluation. New York: McGraw-Hill, 1972.

Nunnally, J.C. Psychometric Theory. New York: McGraw-Hill, 1967.

Phillips, L.D. Bayesian Statistics for Social Scientists. London: Thomas Nelson & Sons, 1973.

Popham, W.J. An Evaluation Guidebook. Los Angeles: The Instructional Objectives Exchange, 1972.

Popham, W.J. (Ed.) Evaluation in Education. Berkeley: McCutcheon, 1974.

Raiffa, H. Decision Analysis: Introductory Lectures on Choices under Uncertainty. Reading, Mass.: Addison-Wesley, 1968.

Rajaratnam, N., Cronbach, L.J. & Gleser, G.C. Generalizability of stratified parallel tests. Psychometrika, 1965, 30, 39-55.

Raju, N.S. A generalization of coefficient alpha. Psychometrika, 1977, 42, 549-565.

Rogosa, D. Politics, process and pyramids. Journal of Educational Statistics, 1978, 3, 79-86.

Rossi, P.H., Freeman, H.G. & Wright, S.R. Evaluation: A Systematic Approach. Beverly Hills: Sage Publications, 1979.

Rutman, L.R. Planning Useful Evaluations. Beverly Hills: Sage Publications, 1980.

Salasin, S. Integrating evaluation findings. Evaluation, 1977, 4, 178-188.

Scriven, M. Evaluation perspectives and procedures. In W.J. Popham (Ed.), Evaluation in Education. Berkeley: McCutcheon, 1974, 3-93.

Shepard, R.N. On subjectively optimum selections among multivariate alternatives. In M.W. Shelley & G.L. Bryan (Eds.), Human Judgment and Optimality. New York: John Wiley, 1964.

Slovic, P. & Lichtenstein, S. Comparison of Bayesian and regression approaches to the study of information processing in judgment. Organizational Behavior and Human Performance, 1971, 6, 649-744.

Solomon, H. (Ed.) Studies in Item Analysis and Prediction. Stanford: Stanford University Press, 1961.

Stake, R.E. Evaluating Educational Programmes. Washington, D.C.: Organization for Economic Cooperation and Development, 1976.

Thorndike, R.L. & Hagen, E. Measurement and Evaluation in Psychology and Education. New York: John Wiley & Sons, 1961.

Tryon, R.C. Reliability and behavior domain validity: Reformulation and historical critique. Psychological Bulletin, 1957, 54, 229-249.

von Winterfeldt, D. & Fischer, G.W. Multivariate utility theory: Models and assessment procedures. In D. Wendt & C. Vlek (Eds.) Utility, Probability and Human Decision Making. Dordrecht, Holland: D. Reidel, 1975, 47-85.

Winkler, R.L. Introduction to Bayesian Inference and Decision. New York: Holt, Rinehart and Winston, 1972.

Wolf, R.M. Data analysis and reporting in evaluation. In W.J. Popham (Ed.) Evaluation in Education. Berkeley: McCutcheon, 1974, 203-242.

Worthen, B.R. & Sanders, J.R. Educational Evaluation: Theory and Practice. Worthington, Ohio: Charles A. Jones, 1973.

Yntema, D.B. & Klem, L. Telling a computer how to evaluate multidimensional situations. IEEE Transactions on Human Factors in Electronics, 1965, HFE-6, 3-13.