# ASSESSING TRADEOFFS IN PLAN EVALUATION: A BEHAVIORAL APPROACH

by

J.J. Louviere

and

Jay Baker

January 1979

Technical Report 108

Institute of Urban and Regional Research
University of Iowa

Jordan J. Louviere is an Assistant Professor of Business Administration and Geographpy and a Research Associate at the Institute of Urban and Regional Research at the University of Iowa in Iowa City, Iowa

Jay Baker is an Assistant Professor in the Department of Geography and a Research Associate at the Florida Resources and Environmental Center at Florida State University in Tallahassee, Florida.

# ASSESSING TRADEOFFS IN PLAN EVALUATION: A BEHAVIORAL APPROACH

## Abstract

Increasingly popular matrix-type plan evaluation procedures are often applied without sufficient attention being given to derivation of weights placed on goals and to the form of the "scoring" function used in integrating the various effects of the plans. Two empirical studies indicate the sensitivity of the procedures to choice of weight derivation technique and challenge the use of a linear scoring function. A detailed discussion of the theoretical motivation for employing a functional measurement approach in weight assessment and scoring function specification is presented and then illustrated in a third empirical comparison analysis. The theory-based procedure is shown to be superior on several grounds.

# INTRODUCTION

A common problem faced by planners in both public and private agencies is the evaluation of alternative plans and the selection of some one or more which "best" satisfy some stated objective function. A variety of procedures for solving this problem have been used over the years, ranging from cost-benefit analysis to informal consensus among those individuals responsible for making the decision.

The attraction of benefit-cost studies has been that measurement of all effects is in the same metric (thereby facilitating tradeoffs) and is firmly grounded in theory. (See Prest and Turvey, 1965 for a review of the theory, mechanics, issues, and applications of cost-benefit analysis.) The technique has become increasingly unpopular in recent years due to two principal criticisms: (1) It is maintained that certain aspects of alternatives cannot be measured monetarily. Although we do not endorse this objection (see Baker, 1975 for a discussion of the issue), it is sufficiently widely held to occasional distrust and suspicion of results. Perhaps it is more accurate to state that monetary measurement of certain entities is very difficult and can be very demanding of time and expertise. (2) The theory itself is not sensitive to distributional effects; that is the question of "who benefits and who pays" is not adequately treated, particularly at the individual or micro level.

To attempt a rigorous evaluation of planning alternatives and yet overcome the criticisms of benefit-cost analysis, a number of procedures commonly termed "multi-objective" or "multi-attribute" approaches have been proposed and are gaining in favor. (See Lichfield, Kettle and Whitbread, 1975; Nijkamp, 1975; and Keeney and Raffia, 1976 for reviews and discussions.) These procedures have two aspects in common: (1) establishing the relative importance of the various goals, and (2) integration of the effects of the plan on the goals into what is termed a "scoring function." If the weights are not accurate or if the function is incorrect, the results of the evaluation must be questioned. Unfortunately, practitioners have been relatively uncritical of these aspects in many applications of the multi-criteria procedures. For example, Lichfield, Kettle, and Whitbread (1975) have noted the lack of rigor and theoretical motivation particularly in the assignment of relative weights in many applications. It seems critical to application to be able to test whether weights or functions are theoretically adequate, and if not to provide an adequate theory for their treatment.

Hence, it is these two steps--assessment of trade-off values and weights and derivation of scoring functions--which this paper addresses both theoretically and empirically. The discussion will concentrate on the goals-achievement approach (Hill, 1968) because it is exemplary of a number of cross-tabular or matrix-based weighting and scoring procedures. While there are a number of more sophisticated refinements of Hill's approach, his is by far the best known and most commonly applied. Thus, by shedding light on the weights/function issue for this approach, it is possible to generalize to a wide range of related approaches.

Hill's procedure requires specification of goals that the "best" plan should be designed to achieve. For example, a coastal zone management plan might aspire to achieve the following goals: (A) minimize destruction of wetland ecosystems; (B) minimize damages from coastal storms; (C) minimize direct pecuniary costs of the plan to the public sector; and so forth.

Goals specification thus represents the initial input into the process. The next procedure step requires an estimate of the relative importance of each goal.

Measurement of "relative importance" requires subjective assessment on the part of some set of actors in the process, usually the planners themselves or outside "experts." Frequently a scale ranging from one to ten points is employed to elicit judgments of "relative importance" from the actors. That is, the actors are required to estimate "how important" each goal should be relative to every other goal. Ratios of such judgments are typically considered meaningful: for example, if goal A receives an '8' and goal B a '4', it would usually be assumed that A was twice as important as B. It must be noted at this point that the weight assignment process intimately depends upon a later step in the process--the selection and development of a "scoring function" or combination rule.

Concurrent with the assignment of "weights" is the assignment of levels or scale values for "how well" a given plan (or alternative) achieves each specific goal. Thus, subjective scaling estimates for "how well" each plan achieves each goal is a necessary input. Various scales are employed in this process, but the essential intent is always the same: for each individual actor or a group of actors measure the achievement levels of plans through individual judgments.

The final step is to develop a "scoring function" or combination rule which combines the weights and levels of achievement into an overall or composite "weighted achievement" score. The "best" plan is assumed to be that plan which achieves the highest composite score. Several points may be raised regarding the validity of this process:

1. If either the weights or the achievement levels or the scoring function is wrong, the entire process is wrong.
2. There is only one possible scoring function which is theoretically consistent with this procedure--a linear, additive (or averaging) composite rule. The empirical validity of this assumption, however, may be tested; and hence, it may be rejected if false.
3. If addition is not the correct empirical composition rule for this process, then current procedures make absolutely no sense whatsoever.

The remainder of this paper addresses these three issues in particular through a blend of empirical and theoretical work. It is first demonstrated that weight values cannot be empirically defended as adequate using frequently applied methods. It is next demonstrated that the composition function is not linear and additive. Following these empirical results a new procedure for the assessment of goals achievement is developed which is firmly rooted in theory and applied in a second empirical case study. Finally, the generalizability of the method to various issues of interest in goals assessment, impact assessment and effectiveness assessment is explored.

## EMPIRICAL ANALYSIS OF WEIGHTS AND SCORING FUNCTIONS

In this section four different procedures for assessing relative importance or "weight" are investigated and empirically compared. Similarly, the adequacy of the linear scoring rule is examined within the context of one of the procedures for weight assessment. The following procedures are the object of analysis:

1. Informal Consensus. In this procedure the evaluator exercises his own judgment based upon his expertise, knowledge of the issues, impressions of community values, and other inputs. Most often, a team of evaluators is responsible for decision-making, and they, as a group will arrive at some consensus weighting through informal discussion. This is probably the most commonly-practiced means of assigning weights in a goals achievement application.

2. Individual Polling. Although infrequently used, in this procedure a number of individuals are asked to independently assign weight values to the goals in question, using some scale, such as the one described earlier. These individuals might be experts, planners, interest group members, or the public at large. This technique requires some means for determining group or aggregate weight for each goal; frequently some statistic representative of central tendency is employed such as the mean, median or mode.

3. Delphi Paneling. This approach is similar to a polling procedure (see Dalkey, 1968), but is usually applied only to so-called "experts" and involves several rounds of evaluation accompanied by feedback. Individuals who are polled are asked to provide a numercial judgment or prediction about some event or set of events. Applied to the present context, individuals would be asked to assign relative weights to the various goals. Feedback would then be provided in the form of descriptions of the responses of other panel members. Usually, some measure of central tendency, such as the mean, is provided to the panelists, and they are asked to re-evaluate the alternatives (to reassign relative weights). Rounds of assessment are continued until the aggregate judgments stabilize. Group weights are usually

derived by taking the mean of the individual judgments. To our knowledge, this procedure has not been directly applied to the estimation of relative weights in plan evaluation, but its extension seems clear.

4. Policy Capturing. This approach is a member of the family of procedures which are employed to describe the weighting and composition of values in subjective judgment situations. Applied to the goals achievement problem, this procedure seeks to simultaneously estimate the relative weights of the goals and the manner by which the achievement levels and weights are combined. It has seen wide application to policy assessment problems in the State of Colorado (USA) and is reviewed in Slovic and Lichtenstein (1973), Stewart and Gelberd (1976), and Slovic, Fischoff and Lichtenstein (1978). In this procedure the levels of achievement for each plan (alternative) vis-a-vis each goal are given, but weights are left unspecified. Those involved in the assessment process are asked to judge each plan regarding "how closely it comes in their mind to best satisfying the stated objective function." These judgments are made on numerical scales similar to those currently employed to measure the levels of achievement of each plan.

For example, suppose in the case of a coastal zone management plan one plan is described as involving the destruction of 250,000 acres of wetland ecosystems, preventing $10,000,000 of damages from coastal storms, and costing $5,000,000 to implement. There would be other plans each of which would have different values (achievement levels) for each of these goals. Then, each individual or evaluator is asked to evaluate each plan on a scale representing "how well" each plan performs overall with respect to all goals.

Formally, therefore, this is a controlled experimental design in which the levels of achievement of the goals are fixed and measured without error;

hence, the goals themselves are the independent variables. Both the relative weights and the functional form of the evaluation strategy or model employed by the "judges" or evaluators can be assessed, given this kind of experimental design. In particular, it can be demonstrated (Hammond, et.al., 1975) that multiple linear regression can be applied to this problem to derive relative weights and function forms for each goal. Furthermore, this type of procedure has the advantage that it can be easily and efficiently applied to a large number of people, each of whom can be directly compared because the design is "fixed" for all. The coefficients or parameters can be estimated with reference to some theoretical model (e.g., a linear function) and they can be derived for both individuals or groups. It is also possible to derive estimates of "weights" for individuals and/or groups, given the restriction that they apply only over the range of the goals employed in the judgment or evaluation task;* Hence, the weights are truly relative in that they apply solely to the context of the alternatives--any new alternatives could well change the relative weights, as could the addition or deletion of goals. Policy capturing has begun to see wide application in psychology; examples include the assessment of public sentiment in community goal-setting (Stewart and Gelberd, 1972; Hammond, et al., 1975); consumer assessments of public bus alternatives (Norman and Louviere, 1974); and evaluation of recreational opportunities by resource management personnel (Louviere, 1974), among a number of growing applications.

## An Empirical Comparison of Weights and Composition Rules: An Example in Flood Plain Management

The four procedures discussed above were compared by employing the same set of students in undergraduate geography courses at Florida State University

*Technically, weight for goal i = variance attributable to goal i ÷ total "explained" variance and the University of Wyoming as plan evaluators in two empirical studies. It is not suggested that the values of students are representative of those of any group other than their own, but the procedures may be tested on these individuals as well as any other group or groups of interest. More importantly, however, differences in weights as a function of differences in methods of procedure may be tested equally as well on students as any other group. Hence, their use is merely for convenience in illustration of application.

## Discussion of Procedures

1. Informal Consensus. In the first procedure weights were derived from informal consensus by assigning 60 students at Florida State University (FSU) to 15 groups of four students each. Each group was presented the following problem: "You are to compare and eventually choose from among several plans which have been designed to deal with the hazard presented by floods. Each plan is designed to accomplish, to some degree, each of four predetermined goals. The goals are to a) minimize total project costs; b) maximize the number of lives to be saved; c) minimize the destruction of natural habitat, and d) minimize the destruction of property. Some plans will do well with respect to some goals and not so well with respect to other goals. In order to select the best plan, you must decide how important each goal is, relative to the remaining goals." While the actual instructions were more detailed, the groups were requested to arrive at a consensus weight for each goal through informal discussion. Absolute weights were expressed with reference to a zero to 20 scale, and the aggregate weights were obtained by averaging over the 15 groups. Relative weights were obtained by dividing by the sum of the separate weights.

2. Polling. The same instructions were used in polling as in informal consensus but 30 FSU individuals were instructed to independently arrive at their own estimates of the weights with reference to the same zero to 20 scale. Aggregate weights were obtained by averaging over all 30 individuals. Relative weights were derived as in informal consensus.

3. Delphi Paneling. The same group of 30 students used in polling were employed as Delphi panelists. The average group weights were fed back to the individuals, and they were invited to reassess their weight estimates. The same group means were obtained in round two as were found in the first round (polling); hence, the process was terminated. Relative weights were derived as above.

4. Policy Capturing. In the policy-capturing condition 30 of the FSU students were presented with 45 different plans, each of which differed in the degree to which it achieved each goal. For example, one plan might cost $20,000,000; save ten lives annually; result in 10,000 acres of natural destruction; and avert $5,000,000 worth of property damage per year. The remaining plans each differed in their combinations of the levels of achievement of these goals. The 45 plans were chosen from among a very large number of possible plans so as to encompass a wide range of achievement levels on each goal; however, the design was not optimal for testing for a non-additive composition rule among the goals. That is, this design cannot reject a linear scoring function. Students were instructed to rate each plan on a scale from zero to 20, where the former score represented the worst plan they could imagine, and the latter represented the best plan regarding "how well" each plan fulfilled the stated goals. A multiple linear regression analysis was applied to the 1350 observations (45 plans x

30 evaluators) to obtain relative weights. The disadvantage of this procedure is that one must assume a linear statistical model without interaction effects to be true _a priori_. The advantage is that all evaluators are forced to make trade-offs such that the relative weights will reflect the degree to which that parameter depends on the values of the other evaluation criteria. Of course, if the experimental design is not balanced or if the model is not strictly linear, the regression weights will be biased. In order to test for this eventuality, a similar study was executed at the University of Wyoming.

5. _Policy Capturing Study II: A Test for Additivity of the Composition Rule_. In this study, 15 students at the University of Wyoming were presented with 48 different plans in the same format as described above. The principal difference in design was that the values were assigned to the goals so that a 2 x 2 x 2 x 2 (High-Low) cross-classification analysis could be performed on the data to assess dependencies. In particular, levels of achievement were assigned to each goal such that the median level of achievement for each goal represented a dividing point between "High" and "Low." Combinations of levels of achievement of goals then were deliberately chosen so that three plans would fall into each of the 16 cross-classes ($2^4$ cross-classes).* Individual judgments were made in precisely the same manner as in condition number four (4) above. The creation of a factorial design is the important difference because it contains sufficient variation to do an error analysis to retain or reject additivity.

*These cross classes represent the 16 possible combinations of High and Low achievement on each of the four goals.

Results of the Five Conditions

The relative weights assigned each goal by each evaluation procedure are presented in Table 2. While the values may appear to be similar, they are significantly different statistically at the 0.01 level and some yield different orderings of goals. The appropriate test is to treat Table 2 as a one-way-analysis of variance with the goals as levels of factor one and the procedures as replications. The critical hypothesis of no significant difference in procedures must be rejected for Table 2. A visual assessment of this test can be obtained by plotting the weight for each procedure against the corresponding goal, arbitrarily placed on the X-axis. Equality of procedures in determining weights demands that all lines of data be parallel to the X-axis and to each other. This was not the case. Thus, the procedures yield different weights. This result permits the conclusion that different procedures yield different relative weights, and hence, they cannot all be correct. More importantly, however, because there is no theory underlying any of these procedures, with the possible exception of the policy capturing procedure, there would be no basis for choosing among them and each could generate potentially different results.

Equally important to the empirical analysis is the test of the assumption of an additive scoring function or composition rule. The Wyoming study specifically permits one to reject additivity if it is inappropriate, although no comparison of the weights with the weights derived in the first four conditions is appropriate because the sample is different. The test for additivity is the retention of the null hypothesis of no significant interaction or dependency effects among the four goals. Because there are three replications per individual and 15 individuals, there is more than sufficient within and between individual variation to perform an analysis of variance test of the interaction effects within the 2 x 2 x 2 x 2 experimental design formed by the goals.

The results of an analysis of variance on the response data from the Wyoming sample revealed a number of significant interactions in the data. Graphical plotting of the appropriate interaction means (See Anderson, 1974; Louviere, 1978; Lerman and Louviere, 1978) revealed that all interaction effects were of the same form: when the two curves for high and low levels of achievement on one goal were plotted against high and low levels of achievement of a second goal, the two curves converged toward the low level of achievement and diverged toward the high level of achievement; additivity would require that the curves be parallel. A typical data plot is given in Figure 1.

This finding clearly precludes the use of a strictly additive model to interpret the data and requires further analysis because two values of a factor are sufficient only to reject linearity; more values are required to test for alternative function forms and to assist in interpretation. Of course, one possibility is that the interactions are merely artifacts of the judgment scale that was employed. We shall demonstrate that this is not a viable interpretation; and in fact, a considerable body of empirical and theoretical results suggest that these response scales are valid interval scales (Anderson, 1972, 1974, 1976).

The results of the research reported thus far, therefore, suggest that it is likely that these various procedures yield <u>different</u> relative weights, and therefore cannot all be correct. Moreover, it seems clear that relative weights should <u>depend</u> upon the values or levels of the evaluation variables

Figure 1

Louviere and Baker

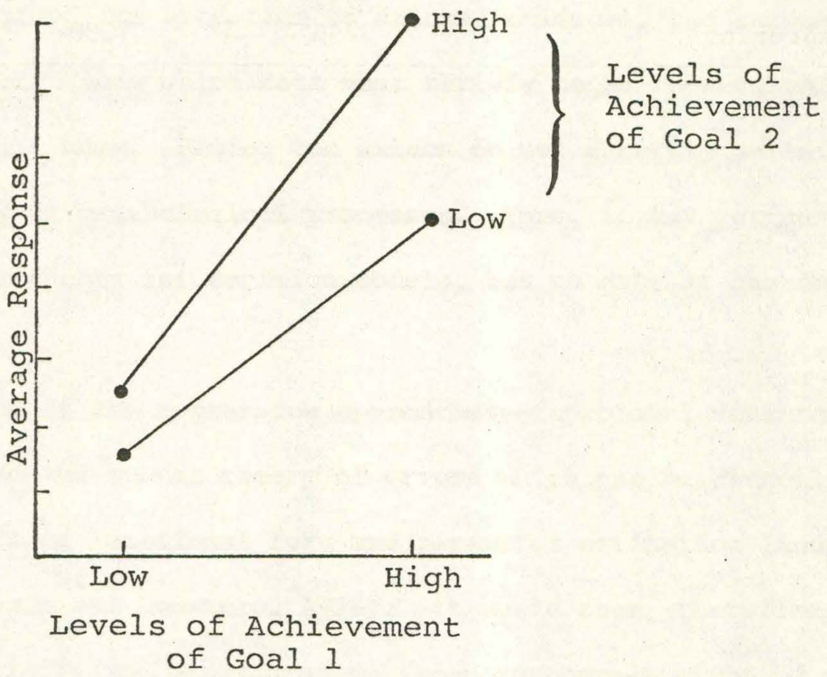

Table 2

Relative weights estimated by various procedures: Empirical Conditions 1-4.

| Procedures | Goals | | | |
|---|---|---|---|---|
| | Average Percent of Total Sum of Weights | | | |
| | $ Costs | Deaths Averted | Habitat Destroyed | # Damages Averted |
| Informal Consensus | .22 | .35 | .22 | .24 |
| Polling | .18 | .32 | .28 | .23 |
| Delphi | .17 | .31 | .29 | .22 |
| Policy Capturing | .19 | .38 | .30 | .14 |

(or goals); and hence, any approach that does not require individuals to explicitly express these dependencies (or trade-offs) must be biased. Most importantly, as Arrow and others have demonstrated (1963) individual values cannot be aggregated without doing violence to the notion of incommensurability of different value systems except in very highly restricted situations not likely to be true in practice. Therefore, any procedure which cannot develop unique weights and function forms for each individual must also be rejected as not being theoretically acceptable. The procedure tested in the Wyoming "Policy Capturing" study overcomes all of these objections and can be theoretically justified as we proceed to demonstrate.

## THEORETICAL FRAMEWORK

Axiomatic utility theory in economics, and behavior decision theory in psychology (see reviews in Keeney and Raiffa, 1976; and Slovic, Fischoff, and Lichtenstein, 1977) are two available sources of theory upon which to base an individual-level approach to plan evaluation. Although utility has traditionally been held as a normative theory, it is clear that it must also be descriptive in the sense that it is used to describe and, hence or provide an understanding of the decision maker's preference and/or value trade-offs. Thus, because it has the potential to provide a mathematical description of the individual decision maker's value assignments, it must be a candidate for a theoretical base.

On the other hand, behavorial decision theory encompasses a wide variety of general value assignment judgment behaviors and is most concerned with descriptions of these judgment processes, although some normative theory and application is evident. One particular area of behavioral decision theory appears particularly appropriate to the present problem: the area of information processing in judgment, to which the policy-capturing approach belongs. Unlike the axiomatic utility approach, this area has usually relied on linear and multilinear models to describe the judgment or evaluation process but has taken them as given without deducing them from sets of axioms. Although one paradigm--conjoint measurement--has devoted most of its attention to axiomatic issues, the axioms specify mathematical conditions which data must satisfy to be represented in various algebraic ways. Hence, the axioms do not directly pertain to either economic theory or psychological processes. Thus, it may yet serve as a potential error theory for decision models, but to date it has not satisfied this role.

Another one of the regression approaches--functional measurement--has a well-developed and robust theory of errors which can be directly applied to the diagnosis of functional form and parameter estimation (Anderson, 1972, 1974, 1976; Lerman and Louviere, 1978). It would seem, therefore, that the best of both utility and behavioral decision approaches might be successfully blended to develop a new approach to plan evaluation. It is now demonstrated that the algebraic base of the functional measurement approach is consistent with axiomatic utility results.

## Overview of Functional Measurement and Value Assignment

The term functional measurement describes an approach to modeling individual evaluation behavior which is based on an explicit theory of how individuals evaluate alternatives and which employs procedures from the statistical treatment of experimental designs to collect and analyze response data (Anderson, 1972, 1974, 1976). In order to develop the algebraic theory, the following general assumptions are required:

$$v(x_j) = f_j(X_j) \tag{1}$$

where

$X_j$ are controllable levels or values of goals, more generally thought of as attributes (and their levels) of alternatives;

$v(x_j)$ are the values attached to the $X_j$ by the individual during evaluation of alternatives. They are the marginal utilities or values;

$f_j$ are the j different mappings of the $X_j$ into their respective marginal values.

$$V = g(v(x_j)) \tag{2}$$

where

$V$ is the overall evaluation or value (or utility) assigned to the bundle of j attributes or goals;

$v(x_j)$ are the marginal utilities or values as defined above;

$g$ is a mapping defined over all j attributes or goals into V.

Equation (2) therefore asserts that overall evaluations are based on some composition of marginal values. By substitution of equation (1) in equation (2) one trivially derives:

$$V = g(f_j(X_j)) \tag{3}$$

where all terms are as defined previously.

Equations (1)-(3) are too general for modeling purposes, however, they lay the conceptual framework for modeling. In order to develop the algebraic theory for modeling, one must make quite specific assumptions, at least about equation (2). In particular, the question of equation (1) may be left unaddressed temporarily because it is primarily an empirical issue. The theoretical development, therefore, will focus upon equation (2), which is of critical interest. In particular, a most general form for equation (2) is proposed that permits a variety of common value expressions to be derived as special cases. The general form is that of a multilinear equation, a form which has seen considerable application and study in both psychology and economics (see, e.g., Keeney and Raiffa, 1976; Anderson, 1974). This form may be written as follows for the case of three goals or attributes:

$$\begin{aligned} V = k_0 + k_1v(x_1) + k_2v(x_2) + k_3v(x_3) + k_4v(x_1)v(x_2) + k_5v(x_1)v(x_3) \\ + k_6v(x_2)v(x_3) + k_7v(x_1)v(x_2)v(x_3), \end{aligned} \tag{4}$$

where the k's are scaling constants, V and $v(x_j)$ are as defined above.

Equation (4) may be generalized to j( = 1,..., J) independent attributes or goals; it should be noted that strictly additive or multiplicative model forms are special cases of the general form of equation (4) in which certain specifiable k's equal zero. For example, strictly additive specifications require that $k_4 - k_7$ be zero. This latter observation, of course, demands that one have an error theory for testing this expectation, and this is the strength of Functional Measurement: Functional Measurement recognizes the isomorphism between value response models such as equation (4) and models which represent the outcome of multifactor experiments. Hence, the theory posits that value responses can be studied and modeled by employing methods developed to implement the theory of the design and analysis of experiments, which of course include a theory of errors.

The critical role of equation (4) emerges as a consequence of the above discussion: diagnosis or testing of the coefficients of equation (4) is tantamount to being able to reject one or more specifications while retaining another. In order to implement this capability it is necessary to understand the mechanics of the design and analysis of experiments, particularly factorial or fractional factorial experiments (see, e.g., Snedecor and Cochran, 1967; Winer, 1972; Hahn and Shapiro, 1966).

In particular, it is necessary to "design" a utility or value assignment study according to principles of factorial designs. For example, if there are four attributes or goals, one begins by assigning i levels to goal one; j levels to goal two; k levels to goal three; and l levels to goal four. Then one most select combinations of goal levels from the i x j x k x l total possible combination of goal levels. Of course, if the product is small enough, one could use all combinations; frequently, however, the total combinations are too large for evaluators to handle with any facility and some reduction mechanism must be sought.

These reduction mechanisms encompass the class of designs termed "fractional" factorials. They are so termed because the reduction is accomplished by means of collapsing the design over levels of one or more attributes, resulting in a loss of ability to test some of the joint or cross-product coefficients. However, as will be demonstrated, this rarely affects estimates of marginal values. The preceding notions will now be formalized in the forthcoming treatment.

Consider a two attribute or two goal situation in which goal one has i levels and goal two has j levels. There are i x j possible combinations of levels of goals one and two. Rewriting equation (4) for this situation

$$V_{ij} = k_0 + k_1 v(x_{1i}) + k_2 v(x_{2j}) + k_3 v(x_{1i}) v(x_{2j}) \qquad (5)$$

where all terms are as defined previously, except that the levels and combinations (i,j) are subscripted. First, consider the effect of averaging equation (5) over the j subscript, holding $x_1$ constant at level 1:

$$\bar{V}_{1j} = [\sum_j k_0 + k_1 v(x_{1i}) + k_2 v(x_{2j}) + k_3 v(x_{1i}) v(x_{2j})]/J \qquad (6)$$

$$= K_0 + K_1 v(x_{2j}) \qquad (7)$$

where $K_0 = k_0 + k_1 v(x_{1i})$ and $K_1 = [v(x_{2j})][k_2 + k_3 v(x_{1i})]$.
Thus, equation (7) demonstrates that so long as any general form of equations (4) or (5) are true, then the marginal average of any level of any attribute or goal equals the desired marginal utility value up to a positive linear transformation. This is an important result because it suggests that any experimental design which permits unbiased estimates of the marginal means will yield estimates of the desired marginal values up to a linear transformation. Hence, the marginal means are "just as good" as any other estimates of the marginal values.

Moreover, an analysis of variance or a multiple linear regression provide an appropriate vehicle for error analysis because the coefficients of equations (4) and (5) are isomorphic with the "effects" of analysis of variance or the coefficients of a multiple linear regression analysis. Hence, if the evaluation study is designed as a factorial or fractional factorial experiment, it is almost always possible to derive the marginal terms; and, depending upon the design, it is usually possible to test many of the "joint" terms of interest. These "joint" or cross-product terms are of direct interest because they are the keys to the diagnosis or testing of various subset specifications of equations (4) and (5) or any expanded general form.

For example, if the underlying evaluation function is strictly additive in the marginal values it is written as:

$$V_{ij} = k_0 + k_1 v(x_{1i}) + k_2 v(x_{2j}), \quad (8)$$

where all terms were defined previously. It is clear that a strictly additive value function demands that all cross-product or "joint" terms (called Interaction Effects) be equal to zero, or statistically non-significant at some confidence level. This, of course, is the usual assumption in a goals achievement context in that the overall score is a weighted additive combination of goal weights and attainment levels. As will be demonstrated in a later empirical section of this paper, this is a testable assumption which can be rejected if false. Indeed, the notion of assigning weights to goals is only meaningful if equation (8) is true. To understand why, consider equation (5) rewritten as follows:

$$V_{ij} = W_0 + W_1 v(x_{1i}) + W_2 v(x_{2j}) + W_3 v(x_{1i}) v(x_{2j}), \quad (9)$$

where all terms are as before, except for the W's which are weights or constants. Now, examine the partial derivative of $V_{ij}$ with respect to $v(x_{1i})$:

$$\frac{\partial V_{ij}}{\partial v(x_{1i})} = W_1 + W_3 v(x_{2j}). \quad (10)$$

Equation (10) clearly indicates that any form of equation (5) other than a strictly additive specification yields weights which are not independent of the remaining attributes. Put briefly, if equation (5) is not strictly additive, the slope (or weight) of one attribute depends upon the levels of one or more additional attributes and is not assessable by independent questioning of the form illustrated by the initial empirical results and discussion.

If one assumes that equation (5) is strictly multiplicative, of course, the same result obtains as demonstrated immediately above. For example, assume strict multiplication for equation (5).

$$V_{ij} = k_0 + k_3 v(x_{1i})v(x_{2j}), \tag{11}$$

where all terms are as defined previously. Again, the slope for $v(x_{1i})$ depends upon the levels of $v(x_{2j})$ and cannot be independently assessed. A critical test for equation (11) can be derived by substituting equation (7) for both $v(x_{1i})$ and $v(x_{2j})$ in equation (11) to yield:

$$V_{ij} = k_0 + k_1[K_0 + K_1 v(x_{1i})][K'_0 + K'_1 v(x_{2j})], \tag{12}$$

where all terms are as defined previously. Expansion of equation (12) yields a general form equivalent to equation (5) and demonstrates that equations (11) and (5) are mathematically equivalent. Thus, equation (11) can be tested by noting that all terms in a general form of equation (5) must be significant.

The preceding section has demonstrated the theory of Functional Measurement as applied to diagnosis and testing of individual and group evaluation functions. It must be noted that the appropriate statistical tests must be carried out using the marginal means as independent variables; for this reason discussion of equation (1) was omitted so as not to confuse the development. Estimation of equations (1) or (3) is essentially an empirical question, once equation (2) is diagnosed. Indeed, the only value of equation (3) is its ability to interpolate over a larger range of options in the space of alternatives than one has observations. In this respect it can play an important role, but the shape or form of the specifications in equation (1) is clearly unique to an individual or a group of individuals or a particular problem or any combination of these. Furthermore, general procedures for estimation have been treated elsewhere and are not the focus of this paper (Louviere, 1978; Lerman and Louviere, 1978). The equations considered above may be estimated and tested using standard least-squares procedures. However, goodness-of-fit is different than lack-of-fit for these models and a note of caution in the next section is in order.

## GOODNESS-OF-FIT OR LACK-OF-FIT

The equations developed above may be rigorously estimated using least-squares procedures. One would often be tempted to use some measure of goodness-of-fit such as the correlation of model with data to test how well these models fit. However, as has been repeatedly demonstrated in recent years, correlation as a measure of fit can be very misleading indeed. Similarly, rules of thumb from applications of axiomatic utility theory such as the notion that if the scaling constants sum to unity the equation is strictly additive or if they sum to more than one, theoretically the equation is multiplicative (multilinear). The important question is, however, to how much more than one should the constants sum? Or how much more or less than one may one regard as not different than one? These questions require a rejection criterion or lack-of-fit and not goodness-of-fit.

It has been demonstrated that the strictly linear model will correlate highly with data even when some other multilinear form is appropriate. Those who care strictly for predictive ability need search no further--the strictly linear equation will recover the data admirably. Those who are interested in understanding and the ramifications of manipulating the system, however, must delve deeper and examine lack-of-fit. It is precisely in this area where the analysis of variance is most appropriate--it can detect significant departures from strictly linear assumptions, and through graphs and tests onvarious terms in the analysis of variance model, the specific functional form which is appropriate may be tested or diagnosed.

Nonetheless, the value of the strictly additive assumption should not be cast aside. It is now well-known (Anderson and Shanteau, 1977; Birnbaum, 1973; Dawes and Corrigon, 1974; and Wainer, 1976) that a linear equation will reproduce even multiplicative data well; and, more importantly, it will reproduce the rank order of the response data almost as well as the "true" model. This latter point means that one can use the additive assumption to reproduce the rank order of the expected values with very little error. The additive assumption simplifies experimental design because one never needs more than a "main-effects" experimental plan.* Such plans require vastly fewer combinations than complete factorial plans, and fewer than many fractional factorial sampling plans, as well. Hence, the evaluation functions may be assessed with a modest number of alternatives. Despite this optimism, it would not be fair to conclude that linear evaluation functions may be <u>directly</u> applied to goals assessment. They cannot. That is, they cannot be directly interpreted as representative

*This is a plan that permits only estimates of marginal values.

of the individual or the group evaluation functions unless they are truly "correct." But they have a powerful potential application nonetheless.

## A SIMULATION APPROACH

Because they predict rank order almost perfectly, linear functions can be used to simulate the choice or voting process. The planner proposes a number of realistic alternatives. The value of each alternative on each decision factor (goal) is substituted into each individual's equation, and the individual is assigned to that alternative which yields the highest predicted response value. Note that if the linear model reproduces the rank-order well this will be a close approximation. The number of individuals "choosing" or "voting" for each alternative may then be determined by counting and the choice proportions calculated.

These simulated choice proportions need have no linear relationship to the independent variables. Indeed, one strategy would be to use a factorial approach or fractional factorial approach to choose alternatives, simulate the choice proportions, and then re-estimate a model on the simulated result. This final model can serve as the evaluation function. The approach does no violence to Arrow's (1963) aggregation problem because each individual's value system is retained intact in this aggregation. One may, in effect, run a controlled experiment by factorially varying the alternatives and analyzing the resulting choice proportions. In the next section this approach is illustrated in a second empirical study.

## A SECOND EMPIRICAL EXAMPLE

Four attributes or goals for a new hypothetical cross-town expressway in Tallahassee, Florida, were studied in example two. The goals were (1) to reduce total travel time, (2) to lower the accident rate, (3) to minimize the number of displaced households, and (4) to minimize costs of the facility. It is not claimed that these exhaust the list of objectives, only that they are commonly considered and may be used to illustrate the approach. Again various popular methods for obtaining weights are compared with the approach outlined immediately above, principally to resolve unanswered questions regarding their use and for comparison with the theoretical approach advanced above.

### 1. Polling Condition

Fifty Florida State University (FSU) students were interviewed on the campus by trained graduate student interviewers. Criterion for selection was willingness to participate in a follow-up study. Students were interviewed at exits of ten randomly selected buildings around the campus. Interviewees were told that there was a proposal to build a cross-town expressway in Tallahassee, and that there were four major considerations: (1) how much travel time it would save, (2) how many accidents might be prevented, (3) how many families or households might be displaced, and (4) how much it would cost. Interviewees were then asked to judge how important each of these should be in the final decision. Judgments were made on a 0-20 scale which was labelled respectively "not at all important" and "most important." That is, each student was asked to estimate how much weight should be given to each factor. These data were converted to relative weights by dividing each by the total sum.

2. Delphi Condition

The same 50 students were contacted in person or by telephone one week later and informed of the overall group results. They were given their own results and asked if, on the basis of the group weights, they might wish to change their original estimates. The process was terminated after the first round of feedback, and a new set of relative weights estimated because there was no change in the group means after round one.

3. Functional Measurement Condition

Functional measurement is the term used for the evaluation analysis procedure described in the theory section. In this procedure 50 other FSU students were sampled by contacting them individually either in person or by phone, describing the nature of the study and asking them to participate. Selection was based on willingness to participate.

In this procedure the four goals were each assigned three values or levels: (1) potential time savings (3, 9, 15 minutes); (2) potential number of accidents averted (2, 6, 10 per month); (3) potential number of families displaced (100, 400, 700); and (4) potential total project cost ($5, $15, $25 million). All possible combinations of four factors at three levels constitutes a $(3)^4$ design (or 81 alternatives). In order to reduce this to a manageable number for the sample to evaluate, 27 alternatives were selected according to a one-third factorial plan such that all main effects (marginal value estimates) were independent of (not correlated with) all two-way interactions (Hahn and Shapiro, 1966) independently of remaining effects.

The choice of this particular set of 27 alternatives, therefore, yields the following desirable statistical properties:

1. The main effects of each of the four factors or goals are totally uncorrelated with each other and with three of the two-way interactions. Hence, a correlation matrix of the four factors using the values given above would consist of all zeroes off the main diagonal.
2. There are ample degrees of freedom for testing interaction effects (4,196) for the three estimable interactions. Additionally, each separate interaction component may be decomposed into four (1,196 degrees of freedom) separate tests.
3. The independent variables are fixed and measured without error, the only variation is in the dependent variable within and between student evaluators.

The 50 student subjects evaluated each alternative on a 21 category rating scale marked at either end by "very undesirable alternative" (0) and "very desirable alternative" (20). Students were shown two alternatives, one of which had more desirable values than any used in the study and a second which had more undesirable values than any used in the study. Subjects were asked to call these two alternatives 20 and 0, respectively. They were instructed to evaluate each of the alternatives by assigning a number between 20 and 0 reflective of where they felt that alternative fell between the two extremes. As another measure to guard extreme responses, five other alternatives were also inserted into the design for a total of 32 alternatives. These five were selected by assigning two of them either all desirable or all undesirable values. They were less undesirable than the two standard or "anchor" alternatives described above (the "0" and "20" alternatives) but were more desirable or undesirable than any of the 27 target alternatives. The remaining alternatives consisted of combinations which included both the more extreme values and the target values. They are used simply to insure that the subjects see these more extreme values more than once and do not become suspect of their role. The major role of these five "filler" alternatives is to keep the subjects from assigning artificially extreme (high or low) values to the most and least desirable target alternatives. Human subjects quickly learn the alternatives and will give more extreme responses to them than they would if the "filler" alternatives are not used. The procedures are necessary to insure against non-linearities in response scale use which would tend to make the response scale ordinal and not interval in measurement level (Anderson, 1974, 1976).

Data were analyzed by means of analysis of variance. Because each subject completes all 27 alternatives (treatment combinations), this is technically a repeated measures design, and each term in the model will have a separate and unique error term (the mean square of the term by subjects interaction). Because of the design selected, analysis may proceed as if the design were a $3^3$ x 50. This is because the interactions between the three factors of interest are estimable as part of a $3^3$ set of 27 treatment combinations--only the fourth factor is confounded, and only the interactions of this factor are confounded--the main effects of the fourth factor (which is cost) are still estimable. Essentially, the three two-way interactions with cost are perfectly correlated with either all three of the three-way interactions or two of them and the four-way interactions. Thus, only the three two-way interactions and the main effects are reliably estimable.

Based on the results of the overall analysis just described, an estimate of the appropriate group model may be inferred from the pattern of the statistical corresponding graphical results. Separate individual models may then be fit, and an estimate of the relative weights obtained (measured as the proportion of variance accounted for by each factor). These weights may then be compared to the weights derived from the Polling and Delphi procedures and with the weights derived from a strictly linear assumption. A statistical comparison would consist of conducting an analysis of variance (methods by subjects) for the two separate groups of subjects who are the same for each of two procedures (group one: Polling and Delphi; group two: linear and multilinear models). Then the two groups could be compared across the four procedures by means of paired t tests. Because the same subjects were not used in both groups, an analysis of variance cannot be conducted. The models to be fit are the following:

$$V_{ijkl} = k_0 + k_1(V_{i...}) + k_2(V_{.j..}) + k_3(V_{..k.}) + k_4(V_{...l}) + e_{ijkl} \tag{13}$$

$$V_{ijkl} = k'_0 + k'_1(V_{i...} - k'_0)(V_{.j..} - k'_0)(V_{..k.} - k'_0)(V_{...l} - k'_0) + e_{ijkl} \tag{14}$$

where $V_{ijkl}$ is the numerical response observed as a function of the ijkl-th alternative; $V_{i...}$, $V_{.j..}$, $V_{..k.}$, and $V_{...l}$ are the marginal means (estimates of the marginal utilities as demonstrated in equations (6)-(7) and the ensuing discussion) for each factor; the k's are scaling constants and the e's are error terms.

It can be demonstrated that if equation (13) is true, it may be written as:

$$V_{ijkl} = (V_{i...}) + (V_{.j..}) + (V_{..k.}) + (V_{...l}) - (3V_{....}) + e_{ijkl} \tag{15}$$

where $(V_{....})$ is the grand mean. Thus, all scaling constants are known using this approach. Because of the design selected, each vector of marginal means in uncorrelated with each other vector. Hence, the proportion of variance accounted for by each factor can be readily determined.

Likewise, it can be demonstrated that if equation (14) is true, it may be written as:

$$V_{ijkl} = k'_0 + \frac{1}{(V_{....} - k'_0)^3} [(V_{i...} - k'_0)(V_{.j..} - k'_0)(V_{..k.} - k'_0) \quad (V_{...l} - k'_0)] + e_{ijkl} \tag{16}$$

where $k'_0$ is an intercept term necessary to allow for the arbitrary zero in the response scale. Thus, in the case of equation (16) $k'_0$ must be estimated separately for each individual. This may be done, for example, by means of generalized least-squares in a stepped-search algorithm for finding the minimum of the least-squares function. Once $k'_0$ is known, equation (16) may be rewritten as:

$$\ln(V_{ijkl} - k'_0) = \ln(V_{i...} - k'_0) + \ln(V_{.j..} - k'_0) + \ln(V_{..k.} - k'_0) + \ln(V_{...l} - k'_0) - 3\ln(V_{....} - k'_0) \tag{17}$$

which is estimable via log-linear regression and again, the proportion of variance accounted for by each variable may be obtained. Moreover, it might be noted that equation (17) serves also as a lack-of-fit test for the theory in that all constants are predicted to be unity *a priori* and one can test for significant departures from this prediction as a standard test in regression analysis.

Results

Because the factors are statistically uncorrelated and designed to cover the entire range of variation of these factors, the proportion of variance in the dependent variable accounted for by each factor relative to the total "explained" variation is an estimate of its relative "weight." To obtain relative weights that sum to unity, the proportion of variance accounted for by each factor is divided by the total explained variation. Relative weights for each subject are computed under three different model hypotheses:*

1. a linear model using marginal means as estimates of the marginal utility values (Eq. 15)
2. a multiplicative model using marginal means as estimates of the marginal utility values (Eq. 16)
3. a strictly linear model using the experimental values (levels) as predictor values:

$$V_{ijkl} = a + b_1 \text{ Time}_i + b_2 \text{ Cost}_j + b_3 \text{ Accidents}_k + b_4 \text{ People}_l + e_{ijkl}$$

The results of these analyses and the polling and Delphi procedures are contained in Table 3. The null hypothesis is that there are no significant differences due to the three methods associated with this evaluation procedure. This hypothesis is tested by an analysis of variance (method x factor x subjects). Results reveal that the null hypothesis must be retained. Hence, all information-integration computations yield similar results,

*It should be noted that relative weights are confounded with the scale and range of attribute values and are only meaningful if the model is strictly additive. Thus, this exercise is for comparative purposes only. It is important to note, therefore, that inferences regarding "importance" are unfounded; moreover, these weight values are completely relative to this experimental design--another design could yield different results.

Table 3

Relative weights estimated by various procedures: Experiment 2.

| Procedures | Goals | | | |
|---|---|---|---|---|
| | $ Cost | Travel Time | People Displaced | Accidents Averted |
| Polling | .20 | .25 | .30 | .27 |
| Delphi | .19 | .25 | .30 | .27 |
| Information Integration | | | | |
| 1) Additive model, non-linear in marginals | .21 | .36 | .27 | .15 |
| 2) Multiplicative model | .22 | .35 | .32 | .12 |
| 3) Strictly additive model | .22 | .35 | .29 | .14 |

although the averages for factors and methods show a slight tendency for the strictly linear procedure to yield relative weights different than the other two methods.

Comparisons between these results and the polling and Delphi methods are only valid if both groups of 50 subjects represent random samples drawn from the same population, and they are not. For the sake of comparison, multiple t tests were performed on the pooled factor means from the polling and Delphi methods which yielded indentical results and the pooled factor means from the three models which yielded similar results. Results of this analysis suggested that the two sets of methods yield different results. However, as suggested above, a better test of this hypothesis would have been accomplished if all subjects would have participated in all conditions. Thus, the results very tentatively suggest as before that the simple methods yield results quite different from the decision analysis methods. However, although all of the decision modeling methods appear to provide similar estimates of the relative weights, their interpretation would differ depending upon the model specification employed to interpret the decision process. A multiplicative process is supported by the overall results; and it has been repeatedly found in previous research (Louviere, 1978; Louviere and Wilson, 1978; Louviere and Levin, 1978). Furthermore, it makes logical sense: multiplicative processes imply that the values of one or more of the attributes act to modify or intensify the response depending upon the levels of the other attributes with which they are combined. Multiplicative processes suggest that if any one level of any attribute is at an unacceptable level, it matters little what levels the remaining attributes have: the entire bundle is probably unacceptable. The individual level results leave a mixed picture: because each respondent completed only a single replication of the design one cannot technically diagnose individual equations. The individual $R^2$ values are not diagnostic and may even be misleading, although they favor the multiplicative hypothesis. More precise diagnosis would be desirable. If the respondents completed a replication of the experiment there would be sufficient variation within an individual for an error analysis. Future work will explore other possibilities, as well.

## CONCLUSIONS

This paper compared a number of different procedures for deriving relative weights for goals in planning contents. Although a number of issues have been resolved, doubt remains regarding whether the polling and Delphi procedures yield adequate estimates. However, results of Study I suggest that the results are very different, and Study II shows a similar result, although it is weaker than Study I because different respondents were employed under different conditions.

More importantly, empirical results support the theoretical exception that utility functions can be diagnosed and tested for both aggregations and individuals. These functions can be employed to assess individuals' (or the "public" if the sample is judiciously selected) reaction to various alternatives by substituting the values of each goal in each individual's value function and assigning each individual to that alternative with the highest expected value score. Aggregate choice proportions can then be derived by summing the total assignments over all relevant alternatives. Similar to cost/benefit and related methods, the utility functions permit all factors to be measured or assessed in a common metric--the overall response or value metric or the number of choices of "votes." Hence, it appears that these procedures offer considerable promise for solving both the commensuration problem and the aggregation of individuals problem. More research is necessary, however, before such a conclusion can be strongly stated.

## REFERENCES

Anderson, N. H., 1972, "Cross-Task Validation of Functional Measurement," Perception and Psychophysics, 12, 389-395.

Anderson, N. J., 1974, "Information Integration Theory: A Brief Survey," In Contemporary Developments in Mathematical Psychology, Vol. 2, Eds. D. H. Krantz, R. C. Atkinson, R. D. Luce and P. Suppes (W. H. Freeman, San Francisco).

Anderson, N. H., 1976, "How Functional Measurement Can Yield Validated Interval Scales of Mental Quantities," Journal of Applied Psychology, 61(6), 677-692.

Anderson, N. H. and Shanteau, J., 1977, "Weak Inference with Linear Models," Psychological Bulletin, 84, 1155-1170.

Arrow, K. J., 1963, Social Choice and Individual Values, 2nd Ed. (John Wiley, New York).

Baker, E. J., 1976, "Toward an Evaluation of Policy Alternatives Governing Hazard Zone Land Uses," Natural Hazard Research Working Paper No. 28, Institute of Behavioral Science, University of Colorado, Boulder, Colorado.

Birnbaum, M. H., 1973, "The Devil Rides Again: Correlation as an Index of Fit," Psychological Bulletin, 79, 239-242.

Dalkey, M. C., 1968, Experiments in Group Prediction (The Rand Corp., Santa Monica, California).

Dawes, R. M. and Corrigan, B., 1974, "Linear Models in Decision Making," Psychological Bulletin, 81, 95-106.

Hahn, G. J. and Shapiro, S. S., 1966, "A Catalog and Computer Program for the Design and Analysis of Orthogonal Symmetric and Asymmetric Fractional Factorial Experiments," Technical Report No. 66-C-165, General Electric Research and Development Center, Schenectady, New York.

Hammond, K. R., Stewart, R. T., Beehmer, B. and Steinman, D. O., 1975, "Social Judgment Theory," In Human Judgment and Decision Processes, Eds. M. F. Kaplan and S. Schwartz (Academic Press, New York).

Hill, M., 1968, "A Goals Achievement Matrix for Evaluating Alternative Plans," Journal of the American Institute of Planners, 34, 19-29.

Keeney, R. and Raiffa, H., 1976, Decisions Among Multiple Objectives: Preference and Value Tradeoffs (Wiley, New York).

Lerman, S. R. and Louviere, J. J., 1978, "On the Use of Functional Measurement to Identify the Functional Form of Travel Demand Models," Transportation Research Record, forthcoming.

Lichfield, N., Kettle, P. and Whitbread, M., 1975, Evaluation in the Planning Process (Pergammon Press, New York).

Louviere, J. J., 1974, "Predicting the Response to Real Stimulus Objects from an Abstract Evaluation of their Attributes: The Case of Trout Streams," Journal of Applied Psychology, 59, 572-277.

Louviere, J. J., 1978, "Psychological Measurement of Travel Attributes," In Determinants of Travel Choice, Eds. D. A. Hensher and M. Q. Dalvi (Saxon House Studies, Teakfield Farnborough, London).

Louviere, J. J. and Levin, I. P., 1978, "Functional Measurement Analysis of Spatial and Travel Behavior," Proceedings, Association for Consumer Research, forthcoming.

Louviere, J. J. and Wilson, E., 1978, "Predicting Consumer Response in Travel Analysis," Transportation Planning and Technology, 4, 1-9.

Nijkamp, P., 1976, "Stochastic Quantitative and Qualitative Multi-criteria Analysis for Environmental Design," Research Memorandum No. 56, Dept. of Economics, Free University, Amsterdam.

Norman, K. L. and Louviere, J. J., 1974, "Integration of Attributes in Public Bus Transportation: Two Modeling Approaches," Journal of Applied Psychology, 59, 753-758.

Prest, A. R. and Turvey, R., 1965, "Cost-benefit Analysis: A Survey," The Economic Journal, 75, 683-735.

Slovic, P. and Lichtenstein, S. 1971, "Comparison of Bayesian and Regression Approaches to the Study of Information Processing in Judgment," Organizational Behavior and Human Performance, 6, 699-744.

Slovic, P., Fischoff, B. and Lichtenstein, S., 1977, "Behavioral Decision Theory," Annual Review of Psychology.

Snedecor, G. W. and Cochran, W. G., 1967, Statistical Methods (Iowa State University Press, Ames, Iowa).

Stewart, R. T. and Gelberd, L., 1972, "Capturing Judgment Policies: A New Approach for Citizen Participation in Planning," U.R.I.S.A. Proceedings, Urban and Regional Information Systems Association, San Francisco.

Wainer, H., 1976, "Estimating Coefficients in Linear Models: It Don't Make No Nevermind," Psychological Bulletin, 83, 213-217.

Winer, B. J., 1972, Statistical Principles in Experimental Design (McGraw-Hill, New York).