EVALUATION OF COMPLEX STIMULI USING MULTI-ATTRIBUTE UTILITY PROCEDURES

Detlof v. Winterfeldt, et al

Michigan University

Prepared for:

Office of Naval Research
Advanced Research Projects Agency

29 October 1973

that MAUT can improve upon the decision maker's own unaided intuition. The probabilistic procedure was found to be the superior method for predicting simple choices between stimuli.

EVALUATION OF COMPLEX STIMULI USING MULTI-ATTRIBUTE UTILITY PROCEDURES[1]

Technical Report

29 October 1973

Detlof v. Winterfeldt and Ward Edwards

Engineering Psychology Laboratory

The University of Michigan

Ann Arbor, Michigan

This research was supported by the Advanced Research Projects Agency of the Department of Defense and was monitored by the Engineering Psychology Programs, Office of Naval Research under Contract No. N00014-67-A-0181-0049, Work Unit Number NR. 197-021.

A natural and intuitively appealing approach to evaluating multi-attributed choice alternatives such as used cars or job offers is to list the dimensions on which the alternatives vary, to describe each alternative by its states in those dimensions, and then to think about trade-offs between states of different dimensions. This evaluation process is formalized by a number of measurement theoretic models and scaling procedures which are known under the name of multi-attribute utility theory (MAUT).

Additive evaluation models are the best known and most structured MAUT models. They allow a decomposition of the complex evaluation task into simple judgments :

a) determining the value-dimensions,

b) judging the importance of each dimension,

c) assessing utility functions within dimensions.

The overall utility U of an evaluation object $\underline{x}$ can then be determined by

$$U(\underline{x}) = \sum_{i=1}^{n} w_i u_i(\underline{x})$$

where w_i is the importance weight of dimension i and $u_i(\underline{x})$ is the utility of $\underline{x}$ in dimension i.

Additive MAUT models have been analyzed in a variety of experimental and applied evaluation situations such as the evaluation of cars, jobs, graduate students, water quality, and social programs (for a summary, see Fischer, 1972a). Their descriptive validity as models of the actual evaluation process has been established by comparing the decomposition procedure with more direct methods of assessing the worth of an evaluation object like overall ratings, rankings, or pair comparisons. Such a convergent validity approach indicates that additive models approximate intuitive judgments rather well. Correlation

coefficients between model predictions and direct judgments were typically in the mid .80s and higher (Pollack, 1964; Yntema and Klem, 1965; Hoepfl and Huber, 1970; Huber, Daneshgar, and Ford, 1971; Pai, Gustafson, and Kiner, 1971; Fischer, 1972b).

Attempts to determine the normative validity of MAUT as a tool for decision aiding have been less successful. The main problem is the lack of an appropriate validation criterion. The decomposition approach is most useful where unaided intuition fails as a guide for rational decisions. It is well known, for example, that an increase in the number of attributes leads to inconsistent overall judgments and unsatisfactory simplistic strategies like lexicographic orderings or satisficing (see Tversky, 1969; Slovic and Lichtenstein, 1971). In such cases a simple convergent validity approach is clearly inappropriate.

Yntema and Torgerson (1961) tried to circumvent this problem by creating an artificial experimental environment. Their stimuli were ellipses which varied on the three dimensions size, shape, and color. A non-additive worth function which increased monotonically with the values in each dimension was arbitrarily defined by the experimenters. Subjects learned to evaluate the ellipses by making judgments about their worth and then receiving feedback from the externally defined worth function. After training the correlation of the subjects' worth estimates with the external criterion function was .84. An additive decomposition approach resulted in a higher correlation of .89. Apparently an additive decomposition procedure can recover the main aspects of a learned evaluation function, even if this function is non-additive. Since all subjective evaluation strategies can be assumed to be learned in a similar way, Yntema and Torgerson's experiment provides evidence for the usefulness of additive MAUT models for decision aiding.

The present experiment tried to demonstrate the validity and usefulness of additive MAUT models and procedures in a more realistic context employing yet another attack to the validation problem. An apartment evaluation problem was analyzed in which decision aiding was clearly necessary because of the large number of dimensions involved. After the construction of utility functions and importance weights subjects were trained to consistently apply an evaluation strategy with which they felt most comfortable and which reflected their preferences best. The correlation between model predictions and complex judgments after training constituted a general measure of the usefulness of MAUT. Three procedures to assess utility functions and importance weights were tested for their differential validity beyond differences in the correlation with post-training judgments. Descriptions of apartments were constructed from weights and utility functions such that the three procedures would predict different preferences in simple choices.

The three methods of assessing importance weights and utility functions correspond to three distinct approaches to additive multi-attribute utility theory :

1.) A probabilistic procedure, described by Raiffa (1969), which is an outgrowth of additive expected utility theory (see also Fishburn, 1970). This procedure has been compared with others in only one experiment so far (Fischer, 1972b).

2.) A direct rating procedure with importance weights derived from the unstandardized utility functions as described by Sayeki (1971) in the framework of additive conjoint measurement. This procedure has not yet been studied.

3.) The same direct rating procedure as in 2. but with directly assessed weights, similar to the procedures suggested by Edwards (1971) and Hoepfl and Huber (1969). Direct rating procedures have been applied in one form or another quite frequently in multi-attribute utility experiments.

Method

Assessment procedures

In the probabilistic procedure (in the text denoted by 'P') weights are assessed as follows. Let $\underline{x}^*$ be the best object under consideration, $\underline{x}_*$ the worst. Let $\underline{x}_i^*$ have the best state in dimension i, the worst in all other dimensions. The importance weight w_i is defined as the probability which makes the decision maker indifferent between receiving $\underline{x}_i^*$ for sure and the option to play the gamble $(\underline{x}^*, w_i, \underline{x}_*)$ in which he receives $\underline{x}^*$ with probability w_i and $\underline{x}_*$ with probability $1-w_i$. Utility functions are constructed in a similar way. Let $\underline{x}_i^j$ be the object with the worst states in all but the i-th dimension, in which it attains state j. Then the utility of object $\underline{x}_i^j$ is defined as the probability u_{ij} which makes the decision maker indifferent between receiving $\underline{x}_i^j$ for sure and the gamble $(\underline{x}_i^*, u_{ij}, \underline{x}_*)$. Probabilistic utility assessments should lead to the same results if the constant states in the dimensions $k \neq i$ are not the worst states.

In the second procedure (in the text denoted by 'U') single dimension utility functions f_i are obtained from overall ratings of evaluation objects which vary only in dimension i. The endpoints of the overall rating scales are defined by a very bad and a very good object. The importance of dimension i is determined by observing how much the overall ratings change when the dimension states are changed from the worst to the best. Formally, imortance weights are defined as

$$w_i = [f_i(\underline{x}_i^*) - f_i(\underline{x}_*)] / \sum_{i=1}^{n} [f_i(\underline{x}_i^*) - f_i(\underline{x}_*)]$$

The third procedure (D) differs from U only in that the importance weights w_i' are directly rated. These weights are normalized according to

$$w_i = w_i' / \sum_{i=1}^{n} w_i'$$

Validation methods

The correlational validation of the three multi-attribute utility functions differed from the usual convergent validity approach only in the nature of the criterion function and needs no further explanation here. The differential validation of the three procedures should be explained in more detail.

For each procedure, subject and dimension, importance weights were assessed:

wp_i the weight of dimension i from the probabilistic procedure,

wu_i the weight of dimension i derived from the unstandardized utility function,

wd_i the normalized weight of dimension i from the direct rating method.

Since utility functions were the same for procedures D and U, only two utility functions were constructed for each subject and each dimension.

up_i the utility function of dimension i from the probabilistic pro-procedure,

ud_i the standardized utility function from the direct rating procedure.

From weights and utility functions three weighted utility functions were constructed for each dimension:

$$u_{1,i} = wp_i.up_i; \quad u_{2,i} = wu_i.ud_i; \quad u_{3,i} = wd_i.ud_i$$

For non-controversial dimensions (nc) these weighted utility functions were hardly distinguishable, for the more controversial dimensions (c), they differed substantially. Figure 1 gives a hypothetical example. By inspection of the graphs two states of the controversial dimension were picked, say x_c and y_c. Then two states of the non controversial dimension, x_{nc} and y_{nc}, were chosen

- - - - - - - - - - - - - - -

Insert Figure 1 about here

- - - - - - - - - - - - - - -

such that, for example, the following inequalities held:

$$u_{1,c}(\underline{x})+u_{1,nc}(\underline{x})>u_{1,c}(\underline{y})+u_{1,nc}(\underline{y})$$

$$u_{2,c}(\underline{x})+u_{2,nc}(\underline{x})<u_{2,c}(\underline{y})+u_{2,nc}(\underline{y}),$$

where $\underline{x}$ and $\underline{y}$ are objects with the same states in all dimensions except c and nc, where $\underline{x}$ attains x_c and x_{nc} and $\underline{y}$ attains y_c and y_{nc}. The additive model would predict in this case that - no matter what the constant states in the other dimensions - object $\underline{x}$ would be chosen according to the probabilistic procedure and object $\underline{y}$ according to the direct rating procedure. The validity of the different procedures could be checked in this way by determining the percentage of correct predictions for each procedure.

Subjects and material

Three students helped listing the dimensions and served as pilot subjects. Twenty students gave importance ratings of the raw dimension list, and four paid, single, undergraduate students took part in the actual experiment.

Apartments for the overall evaluation were described by the values in the single dimensions on a card. The order of dimensions was randomized to prevent Ss from attending to only a few dimensions. The value for each dimension was determined by E to make sure that the descriptions were realistic and that the apartments covered a wide quality range. Rent was excluded from the descriptions since quality alone was to be judged. Twenty apartments were constructed for the intuitive evaluation before training (set I). Twenty apartments for the evaluation after training (set II) were constructed by exchanging values of apartments which

had been rated as similar in the first evaluation. The judgmental task before and after training therefore had the same difficulty without repeating sets and without allowing for learning effects.

The rating scales were simple 9 cm lines without endpoints and without numerical segmentation. The extremes were described as 'very good offer' vs. 'very bad offer' in the case of utility judgments and as 'very important' vs. 'not important' in the case of importance judgments. The apparatus for the probabilistic procedure was a gambling device in which a spinning wheel could be turned. It would stop on one of two colored segments which could be adjusted in size to represent events. Probabilities were also displayed (see Figure 2).

- - - - - - - - - - - - - - -

Insert Figure 2 about here

- - - - - - - - - - - - - - -

Experimental procedure

In a kind of brainstorming session three Ss and E listed all attributes which might determine the value of apartments to single undergraduate students. Sixty-four attributes were found and logically structured into 34 dimensions. Twenty students classified these 34 dimensions into one of four importance categories. Fourteen dimensions were selected which had a median ranking in the first two categories (see Table 1).

- - - - - - - - - - - - - - -

Insert Table 1 about here

- - - - - - - - - - - - - - -

In the actual experiment all Ss were run individually. Table 2 summarizes the experimental procedure, indicating also the different data analyses done.

- - - - - - - - - - - - - - -

Insert Table 2 about here

- - - - - - - - - - - - - - -

In the first session the four Ss were made familiar with the material and rated the apartments of set I. They were instructed to make their judgments according to how much they would like to live in such an apartment, assuming that enough roommates were available to occupy the apartment on a 1 person per bedroom basis.

During the following meeting Ss ranked the same set of apartments, assessed importance weights to each of the 14 dimensions, and rated the single dimension utilities.

In the next session the notion of a certainty equivalent and the use of the gambling device were introduced by playing gambles for money. With fixed outcomes E adjusted the probabilities in the gamble and asked Ss if they preferred the gamble or a sure outcome. This was done systematically until a range of 5% was determined in which the Ss switched from preferring the gamble to preferring the sure outcome. This idea was pointed out to them and they were asked for the exact probability in that range which would make them feel indifferent. All gambles were played, if chosen, and money changed hands on each trial. The same systematic procedure with apartments as outcomes was then used to scale utilities and importance weights according to the description of the probabilities method given alone.

The following hour Ss were trained in the overall evaluation of apartments. They compared apartments which they had rated as similar before and stated their preference. The reasons they gave were discussed and possible objections were pointed out by E. Then they rated the apartments of set II. After the same time lag as between the first and the second pretraining evaluation they ranked set II.

In the last session each S received 50 pairs of apartments which were constructed from their weights and utility functions to test the differential validity of the procedures. For each pair Ss simply marked the preferred apartment.

Results

Table 3 shows the reliabilities (Spearman's Rho) of the pretraining evaluation (set I). These reliabilities vary between .325 and .817 with an

- - - - - - - - - - - - - - -
Insert Table 3 about here
- - - - - - - - - - - - - - -

average of .561. For the respective judgments after training (set II), the reliabilities were substantially higher with values between .797 and .996 with an average of .893.

The product moment correlations between the utilities derived from the model and the post-training ratings are found for all procedures in Table 4.

- - - - - - - - - - - - - - -
Insert Table 4 about here
- - - - - - - - - - - - - - -

The average values lie between .683 and .726. The differences in these correlations are not large enough to draw any conclusion about which procedure did best. Individual differences are substantial.

It is interesting to note where the procedures produced different results. While differences in the two utility functions were usually very small, the three weight vectors differed substantially. Correlations between weights of the different procedures are shown in Table 5. Only P and U show a moderate agreement.

- - - - - - - - - - - - - - -
Insert Table 5 about here
- - - - - - - - - - - - - - -

Nevertheless, the correlation between the model predictions for sets I and II are high (averaged over Ss between .85 and .89) showing the insensitivity of the additive model against variations in weight parameters.

In the analysis of differential predictions only those cases were analysed in which the differences in the predictions were at least 1 utile (as a comparison: the best apartment had about 90 utiles, the worst about 30). Since the results for different Ss were similar, only the group data are reported. The best pro-

- - - - - - - - - - - - - - -

Insert Table 6 about here

- - - - - - - - - - - - - - -

cedure is the probabilistic procedure which outperforms the direct rating method in 63.4% of the analysed cases. With 58.6% correct predictions the probabilistic procedure is also tendentially better than the direct rating procedure with derived importance weights. Not much can be said about the comparison between U and D since the number of observations is too small (see Table 6).

The data of the differential validation were further analysed with some simple classical and Bayesian hypothesis testing tools. The classical analysis tested the goodness of fit of the data to a theoretical uniform distribution which assumed 50% correct predictions for each procedure in all three comparisons. A χ^2-test revealed that only the data for P vs. D differed significantly from the uniform distribution (see Table 6).

The Bayesian analysis assumed that the data were generated by a Binomial process with an unknown probability p that the procedure favored by the data led to a correct prediction. The Bayesian analogue of the χ^2-test would give a high prior probability to the point p=1/2 and spread the rest of the prior

distribution over other values of p. Since this seemed to be an extremely unrealistic prior distribution, we preferred to work with a uniform distribution over p, i.e. we assumed that all values of p were equally likely. Since a uniform distribution is a special case of a beta-distribution, and since a beta-distribution is conjugate to a binomial data generating process, the posterior distribution is easy to compute as a beta-distribution with the parameters n+1 and y+1, where n is the number of analysed pairs and y is the number of correct predictions of the procedure favored by the data. (For details of this analysis, see Raiffa and Schlaifer, 1961; DeGroot, 1970.) The final step is the computation of the probability that p is larger than 1/2 which is the probability of the hypothesis that the procedure favored by the data has actually a better chance to lead to a correct prediction. These probabilities are found in Table 6.

Discussion

The low pretraining reliabilities demonstrate clearly that the evaluation task was difficult and that some sort of decision aiding was needed. Post-training reliabilities - although far from being perfect - allow some confidence in our criterion function. After training Ss said that they felt much more confident about their judgments and that they reflected their real preferences.

All three procedures predicted post-training judgments to a lower degree than usually found in the literature. However, considering the number of dimensions involved and the unreliability of the criterion, values of .70 are clearly comparable to the higher values found in the literature and they generally confirm the basic results about the validity of MAUT.

Notable are the individual differences. Validities for Ss 3 and 4 are lower than for Ss 1 and 2. In particular, D is much less successful for these two Ss. One possible explanation is that S 3 and S 4 took less care in performing the experimental task. Carelessness could very well have affected D more than U and P, since in P constant consistency checks forced Ss to attend and in U errors and misjudgments could enter only in the utility assessment. According to this hypothesis, S 3 and S 4 should also have lower reliabilities in overall judgments. This is true for S 3 who has the lowest reliabilities in pre- and post-training evaluations. Subject 4, on the other hand, shows the highest and the second highest reliabilities in the respective judgments.

As a natural alternative to the carelessness hypothesis one could claim that Ss 3 and 4 did not follow the additive model in their evaluation. Unfortunately, the data allow no conclusions about these alternatives.

The low pretraining reliabilities lead to a cautious interpretation of the ability of MAUT to improve upon the decision maker's own unaided intuition. Except for S 4 all procedures had validities which are substantially higher than pretraining reliabilities. One can conclude that for these Ss pretraining judgments would have been a worse guide to their decisions than MAUT, since reliabilities determine a natural upper bound for the validity of pretraining judgments.

The results of the differential validation of the three procedures need careful interpretation. They support a ranking: P better than U better than D. This result actually came as a surprise since the highly complex judgmental task in the probabilistic procedure led to the assumption that P could not produce

stable and reasonable decomposed judgments. In the present experiment, however, the probabilistic procedure was very carefully implemented and - although Ss showed some confusion about it in the training phase - in the experimental phase they made their judgments rather securely and with few inconsistencies. The direct preferential choice procedure used here to focus in on the indifference probabilities allowed the experimenter to detect inconsistencies, explain them to the Ss and eliminate them so that the success of the probabilistic procedure might well be the result of forcing the Ss to be more careful about their judgments.

As to the other two procedures, the results are not decisive. Nevertheless, it is of considerable importance for practical reasons that Sayeki's suggested derivation of importance weights did not do much worse than P or D. In fact, the derived weights showed more agreement with the superior probabilistic method than with the directly rated weights. Apparently the Ss could fairly well incorporate the relative importance of the dimensions in their single dimension utility judgments. Weights derived from such judgments can therefore be used to cross check explicit weighting judgments and to uncover inconsistencies in applications of MAUT.

The results of our differential validation indicate that more complex decomposed judgments (like P and U) are not necessarily worse than very simple decomposed judgments.

Referercs

Edwards, W. Social utilities. Paper presented at the Symposium on Decision and Risk Analysis, Americal Society for Engineering Education and American Institute of Industrial Engineers, Annapolis, Maryland, 1971.

Fischer, G.W. Multi-dimensional value assessment for decision making. Technical Report No. 037230-2-T, Engineering Psychology Laboratory, University of Michigan, Ann Arbor, 1972 (a).

Fischer, G.W. Four methods for assessing multi-attribite utilities : an experimental validation. Technical Report No. 037230-6-T, Engineering Psychology Laboratory, University of Michigan, Ann Arbor, 1972 (b).

Fishburn, P.C. Utility theory for decision making. New York : Wiley, 1970.

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

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

Pai, G.K., Gustafson, D.H., and Kiner, G.W. Comparison of three non-risk methods for determining a preference function. University of Wisconsin, 1971.

Pollack, I. Action selection and the Yntema and Torgerson worth function. In Information System Science and Engineering : Proceedings of the First Congress of the Information System Sciences, New York : McGraw-Hill,1964.

Raiffa, H. Preferences for multi-attributed alternatives. Rand Corporation, RM-5868-DOT/RC, April, 1969.

Sayeki, Y. Allocation of importnace: an axiom system. Journal of Mathematical Psychology, 1972, 9, 55-65.

Slovic, P. and Lichtenstein, S. A comparison of bayesian and regression approaches to the study of information processing in human judgment. Organizational Behavior and Human Performance, 1971, 6, 649-744.

Tversky, A. Intransitivities of preferences. Psychological Review, 1969, 76, 31-48.

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

Yntema, D.B. and Torgerson, W.S. Man-computer cooperation in decisions requiring common sense. IRE Transactions on Human Factors in Electronics, 1961, HFE-2, 20-26.

Footnotes

[1]The research reported here was undertaken in the Engineering Psychology Laboratory, Institute of Science and Technology, University of Michigan. The authors thank G. W. Fischer and M. F. O'Connor for helpful comments and criticisms on an earlier draft of this paper.

Table 1

Dimensions for the apartment evaluation (random order)

1. Number of bedrooms
2. Noise form outside and inside the building
3. Size of the living- and dining-room area
4. General cleanliness
5. Distance from campus
6. Furnished - not furnished
7. Type of neighborhood
8. Brightness of rooms
9. Lease - no lease
10. Allowable alterations in apartment
11. Kitchen - kitchenette
12. Closet and storage space
13. Landlord strictness
14. Transportation facilities

Table 2

Summary of the experimental procedure
(the T_i's refer to the tables of results)

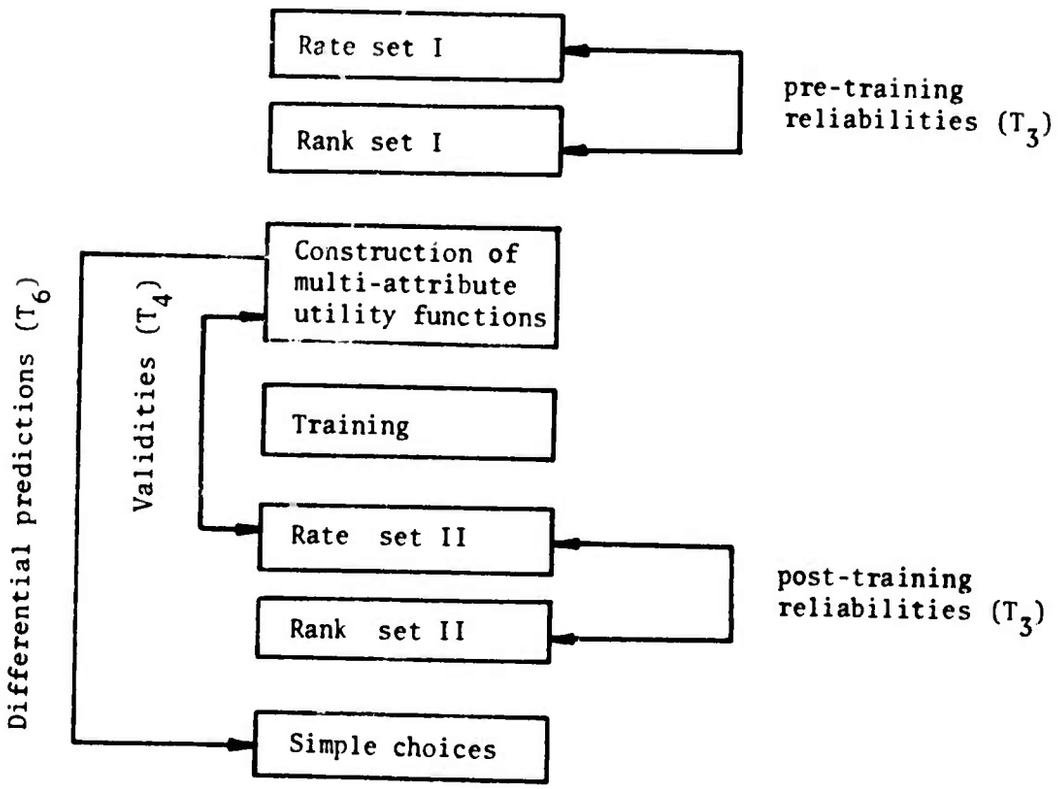

Table 3

Reliabilities in evaluating set I (pre-training) and set II (post-training) (N=20)

set	I	II
subjects		
S1	.633	.996
S2	.467	.840
S3	.325	.797
S4	.817	.933
average	.561	.893

Table 4

Validities of the three procedures (N=20)

set	II		
procedures	P	D	U
subjects			
S1	.812	.819	.753
S2	.745	.809	.726
S3	.697	.488	.652
S4	.681	.492	.640
average	.726	.685	.683

Table 5

Intercorrelations of importance weights between procedures within Ss (N=14)

procedures subjects	P-D	P-U	D-U
S1	.220	.543	.022
S2	.580	.735	.530
S3	.412	.642	.455
S4	.324	.817	.355
average	.394	.699	.365

Table 6

Differential predictions of the three procedures

procedures	P-D	P-U	D-U
analysed number of pairs	82	70	34
number of correct predictions	P : 52 D : 30	P : 41 U : 29	D : 13 U : 21
% correct predictions	P : 63.4% D : 36.6%	P : 58.6% U : 41.4%	D : 38.2% U : 61.8%
χ^2 (with attained significance level)	5.90 (.05)	2.79 (.10)	1.81 (.20)
Posterior probability $Pr(p \geq 1/2)$	.984	.923	.912

Figure 1

Hypothetical example of a controversial and
an uncontroversial dimension for the construction
of differential predictions

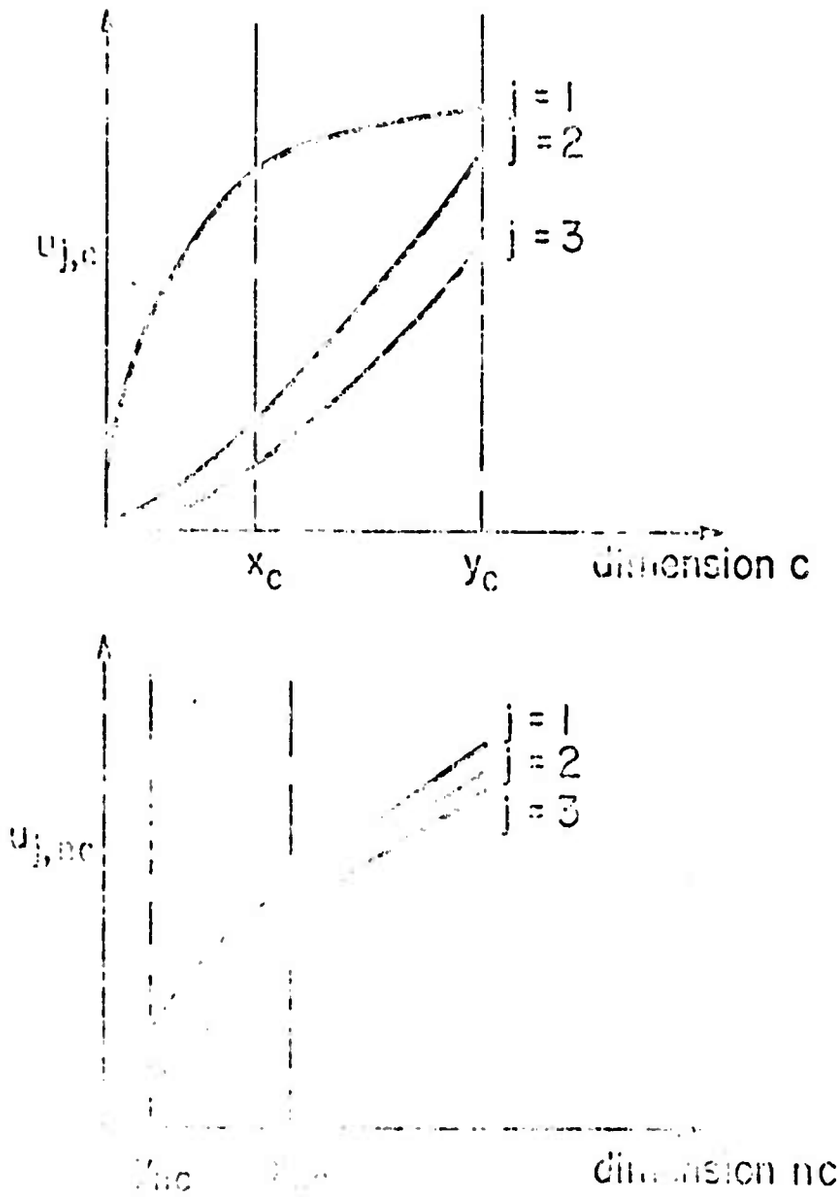

Figure 2

The gambling device for the probabilistic procedure