DOCUMENT RESUME

ED 142 562 TM 006 173

AUTHOR Baron, Joan
TITLE An Exploration of the Implication of the M.A.U.T.-Bayesian Decision-Theoretic Model for Summative and Formative Evaluation and Post-Assessment Organizational Change. Research Report Series.
INSTITUTION Connecticut Univ., Storrs. Bureau of Educational Research and Service.
PUB DATE [Apr 77]
NOTE 27p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *Bayesian Statistics; Data Collection; *Decision Making; *Evaluation Methods; *Formative Evaluation; *Models; Organizational Change; Problem Solving; Program Effectiveness; Program Evaluation; *Summative Evaluation; Values
IDENTIFIERS Decision Theoretic Testing; Multiattribute Utility Bayesian Decision Model

ABSTRACT

The philosophy and assumptions of the Multi-Attribute Utility-Bayesian Decision Theoretic model (MAUT-Bayesian model) are presented. The evaluator uses the MAUT-Bayesian model along with the knowledge of the decision-maker's, and perhaps the evaluator's own values to decide what data should be collected. Appropriate data are presented to the decision-maker, who weighs the alternatives suggested by the multiple aspects of the data. Using an educational policy question as an example, each step of this decision-making process is described. Implications for formative and summative evaluations and post-assessment organizational change are offered. (Author/MV)

An Exploration of the Implication of the M.A.U.T.-Bayesian Decision-Theoretic Model for Summative and Formative Evaluation and Post-Assessment Organizational Change

by

Joan Baron

University of Connecticut

A paper presented to the American Educational Research Association 1977 Annual Meeting, April 4-8, 1977, New York City.

The author wishes to thank Drs. Edward Iwanicki, Robert Gable and Marcia Guttentag for their helpful remarks in an earlier version of this paper.

An Exploration of the Implication of the M.A.U.T. - Bayesian Decision-Theoretic Model for Summative and Formative Evaluation and Post-Assessment Organizational Change

Joan Baron
University of Connecticut

The major goal of this paper is to familiarize the reader with the Multi-Attribute Utility-Bayesian Decision Theoretic model of evaluation. The first part of the paper will contain an exploration of the philosophy and assumptions of the model; the second section will provide a step by step application; the final section will discuss its implications for formative and summative evaluations and post-assessment organizational change.

Philosophy and Assumptions of M.A.U.T. - Bayesian Decision-Theoretic Model

The role of the evaluator in the M.A.U.T. - Bayesian model is that of a facilitator for decision-making. The evaluator collects data and presents it to the decision maker who will then make a decision. Perhaps the most important question an evaluator must answer is, "What data should be collected?" It is in answering this question that the M.A.U.T. - Bayesian model is most useful as it is derived from the assumption that people make decisions by evaluating the various entities (alternatives) on many relevant value dimensions (see Raiffa, 1968, pp IX-X). Generally, people have certain minimum criteria which must first be met. After that, the alternatives are weighed and a decision is made. In a decision to purchase one of two houses, after certain size and price criteria have been satisfied, houses will differ on location, state of repair, amount of insulation, etc. Each of these dimensions will be considered and a final decision will be made. People are routinely called upon

to choose between apples and oranges. And they do it. Returning to the question above regarding what data should be collected, it must be answered that data should be collected on whatever value dimensions the decision maker considers to be important.

If two programs are to be compared, certainly data will be collected on the program's effectiveness as in most program evaluations, this will be the most important dimension. However, many additional factors may also be important. For example, the cost of the program, the amount of training required, attitudinal changes of the participants, etc. may be weighed in the decision making process. The M.A.U.T. - Bayesian model acknowledges the multi-faceted complexity of decision-making and attempts to quantify the process by isolating the values held by the decision-maker and prioritizing them in the same way he or she does when making the decision. Data will then be collected by the evaluator to determine the extent to which the program succeeds on the dimensions which are important. One may ask whether the evaluator should inject his own values into his evaluation and collect data on those. He or she may decide to do so. However, it should be recognized that the decision-maker may elect to ignore those data if he/she does not value the dimension even after being confronted with the data.[1] This may be part of the reason why many well-intentioned evaluations are put into a drawer and never used. The evaluator may have provided information on program effectiveness which is of little importance to the decision-maker. Furthermore, the evaluation may have contained no information on dimensions which were important to the decision.

It should be stated that the M.A.U.T. - Bayesian model encourages the use of experimental and quasi-experimental designs whenever appropriate and possible. Data should be collected using the principles of Campbell and

---

[1] It will be clear after the second section of this paper that the additional data must be presented separately and not included in the matrix.)

Stanley (1963) and Cook and Campbell(1975). The use of control groups wherever feasible is strongly advocated, particularly when evaluating the program's effectiveness.[2]

The Bayesian aspect of the model procedes from the belief that people have ideas regarding the probabilities for certain events to occur, and preferences or utilities for those consequences which are independent from the probabilities. Edwards, et al. (1973) provide the following illustrations:

> What action is wise of course depends in part on what is at stake. Would you not take the plane if you believed it would crash, and would not buy flight insurance if you believed it would not. Seldom must you choose between exactly two acts, one appropriate to the null hypothesis and the other to its alternative. Many intermediate, or hedging, acts are ordinarily possible; flying after buying flight insurance, and choosing a reasonable amount of flight insurance, are [redacted] examples. (p. 214)

The decision maker concerned with a program evaluation generally has ideas regarding the programs' effectiveness prior to the time that the evaluator arives . After the data are amassed, the original probabilities are either confirmed or disconfirmed.[3] Edwards, et al. 1963 (p. 208) wrote:

---

[2] A discussion of "pseudo-experiments" in Edwards et al. (1975, pp. 143-145) urges the reader to be wary of using control groups which do not control. They urge the use of convergent validity to remedy the limitations often confronted in field settings where randomization is not possible and programs change continuously.

[3] "In the Bayesian approach to statistics, an attempt is made to utilize all available information in order to reduce the amount of uncertainty present in an inferential or decision-making problem. As new information is obtained, it is combined with any previous information to form the basis for statistical procedures. The formal mechanism used to combine the new information with the previously available information is known as Bayes' theorem; this explains why the term "Bayesian" is often used to describe this general approach to statistics... When new information is obtained, probabilities are revised in order that they may represent all of the available information." (Winkler, 1972; p. 2)

> "If it were meaningful utterly to ignore prior opinion, it might presumably sometimes be wise to do so; but reflection shows that any policy that pretends to ignore prior opinion will be acceptable only insofar as it is actually justified by prior opinion. Some policies recommended under the motif of neutrality, or using only the facts, may flagrantly violate even very confused prior opinions, and so be unacceptable."

In a later part of their discussion, Edwards et al. points out that work by Hays, et al. (unpublished) (p. 212) that in reality people tend to disbelieve evidence which does not confirm their original beliefs.

> "Subjects are unwilling to change their diffuse initial opinions into sharp posterior ones, even after exposure to overwhelming evidence. This reluctance to extract from data as much certainty as they permit may be widespread. If so, explicit application of Bayes' theorem to information processing tasks now performed by unaided human judgment may produce more efficient use of the available information."

It is for the above reasons that an inferential system which closely mirrors the way in which people process discrepant data would tend to be more useful to the decision maker. The reader who wishes to pursue these issues is urged to read Edwards (1963) et al. in its entirety. Appendix A below reproduced their Figure 2, which graphically illustrates how two very different prior judgments are altered by data so that the posterior curves begin to resemble each other after the data is amassed. An understanding of this concept is essential to understanding the way in which subjective prior judgments are recast into posterior probabilities through the use of data. It will also aid the evaluator in determining how much data would be necessary to alter the prior probabilities. (For an application of the Bayesian approach in an evaluation setting, see Edwards et al. 1975, pp. 175-177.)

An Explication and Application of the M.A.U.T. - Bayesian Model[4]

The second goal of this paper to explicate for practicing evaluators the decision-theoretic approach to evaluation research by applying it to a simulated evaluation problem. The approach as espoused by Edwards, Guttentag, and Snapper (1975) provides a methodological and statistical framework for using evaluation as the input for intelligent decision making. Evaluators frequently acknowledge the existence of the decision-theoretic or otherwise known as the Multi-Attribute Utility Analysis (M.A.U.T.) or Bayesian approach. However, due to its seeming complexity it has not been frequently employed by those evaluators not specially trained in it.

For the purposes of applying the decision-theoretic model, we will use a hypothetical alternative program such as those prevalent in Philadelphia (e.g., Parkway-school-without-walls/storefront-type-school) for culturally and academically disadvantaged potential drop-outs. We will assume that the high school has been in existence for three years and the School Board is making a decision as to whether the storefront alternative should be allowed to continue in its present form. If not, should it be modified to resemble School X or disbanded with the children reentering the traditional high school? (See Roberts (1975) Chapter 6 for descriptions of similar programs).

---

[4]This example uses contrived data and was presented by Baron (1976). The 10 methodological steps of the model were taken from Edwards, et al, 1975. Further elaborations may be found in Guttentag (1973) and Guttentag and Snapper (in press).

By way of preview, the essence of the decision-theoretic approach is to find out what the values of each primary interest group are and then measure the extent to which each of these values is being met by each of the three programs being considered. The one that does the best on the overall basis is the one to be chosen. The M.A.U.T. model delineates a set of 10 steps to follow in achieving this end.

Step 1. It must be determined whose utilities are to be maximized. That is, what are the various prime interest groups affected by and affecting the decision? In this situation, one is concerned with the values of those on the School Board, the teachers and administrators in the school, the parents and the students. Edwards et.al. (p.153) claim that "everyone who has a stake and voice in the decision must be identified and people who can speak for them must be identified and induced to cooperate.

Step 2. One must clarify the purpose for which the evaluation is being conducted, as the same objects or acts may have different values depending on the context and purpose. This has been identified above as the desire to select from among three alternatives: the storefront school, a modified alternative program, and a traditional school program.

Step 3. The alternatives or entities being evaluated should be specified. (This is the same as step 2 in this particular situation.)

Step 4. This is the first technical task. It requires the discovery of what dimensions of value are important to the evaluation of the entities being decided upon. Edwards et.al. recommends stating these as general dimensions eg. acquisition of reading skills, whereas Iwanicki (1976) recommends using specific behavioral objectives focusing on actual student behaviors. In this study, we will attempt the latter. It is critical to mention that a separate list will be drawn up by each separate group. This

step merely lists the important values and dimensions; no attempt is made to judge whether a particular entity or program succeeds on that dimension. It should be noted here that to the extent to which each group has previously done a needs assessment the task will be simplified. Some possible dimensions generated by three of the groups in our storefront school evaluation are listed in Table I.

Table I

Some Possible Value Dimensions generated by
Students, Teachers and Parents

Students' Value dimensions:

We will learn the basic skills.
We will be prepared for a job.
We will know how to solve problems and make decisions.
We will be independent.
We will feel good about ourselves.
We will feel as though the teachers like us and care about us.

Teachers' Value dimensions:

The students will stay in school instead of dropping out.
The students will have a basic sense of self-worth, self-confidence, independence.
The students will learn the basic skills of communication and computation.
The students will have a level of career aspiration commensuate with their ability.
The students will have a sense of social responsibility and dependability.
The students will have mastered some techniques for problem solving and decision-making.
The students will show pride in the quality of their work.
The students will have a basic appreciation of aesthetics and some meaningful options for leisure-time activities.

Parents' Value dimensions:

The students will stay in school instead of dropping out.
The students will be prepared for a good job.
The students will have mastered the basic skills.
The students will be dependable and responsible.
The students will know how to make decisions and solve problems.
The students will be independent.
The students will be confident and feel a sense of self-worth.

It will quickly be noticed that there are some goals which appear on all three lists and some which appear on only one of two.

This step is very similar to what Renzulli recommends in his Front End Analysis. "At the end of the Front End Analysis the evaluator should be able to list the major concerns of each prime interest group and these concerns should be classified and organized according to similarities between the groups." The major difference between Renzulli's approach and this one is that no attempt will be made to merge the different lists. Each list will be evaluated separately and fed back to the group which generated it. Iwanicki (p. 13) also acknowledged the collaborative aspect of developing an evaluation program. At the secondary school level he advocates, that "the persons responsible for planning and implementing the evaluation program should make every effort to involve the school staff in this process." He makes no mention of the students' voice in developing the evaluation program.

Step 5. This step consists of ranking the dimensions in order of importance. This ranking job can be performed either by individuals acting separately or in a group. According to Edwards, Guttentag and Snapper the preferred technique (p. 155) is to "try group process first, mostly to get the arguments on the table and to make it more likely that the participants start from a common base." Disagreements within groups at steps 5 and 6 seem to be due to conflicting values and Edwards et.al. "wish to respect them as much as possible... For that reason, we feel that the judges who perform steps 5 and 6 should either be the decision maker(s) or well-chosen representatives. Considerable discussion, persuasion and information exchange should be used in an attempt to reduce the disagreements as much as possible." They realize that this "will seldom reduce to zero and state that one function of an executive is to resolve disagreements among subordinates. If no resolution

is possible we can only do an evaluation separately for each of the disagreeing individuals or groups, hoping that the disagreements are small enough to have little or no action implications."[5]

For an example of ranking the dimensions in order of importance, refer to Table I under Teachers' Value Dimensions. These were listed in order of importance.

Step 6. In this step, the dimensions will be ranked in order of importance, while preserving the ratios between them. The first step is to assign to the least important dimension an importance weight of 10. The next most important dimension will be assigned a number that reflects its ratio of importance relative to the one below it, assigned a 10. The evaluator will continue up the list recording the group's assigned weights and checking each set of implied ratios as each new judgment is made. Thus, if a dimension is assigned a weight of 20 while the one above it is assigned a weight of 80, this means that the dimension worth 20 is ¼ as important as the one worth 80. By the time the most important dimension is assigned a value, there will have been revisions made to make previous judgments consistent with later ones. Revisions are very much in the spirit of the flexibility, change and openness encouraged by this process. For illustration, weights will be assigned in Table II to the Teachers' Value Dimensions.

---

[5] "A special case arise when one of the dimensions such as cost is subject to an upper bound, i.e., there are budget constraints. In that case, 4-10 should be done ignoring the constrained dimension. Then benefit-to-cost ratios will be calculated. In the absence of budget constraints, cost is just another dimension of value, to be treated on the same footing as all other dimensions of value, entering into U_1 with a minus sign, like other unattractive dimensions." (This will make more sense later.)

Table II

Teachers' Value Dimensions

10	Aesthetics.
15	Pride in work. (slightly more important)
30	Problem-solving and decision making. (double price above, triple aesthetics)
30	Responsibility and dependability. (same as No. 3)
50	Level of aspiration. (5 times more important than aesthetics)
100	Basic skills. (twice as important as level of aspiration, 10 times more than aesthetics)
100	Self-worth. (same as basic skills)
100	Keep students from dropping out. (same as basic skills and self-worth.

Step 7. After the value dimensions have been weighted, Edwards et.al., 1975, define the following "computational step which converts importance weights into numbers that are mathematically rather like probabilities. The importance weights will be summed, each weight will be divided by the sum and multiplied by 100. The choice of a 0-to-100 scale is, of course, purely arbitrary. At this step, the consequences of including too many dimensions at Step 4 becomes glaringly apparent. If 100 points are to be distributed over a set of dimensions and some dimensions are very much more important than others, then the less important dimensions will have non-trivial weights only if there aren't too many of them. As a rule of thumb, 8 dimensions is plenty and 15 is too many. Knowing this, one will want at Step 4 to discourage respondents from being too finely analytical; rather gross dimensions will be just right. Moreover, it may occur that the list of dimensions will be revised later, and that revision, if it occurs, will typically consist of including more rather than fewer." As an illustration the weights listed in Table II will be elaborated in Table III where they will be summed and each will be divided by the sum and multiplied by 100. It can be observed that the ratios of importance have been preserved in this process. Aesthetics with a value of 2.29 continues to be ten times less important than basic skills with a value of 22.98 paralleling step 6 with 10 and 100.

Table III

Illustration of Summing and Dividing Value Weights

Aesthetics	10	10/435=2.29 (These have
Pride in work	15	15/435=3.44 been multiplied
Problem solving and decision making	30	30/435=6.89 by 100)
Responsibility and dependability	30	30/435=6.89
Level of aspiration	50	50/435=11.49
Basic Skills	100	100/435=22.98
Self-worth	100	100/435=22.98
Drop out prevention	100	100/435=22.98
Sum =	435	Sum = 100

Step 8. To recapitulate for a moment: A matrix can now be set up for each primary interest group. The values will be listed across the top, one per column with the value assigned to it in step 7. The rows down the side represent the various alternatives to be weighed in the decision, the store-front school, a modification, a return to tradition school. Our next task is to fill in each cell of the matrix, i.e., "to measure the location of each entity being evaluated on each dimension..."[6] It should be stressed that this matrix is subject to modification at any point in time. Groups can add or delete goals and/or alternatives. This would be in line with Iwanicki's recommendation (p. 13) that "as the evaluation program is being developed and implemented the staff should have the opportunity to systematically review its effectiveness and make modifications where necessary." He and the proponents of the decision-theoretic model share the view that this refinement process is "essential to improved evaluation and decision making."

The next task is to select evaluation instruments to use in collecting the information which will be used in each of the cells in the matrix. The two criteria suggested by Iwanicki would be useful here:

[6] When programs do not yet exist and are potential new options, these judgments are no more than educated guesses. As the program proceeds, data are gathered and the standard techniques of Bayesian statistics can be used to update the initial guesses as data accumulates. (Edwards et.al., 197) The decision-maker and/or program experts may be helpful in recommending instruments they have faith in.

1. The extent to which the instrument accurately measures the objectives of the program being evaluated.

2. The convenience with which the results provided by the instrument can be used to make decisions about the students' achievement of the programs' objectives.

However, he warns that the selection of quality instruments does not always insure that accurate feedback will be collected. Care must be taken to see that the instruments are administered properly.

According to Edwards et.al. (p.156) there are three classes of dimensions--purely subjective, partly subjective, and purely objective:

> The purely subjective dimensions are perhaps the easiest; you simply get an appropriate expert to estimate the position of that entity on that dimension on a 0-to-100 scale, where 0 is defined as the minimum plausible value on that dimension and 100 is defined as the maximum plausible value. A partly subjective dimension is one in which the units of measurement are objective, but the locations of the entities must be subjectively estimated.
>
> A wholly objective dimension is one that can be measured rather objectively, in objective units, before the decision. For partly or wholly objective dimensions, it is necessary to have the estimators provide not only values for each entity to be evaluated, but also minimum and maximum plausible values, in the natural units of each dimension. (p.156)

According to Edwards, et.al. "The final task in step 8 is to convert measures in the partly subjective and wholly objective dimensions into the 0-to-100 scale in which 0 is the minimum plausible and 100 is the maximum plausible. A linear transformation is almost always adequate for this purpose; errors produced by the linear approximations to monotonic nonlinear functions are likely to be unimportant relative to test-retest unreliability, interrespondent differences, and the like." (p. 156)

At the completion of step 8, all entities (alternatives) have been located on the relevant value dimensions and the location measures have been rescaled. Therefore, there will be a number from 0-to-100 in each cell of the matrix.

Before examining a completed matrix, it might be helpful to the reader to go through some of the thinking that generates the numbers inside the cells of the matrix. Each of the teachers' values will be listed with a brief discussion of a possible choice of instrumentation and its scoring procedure.

Dropout prevention: This would be a purely objective dimension. We know our expected drop out rate in the traditional setting. We would calculate the actual drop out rate and compared the two by way of a proportion.

$$\frac{\text{Actual}}{\text{Expected}} = \frac{\text{location score}}{100} \qquad \text{Eg. } \frac{25 \text{ actual}}{50 \text{ expected}} = \frac{50}{100}$$

50 would be entered into that cell in matrix.

Self-worth: Here one might use multiple measures. Standardized self-concept tests might be used. These would have to be compared with the expected scores attained in the traditional setting. Then, as above, the proportion would be rescaled on a 0-to-100 scale. Another powerful measure of self-concept would be the use of interviews with the students, teachers, and parents. Many examples can be found in Roberts (1975) chap. 7 but two examples of parent responses are:

"My son feels very good." "My son is always talking about school."

Basic skills: As in self-worth, standardized tests are one way to determine whether the storefront school is succeeding in teaching basic skills. Work samples and teacher interviews would also be useful. Parent and student interviews might also be relevant. Data will be averaged and put into a 0-to-100 scale for inclusion in the matrix.

Level of career aspiration: This area might need a longitudinal approach with close monitoring. Comparisons could be done within the group (at the beginning and end of the child's experience in the various schools) and between the groups in determining the value between 0 and 100 to put into the three cells of the matrix.

Social responsibility and dependability: Behavioral measures would be useful here. Since the students in the storefront school are out in the community (career opportunities, apprenticeships, working in politics), and assuming administrative and teaching roles within the school, it would be profitable to interview or send questionnaires to the adults working with the students. Criteria might include: tardiness, attendance record, etc. Comparisons might be made against a perfect performance or against the traditional school in filling in the three cells with numbers from 0-to-100.

Problem solving and decision making: Multi-measures would again be useful here, i.e., standardized tests and on-the-job experiences. As above, these would be rescaled so as to be averaged and then included in the cells with 0-to-100 scores.

Pride in work: Standardized tests would not be relevant here. We could look at work samples and interview the students and parents about the students' attitudes toward that work. Comparisons could be made with the traditional setting. Eg. "Teachers care about you because if you don't do your work they 'lean on you.'...They go over your work with you..."

Aesthetics and Leisure Time: Interviews, questionnaires, and logs might be useful in ascertaining whether the students in the storefront school have a different aesthetic appreciation and/or use of leisure time from those in the more traditional settings. These will be scaled from 0-to-100.

At this point, it might be useful to discuss the subject of unintended or unanticipated outcomes. According to Finkelstein and Pollack-Schloss (Chap. 7 in Roberts), the storefront-type of school is not without some costs. "There is a certain degree of role confusion among students, teachers, and administrators... There is some adult hesitancy in setting standards with fear on the part of everyone that the program will be misunderstood and terminated... This in turn produces a high degree of defensiveness which inhibits programmatic self-examination and learning." (p. 83) (An even longer list of positive unanticipated outcomes could also be listed here.) These unexpected outcomes could be incorporated into the matrix if the various interest groups felt that they were important in making a determination of whether to continue the storefront school, i.e., if they considered them to be important value dimensions. It should also be pointed out that as Edwards et.al. claim, the distinction between summative and formative is no longer a meaningful one when using the D-T approach. At every point in time, the data that can be gleaned from the matrix can be used both summatively and formatively. Examples of some hypothetical numbers placed into the cells will be found in Table IV.

Step 9. In this step, utilities will be calculated for each entity. Table IV illustrates this procedure by including two numbers in every cell of the matrix. The first is a number of 0 to 100 which represents the degree to which each entity succeeds on each value dimension. (This is the

output of step 8) The second number in each cell is the product of the first number and the normalized importance weight of each value at the top of each column. These products are then summed across each row.[7,8]

Table IV

Hypothetical Numerical Values in Completed Matrix[9]

Program	Dropout prevent. 22.98	Self-worth 22.98	Basic skill 22.98	Career aspir. 11.49	Social respon. 6.89	Prob. Solv. 6.89	Pride in wk. 3.44	Aesthetics & leisure 2.29	TOTAL
STOREFRONT	(50) 1149	(85) 1953.3	(85) 1953.3	(90) 1034.1	(80) 551.2	(85) 585.65	(95) 326.8	(75) 171.75	7725.1
MODIFIED SCHOOL PROGRAM	(15) 344.7	(30) 698.4	(55) 1263.9	(45) 517.05	(20) 137.8	(40) 206.7	(30) 103.2	(15) 34.35	3297.1
TRADITIONAL SCHOOL	(0) 0	(15) 344.7	(40) 919.2	(40) 459.6	(5) 34.45	(20) 137.8	(10) 34.4	(10) 22.9	1953.05

Step 10. The final step consists of making the decision. If a single alternative is to be chosen, one might look at the totals at the right of each row in Table IV and select the program which has the highest total.

---

[7] The formula for a weighted average. $U_i = \Sigma_i w_j u_{ij}$ $\Sigma_j w_j = 100$

w_j = normalized importance weight for jth dimension (output of step 7)
u_{ij} = rescaled position of ith entity of jth dimension (output of step 8)
The utility for a given entity is proportional to the sum of the probabilities, each multiplied by the appropriate importance weight.

[8] It should be mentioned that utility scores are often useful under non-experimental conditions, i.e., with no control group or randomization. Furthermore, Murphy (1974) has suggested that comparing utilities at different times (Priors and Posteriors) may highlight ways in which the program is not performing as expected. Disparity between the wholly subjective priors and the data-based posteriors could indicate that the program should be modified or that additional research effects might be required. In particular, such analysis may show that a program should be modified to better meet the needs of a specific subgroup of clients." (Edwards, et.al., p. 49).

[9] This hypothetical illustration does not illustrate the use of Bayesian statistics. For an example of Bayesian revisions, see Edwards, et al. (1975, pp. 175-177)

In the above example it is very obvious that we would choose to continue the storefront school. Its total is more than twice that of the modified school's program and almost four times more than the traditional school's program, (i.e., 7725.1 vs. 3297.1 vs. 1953.1). But, it should also be noted that much more than a single highest sum could be derived from a decision-theoretic or M.A.U.T. analysis. One could do many subanalyses as well. If one wanted to know which program maximized a particular value, one need only look down the column which measured that value across the different programs. One could also use it to see where future efforts are needed. For example, one might try to improve the 50 under drop out prevention to more closely reach 100. It should also be noted that because different groups generate different matrices, it is possible for one program to be the most successful for one group and a different program be the most successful for a different interest group. The above matrix was generated from the teachers' values. It is conceivable, though not likely, that the traditional school's program might come out highest if one considers the values of a different subgroup. (see Table I)

In concluding, let us turn to Renzulli's remarks on the five essential ingredients of a well-executed evaluation: (p. 5)

1. To discover whether and how effectively the objectives of a program are being fulfilled.
2. To discover unplanned and unexpected consequences that are resulting from particular program practices.
3. To determine the underlying policies and related activities that contribute to success or failure in particular areas.
4. To provide continuous in-process feedback at intermediate stages throughout the course of a program.
5. To suggest realistic, as well as ideal, alternative courses of action for program modification.

Implications for Summative and Formative Evaluations and Post-Assessment Organizational Change

Edwards et al. (1973) prefer a planning orientation to that of summative and formative evaluations. However, because most evaluators are familiar with the distinction between the two, they will be discussed in turn. In a summative evaluation using the M.A.U.T. - Bayesian model the decision maker will select the entity with the highest utility. This assumes that either the decision maker finds the information consistent with his prior beliefs or that the evaluator has collected sufficient evidence to be convincing in overwhelming his original prior judgments. It may be necessary to collect additional data in order to dispel ambivalence. It may have occurred to the reader that the other important prerequisite for a useful evaluation is that the decision maker be honest concerning his goals and values. If there are hidden agendae and extraneous political pressures influencing the decision, these may render the evaluation inappropriate. These limitations are not limitations of the model; they are, in contrast, real limitations existing in the world in which evaluations are conducted. One might ask whether the traditional mode of evaluation which addresses only program effectiveness is better suited to these limitations. The M.A.U.T. - Bayesians think not. Therefore, it is quite important to determine early in the evaluation process whether a real decision is at stake and whether the decision maker is honestly communicating his goals and priority values. To the extent that these criteria are violated, the evaluation may become a charade.

The goal of formative evaluation is program improvement. In spirit, the M.A.U.T. - Bayesian model closely resembles the model of evaluation called for in Rossi and Cronbach, (1976, p. 18).

> "(1) Evaluation can constructively enter the picture earlier and can be seen as a continuing part of management rather than as a short-term consulting contract. (2) The evaluator, instead of running alongside the train making notes through the windows, can

board the train and influence the engineer, the conductor, and the passengers. (3) The evaluator need not limit his concerns to objectives stated in advance; instead, he can also function as a naturalistic observer whose inquiries grow out of his observations. (4) The evaluator should not concentrate on outcomes; ultimately, it may prove more profitable to study just what was delivered and how people interacted during the treatment process. (5) The evaluator should recognize (and act upon the recognition) that systems are rarely influenced by reports received through the mail. Evaluation thus becomes a component of the evolving program itself, rather than disinterested monitoring undertaken to provide ammunition to the warring factions in a political struggle. Formal reports to outsiders are reduced in significance and research findings become not conclusions, but updating of the system's picture of itself." (p. 18)

The picture of the active participant evaluator drawn by Ross and Cronbach is perfectly consistent with the M.A.U.T. - Bayesian model. However, in order for the above approach to be possible the decision-maker must subscribe to it. To the extent to which the decision-maker is truly open to modifications and improvement and restructuring his program, the M.A.U.T. - Bayesian evaluator can be helpful.

It should be noted that the ways in which a program(s) can be improved often become evident as soon as the goals are listed. Before any data are collected, it may become obvious that unless a particular aspect of the program is modified, there is little or no chance of succeeding on a particular goal. Often, the weaknesses of a program will emerge when one asks the program director for his prior estimates of how well the program is likely to succeed on each dimension.[10]

---

[10] How to phrase the questions when determining prior probabilities is an area ripe for research. One must be careful to ask for the priors commensurate with the time at which the data is to be collected. If priors about the ultimate success of the program are compared with data collected at the beginning of the program there will be unnecessary and possible misleading discrepancies. Edwards, et al. (1963) have an important section on priors, but they do not discuss this issue. Two techniques for generating priors may be found in Raiffa, 1968, (pp. 161-166) and Novick and Jackson, 1974, (pp. 160-166). The Raiffa approach uses fractiles and the Novick and Jackson approach uses a computer-assisted interrogation concerning sample size determination.

If prior estimates are either unrealistic or dishonest, or if the instruments chosen are insensitive, the earliest collected data will make the discrepancy obvious. Certainly, at that point, which is hopefully still early in the evaluation process, the decision-maker and/or program developer may begin to address the discrepancies. It is possible that the prior likelihoods were accurate and new measures are necessary; it is also possible that programmatic changes are needed. The M.A.U.T. matrix will enable the decision-makers to see that the program is succeeding better on certain dimensions than on others and the program can be improved sequentially as the weaknesses are discovered. New priors will be formed from the posteriors and new data will indicate the extent to which the program has been improved. As mentioned above, ambivalence on the part of the decision-maker may make more data collection necessary. The M.A.U.T. - Bayesian approach intuitively conforms to the way in which decisions are made.

Implications for Organizational Change

It has been felt by those using the M.A.U.T. - Bayesian model that post-assessment organizational change occurs as a result of the use of the model. Edwards, et al. (1975) discusses the social psychology of the process.

> "First, each group builds a consensus about its own values vis-a-vis the programs. This makes it possible to exchange information about the relative ordering of values between groups, so that discussions between groups about value differences can be quite explicit and quantified....Second, the same evaluation data can be fed back to each group. The same data, in a matrix in which values, rank order, and/or importance weights differ considerably, will yield very different final conclusions and decisions. Thus, a number of groups can, using the same data, come to very different conclusions about whether a program or programs are meeting their goals. This then provides them with a substantive basic for discussions with one another...In addition it means that decision-makers receive research data on issues that may be foreign to their own values but quite germane to the values of other groups, for example, persons affected by a program. (pp 171-173)"

The M.A.U.T. - Bayesian model involves the various interest groups in

generating values, prioritizing them, giving prior probability estimates, etc. These procedures force the groups to confront their values. For groups which have not yet done this, the process can be very illuminating. The process of the evaluation may potentially open new channels of communication both within groups who are working to achieve consensus and between groups who strive to understand how the different values they bring to bear, affect the outcomes of the program. Furthermore, it leaves in its wake a strategy of decision-making which could be applied within the organization of future planning and development.[11,12]

Implications Summary

The appeal of the M.A.U.T. - Bayesian model is in its flexibility and its fit to the way in which people actually make decisions. According to Ross and Cronbach, (1976, p. 14) the various purposes of evaluation have been defined as, "to assess needs, to guide a "go/no-go" decision, to provide support for a decision already made, to improve program plans and policies, to assist management by monitoring daily operations, to test social theories—all imply different criteria for excellence in evaluation and different, often contradictory, research tactics." In theory, at least, the M.A.U.T. - Bayesian model

---

[11] Keeney and Raiffa (1972) wrote of decision analysis: That "it serves as a learning experience for the participants. By virtue of explicitly examining many of the difficult issues of a particular problem, their abilities to think systematically about complex aspects of public problems will likely improve. In addition, the mathematical reasoning, measurement techniques, and general approach to problem solving might be transferable to different areas of application." (p. 70)

[12] The author (1977) recently used the M.A.U.T. model as a needs assessment and planning model. After generating separate matrices for teachers, superintendents, and principals; representatives from the various groups met to discuss each other's values and priorities and build programs accordingly. It is further suggested that the various groups have a greater sense of ownership in the resulting program after contributing in this way.

can be tailored to fit any of these purposes. Whether this will prove true in practice remains to be seen. This area is young and the number of evaluations which have used this approach is small. It is acknowledged that there are still many questions both theoretical and procedural which need refinement.[13] As each new problem is isolated, the solution results in a stronger technology. The M.A.U.T. - Bayesian framework encourages and stimulates creative problem solving and decision-making for the evaluator as well as the decision-maker.

---

[13] Keeney and Raiffa (1972) wrote of the methodology of a decision analysis what can be directly applied to the M.A.U.T. - Bayesian model: "Although there clearly needs to be a great amount of significant work done...we feel the techniques and procedures that are currently available are sufficiently developed to be an important aid to the decision-maker...It is important to accumulate critical experiences with the use of those techniques on societal problems. And if this effort is to make any sense at all, it is imperative that public officials and members of their staffs begin to use formal analysis on projects of importance to them. The difficulty of such efforts, as well as their possible benefits, should not be underestimated. Often the total value of such analyses is not immediately apparent but rather accrues over time as successive analyses improve in quality and relevancy and as people learn how to interpret and implement such efforts better. We should not become disillusioned if initial attempts are somewhat feeble; the achievement of quality is an evolutionary process. Thus, we believe it is important to start doing, documenting and critically reviewing these attempts with the spirit of learning. How can it be done better next time." (pp. 71-73)

Appendix A*

BAYESIAN STATISTICAL INFERENCE

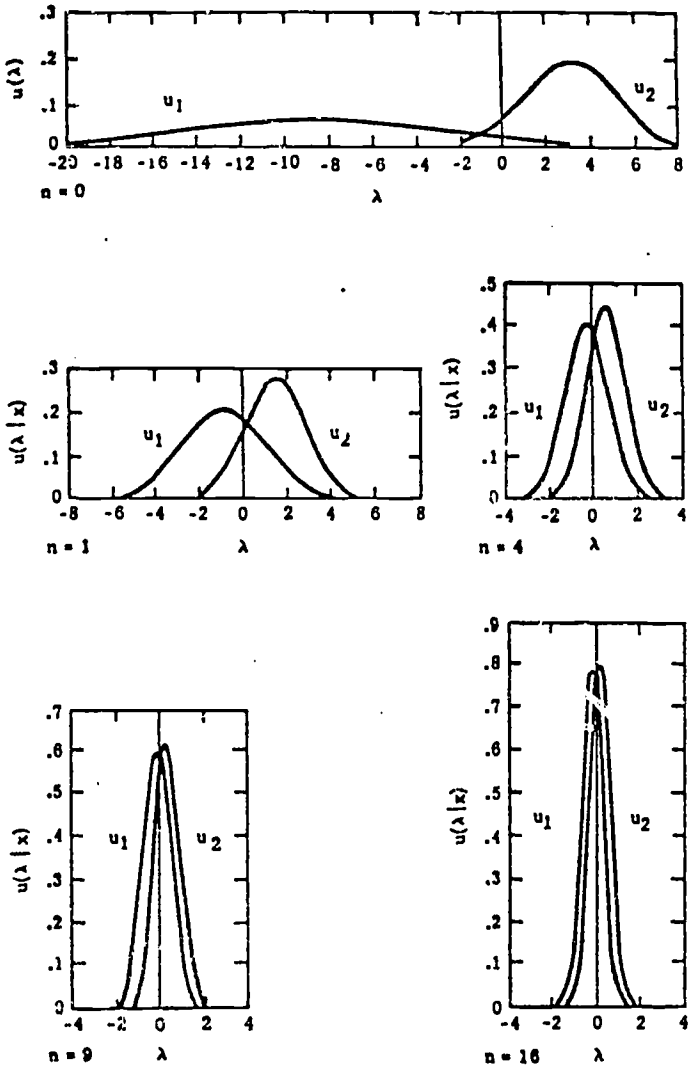

FIG. 2. Posterior distributions obtained from two normal priors after n normally distributed observations.

To illustrate both the extent to which the prior distribution can be irrelevant and the rapid narrowing of the posterior distribution as the result of a few normal observations, consider Figure 2. The top section of the figure shows two prior distributions, one with mean -9 and standard deviation 6 and the other with mean 3 and standard deviation 2. The other four sections show posterior distributions obtained by applying Bayes' theorem to these two priors after samples of size n are taken from a distribution with mean 0 and standard deviation 2. The samples are artificially selected to have exactly the mean 0. After 9, and still more after 16, observations, these markedly different prior distributions have led to almost indistinguishable posterior distributions.

*Edwards, 1963, (pp. 210-12)

References

Baron, J. "A Decision-Theoretic Approach to Evaluation Research: An Explication and Application." Paper presented to the Northeastern Educational Research Associations Seventh Annual Convention, October, 1976, Ellenville, New York.

Baron, J. and Martin, P. The M.A.U.T. - Model: An Application to Needs Assessment (unpublished manuscript).

Campbell, D. T. and Stanley, J. C. Experimental and Quasi-experimental Designs for Research. Chicago: Rand McNally, 1966.

Cook, T. D. and Campbell, D. T. The design and conduct of quasi-experiments and true experiments in field settings. In Dunnette, M. D. Handbook of Industrial and Organizational Research. New York: Rand-McNally, 197[illegible].

Edwards, W., Lindman, H. & Savage, Leo. Bayesian statistical inference for psychological research. Psychological Review, 1963, 70. (3) 193-242.

Edwards, W., Guttentag, Marcia and Snapper, Kurt. "A Decision-Theoretic Approach to Evaluation Research" in Guttentag, M. & Struening, E. L. Handbook of Evaluation Research. Sage Publications, 1975.

Guttentag, M. "Subjectivity and its use in evaluation research." Evaluation, Vol, 1, No. 2, 1973.

Guttentag, M. and Snapper, K. J. "Plans, evaluations, and decisions." Evaluation. in press.

Iwanicki, Edward F. "Some Considerations in the Development of a Secondary School Evaluation Program" (submitted for publication.)

Keeney, R. L. and Raiffa, H. A critique of formal analysis in public decision making in Drake, A., Keeney, R. L. and Morse, P. (Editors) Analysis of Public Systems. Cambridge, MA: The Massachusetts Institute of Technology Press, 197[illegible].

Novick, M. R. and Jackson, P. H. Statistical methods for educational and psychological research. New York: McGraw Hill Book Company, 1974.

Raiffa, H. Decision Analysis: introductory lectures on choices under uncertainty. Reading, MA: Addison-Wesley, 1968.

Renzulli, Joseph S. A Guidebook for Evaluating Programs for the Gifted and Talented. (working draft) Office of the Ventura County Superintendent of Schools. 1975.

Roberts, Arthur D. Educational Innovation: Alternatives in Curriculum and Instruction. Allyn and Bacon, Inc., Boston, London, Sydney. 1975.

References (Continued)

Ross, L. & Cronbach, L. (Eds.) Handbook of evaluation research. Essay review by task force of the Stanford Evaluation Consortium. Educational Researcher, 1976, 5 (10), 9-19.

Winkler, R. L. Introduction to bayesian inference and decision. New York: Holt, Rinehart & Winston, 1972.