ORGANIZATIONAL BEHAVIOR AND HUMAN DECISION PROCESSES
Vol. 64, No. 1, October, pp. 9–21, 1995

Numerical vs Cardinal Measurements in Multiattribute Decision Making: How Exact Is Enough?

O. I. LARICHEV

Institute for Systems Analysis, Russian Academy of Sciences

D. L. OLSON

Department of Business Analysis & Research, Texas A&M University

AND

H. M. MOSHKOVICH AND A. J. MECHITOV

Institute for Systems Analysis, Russian Academy of Sciences

Multiattribute decision making can involve consideration of both quantitative and qualitative measures of criteria attaintment. Some decision support systems (decision aids) to help multiattribute decision making quantify value functions. One of the most popular of these systems, multiattribute utility theory (MAUT), requires two types of input. Decision makers need to express the relative value of different attainment levels on each criterion, as well as express the relative importance of these criteria. Some systems (such as DECAID) require simple direct graphical input of value and criterion importance. Other systems (such as LOGICAL DECISION) use more complex means of expressing relative value. Either way, MAUT converts expressions of criterion importance into quantitative form. This study compares the relative stability of numerical results obtained through two decision support systems, DECAID and LOGICAL DECISION (LD), used in the task of evaluation of multiattribute alternatives. Additionally the relative stability of results was measured by comparison with results obtained using an ordinal method, ZAPROS. ZAPROS is a decision support system for construction of a partial order over the set of alternatives. It does not require conversion of qualitative measures into quantitative form. The relations among alternatives are close to those based on ordinal dominance. The results of experiments show that ordinal relationships between task parameters are much more stable than those obtained from quantitative measures. Results from DECAID and LD are much less coincident with each other than with results obtained through ZAPROS. Many inconsistencies were found in subject responses. It is concluded that more attention should be given to the means of testing judgment consistency, and that in some cases, attempts to solve decision tasks through more "exact" judgments of value function parameters may lead to erroneous results. © 1995 Academic Press, Inc.

INTRODUCTION

There have been many papers that have compared different methods and systems for decision making under conditions of multiple criteria (Timmermans, 1991; Buede and Choisser, 1992; Larichev, Moshkovich, Mechitov, & Olson, 1993; Olson, 1992). A number of different indices have been proposed for comparison of multicriteria systems (Rohrmann, 1986; Timmermans, 1991; and others). In spite of the wide variance in these approaches, we think that there is a consensus that one of the most important criteria for evaluation of a decision method (or system) is obtaining the "right" decision. By right decision, we mean choosing the truly most preferred alternative or obtaining the true rank order of alternatives, true being defined relative to the decision maker's preference function. However, identifying the right decision is very difficult. Usually the right decision is considered to be that choice reached by subjects after a holistic assessment of available alternatives. However, this is not very logical, as all decision methods appear to help decision makers solve tasks that these same decision makers do not easily

This research is partly supported by NSF Grant DPP9213392. Address reprint requests to David L. Olson, Department of Business Analysis and Research, Texas A&M University, College Station, TX 77843-4217.

solve without such methods. For many multicriteria decision tasks, there is no objectively obvious best decision. The preferability of the selected alternative is dependent on the individual preference system of a decision maker, and this system, as a rule, is implicit and has no exact description. Furthermore, the majority of existing methods and systems focus on procedures to elicit some fragments of this individual preference system in order to identify some solution to a specific task. This is true of MAUD (Humphreys & McFadden, 1980), EXPERT CHOICE (implementing AHP, Forman, 1992), LOGICAL DECISION (Smith & Speiser, 1991), and so on.

The process of eliciting information about the decision maker's preference structure varies across methods. This elicitation can be rather complex. The decision maker is supposed to compare and/or evaluate relative criteria importance (using verbal and/or numerical ratings on criterion scales), stating the relative attainment of alternatives on each of the criteria, giving probabilistic assessments of outcomes, and so on. Many papers have discussed human errors, biases, lack of comprehension, and inconsistencies in this process (Edwards, 1983; Montgomery, 1977; Montgomery, Garling, Linberg, & Selart, 1990; Payne, 1976; Shoemaker & Waid, 1982). Thus, information received from a decision maker in the process of solving a multicriteria task may include inaccuracy, and we would have little assurance of obtaining the right result.

TASK FORMULATION

One of the most popular approaches in the area of multiple criteria decision making is multiattribute utility theory (MAUT). Many authors (Humphreys & McFadden, 1980; Keeney and Raiffa, 1976; Keeney, 1992) note that people facing difficult decision problems appreciate MAUT because of its systematic analysis of the decision task. Nevertheless, Timmermans (1991) noticed that despite the rather large number of comparison studies using MAUT, it is difficult to reach definite conclusions, largely because of the impossibility of evaluating the quality of the resulting decision.

In this work we describe the results of experiments using two decision support systems based on MAUT. These systems are LOGICAL DECISION (Smith & Speiser, 1991) and DECAID (Pitz, 1987). Both systems support tasks involving risky outcomes, as well as multicriteria tasks under conditions of certainty. In our experiment, we used the latter situation. Both systems were used to apply additive value functions as a means to reflect decision maker preferences.

Let there be Q criteria, upon which N alternatives are evaluated. Each alternative a_i $(i = 1,\cdot \cdot,N)$ corresponds to the vector $a_i = (a_{i1},a_{i2},\cdot \cdot,a_{iQ})$. The decision context was for a college graduate selecting a job offer from five available opportunities. Each alternative was acceptable, but one alternative was better on one aspect, while relatively weaker on other aspects. The subjects were college students nearing graduation, who were in the job search process, facing opportunities similar to those given in the study. Four criteria are used as the focus for the study: SALARY, JOB LOCATION, JOB POSITION (type of work involved), and PROSPECTS (career development and promotion opportunities). The following alternatives were used:

Firm	Salary	Job location	Position	Prospects
a1	$30,000	Very attractive	Good enough	Moderate
a2	$35,000	Unattractive	Almost ideal	Moderate
a3	$40,000	Adequate	Good enough	Almost none
a4	$35,000	Adequate	Not appropriate	Good
a5	$40,000	Unattractive	Good enough	Moderate

There were three possible values on each criterion. The greater the salary, the more attractive it would be to a rational subject. There were four criteria with three possible values each, and the values on each criterion could be rank-ordered from the most to the least preferable (see Appendix 1). There were no dominated alternatives. Therefore, comparison of these alternatives required some value function, which would consider the advantages and disadvantages of each alternative on each criterion.

Both decision support systems LOGICAL DECISION and DECAID were used for this task. Both systems are easy to use, with flexible dialogues and graphical tools to assist in the elicitation of decision maker preferences. Both of these systems implement ideas of multiattribute utility theory and give support for the constuction of an additive utility function for the case of risky decisions and an additive value function for decision making under certainty. In our study, we used only additive value functions. The value function obtained from both systems would therefore have the linear form,

$$v(a) = \sum_{i=1}^{Q} k_i v_i(a_i)$$

where a is an alternative, estimated over each of the Q criteria, k_i is the coefficient of importance for the ith criterion, a_i is the value of alternative a on criterion i, and v_i is the value function for the ith criterion.

Besides different interfaces, the primary difference in the systems is the way in which numerical values $v_i(a_i)$ on each of the criteria Q are determined, as well as how relative criteria weights k_i are determined. In LOGICAL DECISION, the relative value of various attainment levels on each criterion are given to the sys-

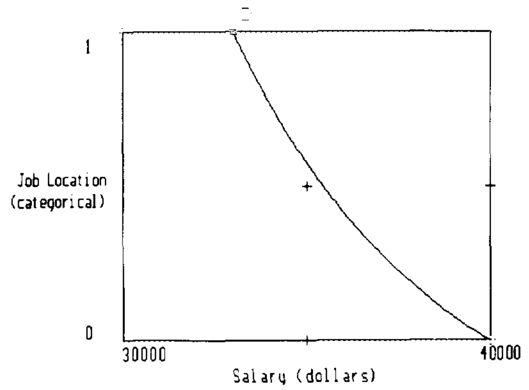
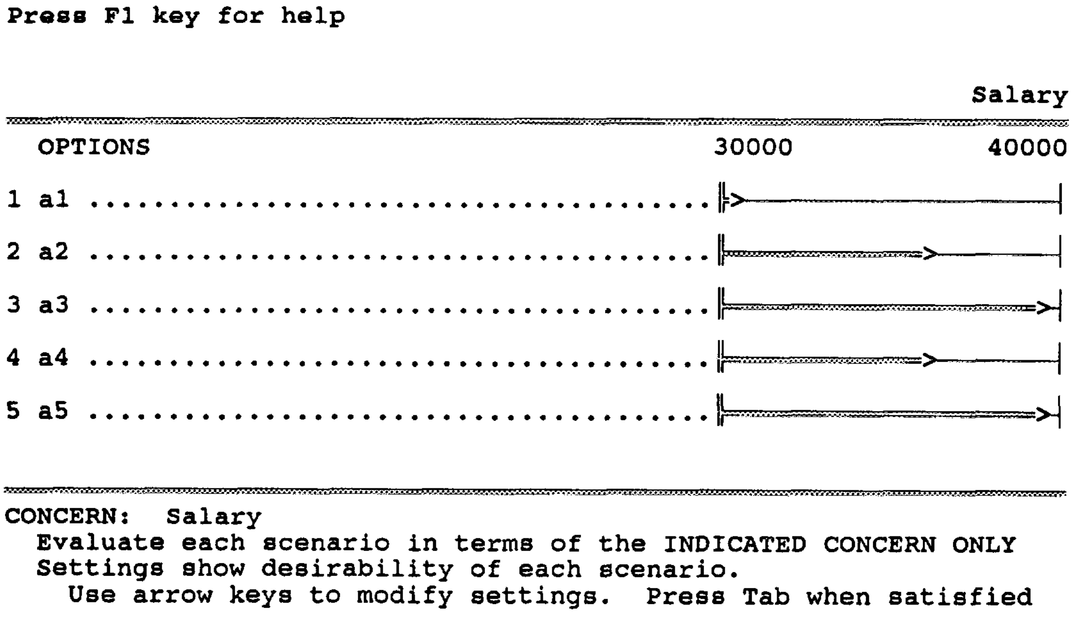


FIG. 1. Tradeoff for two criteria in LOGICAL DECISION (LD). Alternative B: salary = $33,000 and location = 1. Salary weight:job location weight = 1.9999:1.

tem by a number of options, including an option to graphically set the attainment level to be assigned a value of .5, given anchored points for values of 0 and 1 on each criterion. Then a curve is fit to the functional form $y = a + be^{-cx}$. The default is linear, as in SMART (von Winterfeldt & Edwards 1986). LOGICAL DECISION allows the subject to express relative importance of criteria through lottery selections, allowing the subject to balance attainment levels for pairs of criteria (see Fig. 1). DECAID operates entirely through the mechanism of the subject setting relative preferences and importances graphically on unit scales. DECAID operates in a much more direct manner, with the decision maker using the cursor to directly enter both the relative attainments of each alternative on each criterion on scales ranging from zero to one, as well as directly entering the relative importance of each criterion on similar scales ranging from zero to one. DECAID can thus be viewed as a direct elicitation method.

Von Winterfeldt and Edwards (1986) presented SMART as appropriate for use in obtaining direct rating of alternatives on single attributes (direct statement of $v_i(a_i)$, how well each alternative does on each criterion), as well as for ratio estimation of weights k_i, using the resulting additive value function as a means to rank alternatives. DECAID uses direct statement of both $v_i(a_i)$, and k_i by the decision maker obtained through graphical means (using the most important attribute score as an anchor—see Figs. 2 and 3). Von Winterfeldt and Edwards also stated that there are more sophisticated versions of SMART that allowed marking the best and worst values for $v_i(a_i)$ using a linear function anchored on these extremes and allowing curved value forms if there is a suggestion of nonlinear dependence. This is the form to elicit $v_i(a_i)$ used by LOGICAL DECISION (see Fig. 4).

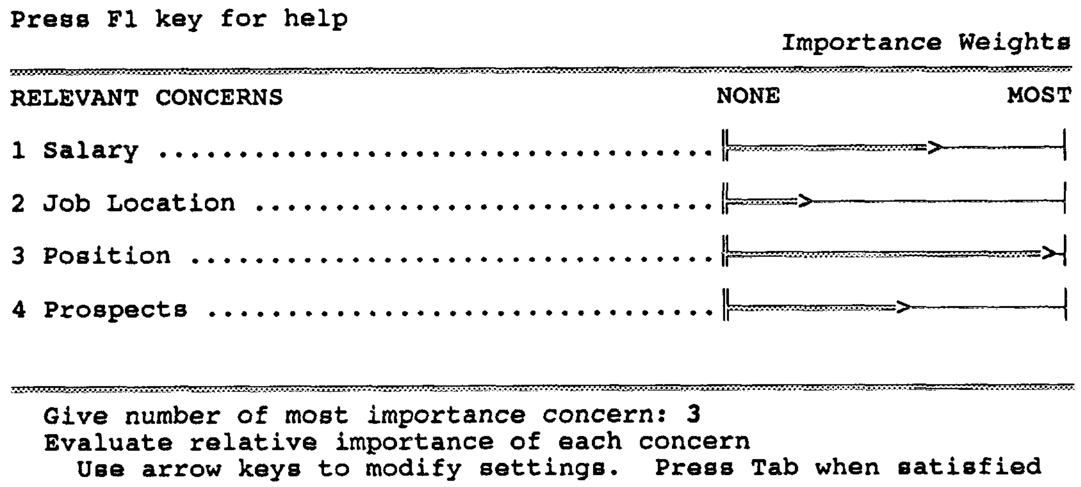
With respect to assessing k_i, Edwards (1992) has emphasized the need for the use of swing weights. Swing weights involve presenting two alternatives with the best and worst values on two attributes while holding all other attribute levels at some common level, asking the decision maker which is the preferable choice and the degree of preference. This would be followed by reassessing the ratio of relative importance through repeated judgments using different attribute bases. In DECAID and LD the best and worst values on each attribute are presented to the decision maker. However, in DECAID direct rating is used. In LOGICAL DECISION, lotteries are presented giving the best per-


```
Press F1 key for help

                                                      Salary
 OPTIONS                                      30000     40000
1 a1 ....................................
2 a2 ....................................
3 a3 ....................................
4 a4 ....................................
5 a5 ....................................

CONCERN: Salary
  Evaluate each scenario in terms of the INDICATED CONCERN ONLY
  Settings show desirability of each scenario.
     Use arrow keys to modify settings.  Press Tab when satisfied
```

FIG. 2. Graphical assessment of attribute values (for attribute 1) in DECAID.



```
Press F1 key for help
                                              Importance Weights

RELEVANT CONCERNS                             NONE            MOST

1 Salary ...................................

2 Job Location .............................

3 Position .................................

4 Prospects ................................


   Give number of most importance concern: 3
   Evaluate relative importance of each concern
     Use arrow keys to modify settings.  Press Tab when satisfied
```

FIG. 3. Screen for marking attribute weights in DECAID.

formance on one attribute combined with the worst performance on the other attribute, while all other attribute values are held constant. The decision maker is asked to select from this pair (as well as a choice where both are considered equally preferable). For the alternative that was not selected, the decision maker is asked how much improvement on the worst attribute value would be necessary to make the tradeoff equal in value. While neither the DECAID nor the LOGICAL DECISION procedure require the decision maker to use swing weights, the decision maker has the opportunity in both packages to view the tradeoff and to adjust it. In our study none of the subjects used this option.

Considering the common features of both systems (DECAID and LD), as well as the similarity of information given by the decision maker during task solution, solution of this task using either system might be expected to yield the same result. Any discrepancies noted would require more detailed analysis to define the steps in preference elicitation. Such discrepancies are the focus of this study.

PRIOR EXPECTATION

Using the MAUT approach, work usually begins with identifying relevant attributes from the decision maker using ordinal scales, which are later assigned quantitative estimates (e.g., Keeney, 1992; von Winterfeld & Edwards, 1986). Analogously, the decision maker is often asked to rank-order attributes on their importance and then calculate their relative criterion weights. This sequence in task formulation and results of a number of experiments (Nikiforov, Rebrik, & Sheptalova, 1984; Larichev, 1992) allow us to assume that qualitative (ordinal) judgments are more stable (consistent) than quantitative estimates (numerical data of tradeoffs and attribute values, including selection of the "middle" point on an anchored scale as used in LOGICAL DECISION.

In his paper "Toward the demise of economical man and woman" (in Edwards, 1992), Edwards wrote of the key principle of ordinal dominance. Ordinal dominance was stated to be the most intuitively compelling and the most rarely violated of the axioms of rationality. In accordance with this definition, we use ZAPROS as the system identifying relationships closest to those of ordinal dominance. In ZAPROS swing weights are used in the sense that the user compares alternatives having the best attainment levels on all attributes but one and the worst value on this other attribute. Unlike LOGICAL DECISION, in ZAPROS full alternatives are presented. In ZAPROS, indifference points are not elicited. User selections among pairs of fully described alternatives are used as the basis for ranking attributes. All three systems, DECAID, LOGICAL DECISION (as used in this study), and ZAPROS, assume an additive overall value function. ZAPROS uses ranking

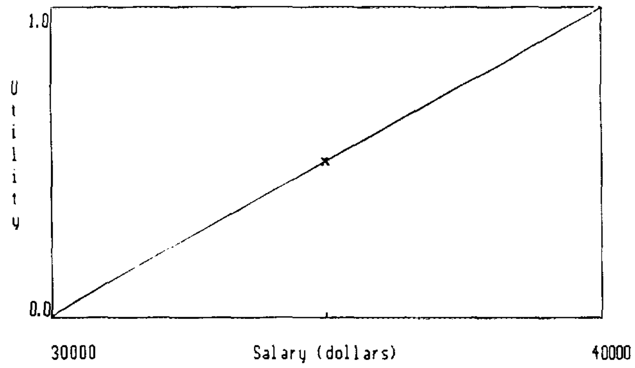


FIG. 4. Assessment of attribute levels (for one attribute) in LD. Options listed: return to SUF menu, split range, set split utility, set midpoint, delete split, initialize range, assign utility, choose range.

rather than rating information, but the additive overall value rule is correct if there is an additive value function. In ZAPROS the additive rule does not provide the summation of values, but rather the means of obtaining pairwise compensation between components of two alternatives.

To construct valid and consistent rank-orders of alternatives from list **L** (Appendix 2), a special procedure for formation of a joint ordinal scale used in method ZAPROS-LM (Larichev & Moshkovich, 1991; Larichev, 1994) was implemented. This procedure involves subject pairwise comparison of all pairs of alternatives from list **L.** The procedure also provides verification of the transitivity of selections given by the subject and allows the subject to change some responses to eliminate any intransitivity. This system guarantees that comparison of each pair of alternatives from list **L** is supported by at least two responses from the user.

People often make mistakes, and so there is the possibility of error even when using ordinal judgments. That is why we sought a stable preference system for the experimental task. We thus used the following procedure. Subjects were asked to compare several specially formed alternatives, with each member of the pair having the best attainment levels on all but one criterion and each of the pair lowering the attainment level on two different criteria (a list of such alternatives is given in Appendix 2). This procedure was used because people have been found to be more accurate when comparing alternatives varying on no more than two criteria (Russo & Rosen, 1975). Subjects were asked to choose from the following responses:

1. alternative 1 is more preferable than alternative 2;
2. alternative 2 is more preferable than alternative 1;
3. alternatives 1 and 2 are equally preferable.

Implementation of this simple system for comparison of simple pairs of alternatives gives us a simple check of the transitivity of comparisons,

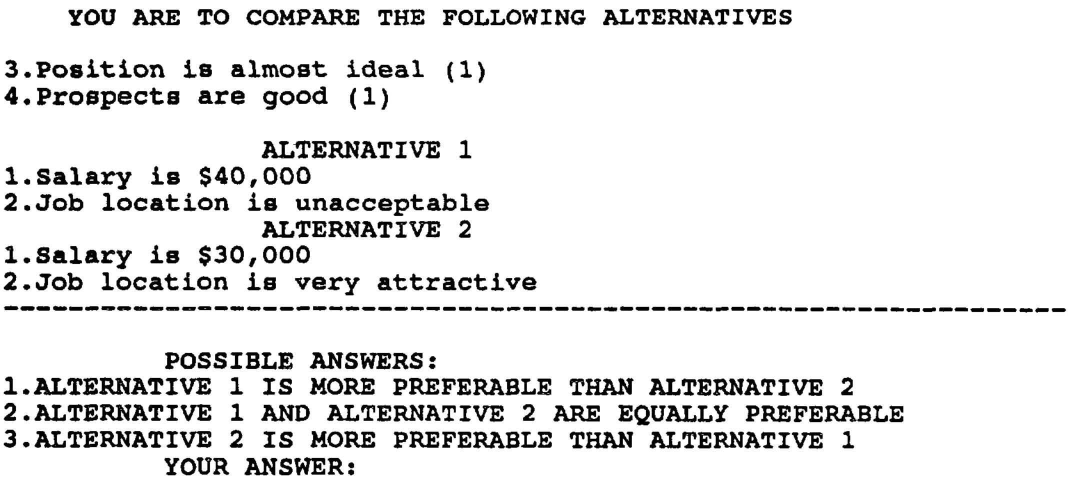
(i) if $a > b$ and $b > c$, then $a > c$;
(ii) if $a > b$ and $b = c$, then $a > c$;
(iii) if $a = b$ and $b = c$, then $a = c$;

where a, b, and c are alternatives, and the symbol > means more preferable, and the symbol = means equal preference.

ZAPROS-LM gives us a valid rank ordering of some alternatives. We can then try to rank-order the same alternatives with the help of value functions constructed using LOGICAL DECISION and DECAID. We note that it has been proven that if an additive value function exists, then the rank ordering of alternatives from list **L** (a joint ordinal scale) may be used for comparison with other alternatives described by vectors that are combinations of the same criterion values (Larichev & Moshkovich, 1991, 1994; Larichev *et al.,* 1993). The idea of pairwise compensation is illustrated by an example comparing two alternatives using ZAPROS in Fig. 5. The algorithm for comparison is presented and accuracy proven in Larichev and Moshkovich (1991). Therefore, this rank ordering may be used for comparison of the initial five alternatives, because in our task the additive value function is supposed to be correct and criteria were formed to be preferentially independent. This algorithm does not guarantee comparison of all alternatives, but those comparisons that are made can be considered as the basis for comparison of results obtained through LOGICAL DECISION and DECAID.

EXPECTATION

As a rule, pairwise comparisons of alternatives compared on ZAPROS-LM will be the same as those ob-



```
YOU ARE TO COMPARE THE FOLLOWING ALTERNATIVES

3.Position is almost ideal (1)
4.Prospects are good (1)

                ALTERNATIVE 1
1.Salary is $40,000
2.Job location is unacceptable
                ALTERNATIVE 2
1.Salary is $30,000
2.Job location is very attractive
------------------------------------------------------------------

          POSSIBLE ANSWERS:
1.ALTERNATIVE 1 IS MORE PREFERABLE THAN ALTERNATIVE 2
2.ALTERNATIVE 1 AND ALTERNATIVE 2 ARE EQUALLY PREFERABLE
3.ALTERNATIVE 2 IS MORE PREFERABLE THAN ALTERNATIVE 1
          YOUR ANSWER:
```

FIG. 5. An explanation of comparisons in ZAPROS.

tained with LOGICAL DECISION and DECAID, but pairs of alternatives that were not compared using ZAPROS-LM will more likely differ in ranking using LOGICAL DECISION and DECAID.

METHODOLOGY

The experiment was conducted as an assignment for student subjects at Texas A&M University in a course presenting decision support and expert systems. Each student had to work with systems DECAID (DC), LOGICAL DECISION (LD), and ZAPROS (Z). Subjects had to use these systems to solve the task of ranking the five alternatives given in Appendix 3. After this was completed, the subjects completed a questionnaire given in Appendix 4, which was used to characterize subject attitudes toward the systems and the results obtained. The following data were obtained for each subject:

(1) answers for the questionnaire for systems DC, LD, and Z;

(2) the order in which systems DC and LD were used;

(3) aggregated numerical values for alternatives from Appendix 3 obtained through DC and LD;

(4) numerical weights for all attributes, obtained through DC and LD;

(5) numbers for attribute values for the alternatives obtained from DC and LD;

(6) rank ordering of alternatives from list **L** (Appendix 2) obtained from Z;

(7) pairwise comparisons of alternatives from Appendix 3 obtained from Z.

To analyze differences in results from DC and LD we could use the numerical values obtained (normalized for DC). To compare these results with results obtained from ZAPROS, it is necessary to elaborate ordinal dependencies obtained through DC and LD, as ZAPROS works with ordinal judgments. The following data transformations were carried out for this purpose.

Alternatives were rank-ordered from the results of each method. This is presented as a matrix of pairwise comparisons of alternatives $M(a) = 5 \times 5$, where

$$\begin{aligned} m_{ij}(a) &= 1 && \text{if} \quad v(a_j) > v(a_j); \\ m_{ij}(a) &= 0.5 && \text{if} \quad v(a_j) = v(a_j); \\ m_{ij}(a) &= 0 && \text{if} \quad v(a_j) < \mathrm{v}(a_j). \end{aligned}$$

Such matrices were built for each subject for each system (for ZAPROS, the system built the matrices, with $m_{ij} = 3$ if a_i and a_i were incomparable on ZAPROS).

The numerical values obtained through DC and LD and attribute weights were used to calculate aggregated values of alternatives from list **L** (Appendix 2) using formula (1). The results were used to build matrices of pairwise comparisons of alternatives from Appendix 2 $M(\mathbf{L})$ in the same manner as $M(a)$ was built. Analogous matrices from ZAPROS were obtained from that system. Analogous matrices for the criteria M(c) were built based on the weights obtained through DC and LD. For ZAPROS, the following method was used. Let us consider alternatives I5 through I8 from list **L** (Appendix 2). Comparison of any two of these alternatives will show the order of weights for corresponding attributes. Let us assume that I5 was prefered to I6. This means $v(\mathrm{I5}) < v(\mathrm{I6})$. If we use formula (1) for calulation of these values, the following expressions will result:

$$w_1v_1(3) + w_2v_2(1) + w_3v_3(1) + w_4v_4(1) > w_1v_1(1) + w_2v_2(3) + w_3v_3(1) + w_4v_4(1).$$

If we recall that $v_1(3) = v_2(3) = 0$ (following MAUT), the following results:

$$w_2v_2(1) > w_1v_1(1).$$

Taking into account that $v_2(1) = v_1(1) = 1$ (following MAUT), we obtain

$$w_2 > w_1.$$

Thus, based on data from pairwise comparison of alternatives from **L**, we are able to construct a matrix of pairwise comparison of attribute weights based on ZAPROS results.

These matrices are used to evaluate the **number of reversals** in pairwise comparison of alternatives and attribute importance based on the different methods. This number of reversals is equal to the number of cases where element m_{ij} for one system is equal to 1, and m_{ij} for another system is equal to 0 (we do not consider as *evident reversals* cases where on one method values are equal and on another method these values are different). Therefore, the number of reversals will characterize the primary differences in evaluation of alternatives through different selection techniques.

ANALYSIS OF RESULTS

Thirty subjects took part in the experiment. Only 22 of these worked with all of the systems and fulfilled all tasks. Data on the averages of their answers to the questionnaire are given in Table 1. Analysis of the data shows that subjects prefer system DECAID (although objective results do not support this result, as will be seen in further analysis). The DECAID system asks questions that seem simple and understandable. Only direct estimation of alternative values and attribute weights are required, expressed in simple graphical

TABLE 1

Average Responses to the Questionnaire

Systems/Questions	1	2	3	4	5	6
DECAID	1.3	1.8	1.2	1.7	1.8	1.8
Logical decision	2.0	2.1	1.7	2.4	2.1	1.9
ZAPROS	2.0	2.3	2.5	2.3	2.5	2.1

mode. The result is quick, and the process is throughly understandable. The LOGICAL DECISION system, although sounder from a theoretical point of view, is less attractive to subjects, as more questions are asked, and the questions the system asks are not always as transparent and easy to understand as those asked by DECAID. The least attractive system was ZAPROS. ZAPROS asks a lot of questions of subjects on the comparison of hypothetical alternatives, and any resulting inconsistencies are analyzed by further questions. Also the relationship of the questions asked by ZAPROS to the ultimate ranking of alternatives is much less clear than with the other systems. These results show that naive users prefer simple systems, in which they feel that they understand all of the "inside mechanisms." This preference is not that distinctive when the merit of the result is addressed.

Further results of using the different systems were analyzed. First, it was found that six of the subjects, while using DECAID, marked reversals when assigning alternative values. As was demonstrated above, the alternatives were specially constructed so that three possible values were used for each attribute, and these values were rank-ordered from the best to the worst. The formulations of these possible values were such that there was no logically consistent way to consider the second value as more preferable than the first value (see the list of attributes in Appendix 1). These facts show that the subjective impression of the users is not always supported by real results. When a task is easy, serious mistakes may result if there are no special means for detecting errors. The LOGICAL DECISION and ZAPROS systems do not allow the user to make such mistakes, and therefore this problem did not occur when subjects used those systems.

These six subjects demonstrated complete misunderstanding of the experiment. This could have been due to lack of appropriate attention to the assignment or some drastic changes in preferences while working with the system, or some other reason. In any case, it is evident that the results from those subjects who encountered this type of problem do not deserve further analysis for our purposes.

For the remaining 16 subjects, all previously described parameters were estimated. Of these 16 subjects, 9 first worked with DC followed by LD, while 7 used LD first and then DC. All used ZAPROS last. Analysis of results showed that the order of work did not influence the result. Therefore, we will not concern ourselves further with the order of system use. The initial analysis was to examine the correspondence of results obtained from DC and LD. For 6 of 16 subjects, the best alternative was the same in both cases. Ten subjects had different first selections with the two systems. This indicates a very low correspondence of results. Therefore, we analyzed the source of differences, as both systems used the same theoretical model as the basis for alternative evaluation.

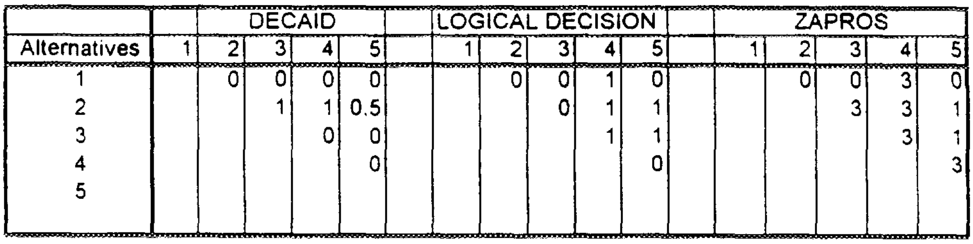
ANOVA was used to analyze the correspondence of data on alternatives, attribute weights, and attribute estimates obtained from both systems. Only numerical estimates for the second value on each attribute scale was available to analyze relative to attribute estimates, as the best value on the scale is set equal to one, and the worst value on the scale is set equal to zero, following MAUT. Therefore, differences could occur only in the estimation of the middle value on the unit scales. The analysis was conducted for each alternative and each attribute for the whole group of subjects. Results are presented in Table 2.

As can be seen, the ANOVA test for alternatives 3 and 5, weights for attributes 1, 2, and 4, and values for attributes 1, 3, and 4 failed. Nevertheless, the result for the group (successful ANOVA test for alternatives

TABLE 2

***F*-test for Alternatives' Value, Attribute Weights, and Attribute Values for Systems DECAID and LOGICAL DECISION**

	Alternatives					
	1	2	3	4	5	*F*-crit.
F for alternatives' value	21.6	11.68	0.2848	10.7	0.3857	4.17 alpha = 5% 2.88 alpha = 10%
	Attributes					
	Salary	Location	Job type	Perspectives		
F for attribute weights	0.0083	0.0116	4.205	1.212		4.17 alpha = 5%
F for attribute values	1.919	9.1168	0.2408	0.24		2.88 alpha = 10%



	DECAID						LOGICAL DECISION						ZAPROS				
Alternatives	1	2	3	4	5		1	2	3	4	5		1	2	3	4	5
1		0	0	0	0			0	0	1	0			0	0	3	0
2			1	1	0.5				0	1	1				3	3	1
3				0	0					1	1					3	1
4					0						0						3
5																	

FIG. 6. Matrices of pairwise comparisons of real alternatives for subject 9. 1, alternative in the row is more preferable than alternative in the column; 0, alternative in the column is more preferable than alternative in the row; 0.5, alternatives in the row and in the column are equal in preference; 3, alternatives in the row and column have not been compared.

1, 2, and 4) support the idea that subjects tried to implement their real preferences in working with both systems. This is in spite of the fact that there was only one subject who managed to rank order the five alternatives identically using both systems.

The analysis showed that both systems resulted in different estimates of overall alternative values and that there is a significant difference in how subjects estimated attribute weights and attribute values across the two systems. Even for the entire group, there is only one attribute (attribute three—JOB TYPE) upon which the subjects were sufficiently stable for an accurate estimate of importance. Furthermore, only on attribute two (LOCATION) were subjects sufficiently stable to allow estimation of attribute values.

To analyze our hypothesis that the differences in measured parameters must be less in ordinal judgments, the analysis was conducted for ordinal dependencies resulting from the numerical values used. Then these data were compared with those obtained from the ZAPROS results. For each subject for each system (DC, LD, and Z), matrices of pairwise comparisons of the five real alternatives (Appendix 3), attribute importance (weights), and alternatives from list **L** (Appendix 2) were built. An example of the matrices for DC, LD, and Z for subject 9 over the comparison of the five real alternatives is given in Fig. 6.

It can be seen that it is easy to calculate the number of reversals in preferences between elements of two matrices. We note that for the case of comparing five real alternatives (as presented in Fig. 6), the reversals with results of ZAPROS were calculated only for pairs compared on ZAPROS.

In Table 3, averages for the number of reversals between results obtained through systems DECAID, LOGICAL DECISION, and ZAPROS for real alternatives, criteria weights, and hypothetical alternatives near the ideal alternative are given. The asterisk indicates data for DC and LD that are calculated for pairs and compared on ZAPROS. As these data correspond to a different number of pairwise comparisons, the percentage of reversals is also given. Data for all subjects for systems DC and LD are presented in Fig. 7.

These data show that for almost all parameters, results obtained through DECAID and LOGICAL DECISION differ from each other in about one-third of the cases, with the largest variance for real alternatives and the least variance for attribute weights. Nevertheless, if we analyze the number of pairwise reversals caused by DC and LD for those pairs of alternatives that were compared on ZAPROS (DC–LD*), we see that the correspondence in the result is larger here. We roughly conclude that only one of three reversals in comparison of alternatives using DC and LD is due to the pair of alternatives compared on ZAPROS.

ZAPROS allows only ordinal judgments, which are tested and corrected during subject use of the system. That is why we are able to consider comparisons made on ZAPROS to be accurate reflections of subject preferences. Analyzing the data, it is interesting to notice that comparisons of alternatives using DC and Z coincide more often than do the results of LD and Z. At the same time, the rank ordering of attributes by their importance is stronger between LD and Z. This result is more easily understood if we recall how attribute importance is measured in different systems. DC just elicits attribute weights directly (a point on a line of unit length). There is the possibility in the system to conduct tradeoff analysis, but subjects did not use this feature. In this sense they were asked only to estimate attribute importance directly.

In LD and Z the subjects were asked to make tradeoffs between pairs of attributes (in ordinal form in Z

TABLE 3

Average Number of Reversals in Pairwise Comparisons in Three Systems (DECAID, LOGICAL DECISION, ZAPROS)

	Pairs of systems			
Parameters	DC–LD	DC–Z	LD–Z	DC–LD*
Real alternatives	3.56	1.19	1.5	1.25
Percentage	35.60	19.30	29	26.50
Attribute weights	1.88	1.69	1.38	
Percentage	31.3	28.20	23	
Hypothetical alternatives	9.75	6.9	8.4	
Percentage	34.80	24	30	

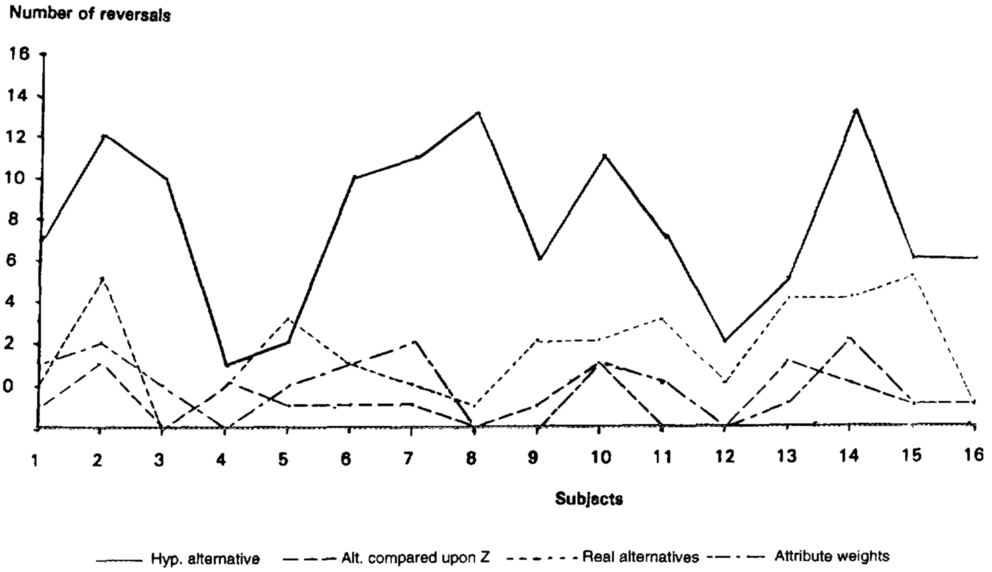


FIG. 7. Number of reversals for DC and LD.

and in numerical form when using LD). In this sense the subjects were not asked to estimate attribute importance at all. There is a great deal of evidence (Belton, 1986; Schenkerman, 1991) that people may think of different things when they are asked to estimate attribute importance. However, in the models used, attribute importance has a very definite meaning (how much of an attribute of unit A would you give up in exchange for a unit of attribute B?). That is why attribute importance elicited in an indirect tradeoff manner is closer to the real goal than is direct asking of attribute importance. Therefore, we conclude that differences between LD and Z are caused by eliciting numerical or ordinal judgments, while differences betwen DC and Z (as well as between DC and LD) are caused by eliciting different things.

In these circumstances, the closer results in comparison of alternatives obtained through DC and Z (compared to the closeness of results between LD and Z) may be due to differences in estimation of the second attribute values. Judgments elicited through ZAPROS also put some limitations on the possible estimation of these second values.

These results show that dependencies for the second values obtained through ZAPROS are closer to those obtained through DC than those obtained through LD. In DC we have direct estimation of all values for each attribute (as a point on a line). These values are later normalized by the system to esimates ranging from 0 to 1. In LD we have the standard approach of marking the "middle" point for the range of values for each attribute. The data measuring results allows us to conclude that when we ask a subject to mark the middle, it very often may be "more close" to the physical midpoint between 0 and 1 than the middle value point based on preferences. Data obtained in the experiment indirectly confirm this conclusion. In Table 4 we have marked the number of estimates equal to 0.5 for the second attribute value for each attribute for LD and for DC. We see that this is very rare for DC and occurs rather often for LD (especially for "qualitative" attributes such as JOB POSITION or PROSPECTS). The same table also gives the number of times that subjects had estimates of more than 0.5 for one system and less than 0.5 for the other. This number is quite high. This supports the assumption that numerical estimation of attribute values is rather tricky and can lead to very different results. The widely approved method of marking the middle point on a value scale may be questioned from a judgmental point of view. Our results

TABLE 4
Comparison of Estimates for Second Attribute Values on DC and LD

	Attributes			
	Salary	Location	Job type	Perspectives
Number of estimates equal 0.5				
Upon DC	0	0	1	1
Upon LD	3	6	8	8
Number of cases with < 0.5 upon one system and >0.5 upon the other (DC–LD)	9	12	7	7

show that it is not that easy to mark the middle point (even for such a "quantitative" attribute as SALARY).

The last analysis was conducted for the entire group on pairwise comparison of attribute weights and real alternatives. The ANOVA technique for measuring the similarity of results for pairwise comparison of attribute weights and real alternatives was used. The results are presented in Table 5.

These results show that for both parameters, the hypothesis is rejected for LD and DC. However, data obtained comparing LD and Z and DC and Z is supported (the *F* test is passed in both of these cases). This again confirms our hypothesis that the decision analysis is more stable when judgments are elicited in an ordinal form. The assignment of quantitative estimates to ordinal values often creates a false illusion of exactness. Attempts to elicit numerical data (or use numerical approximation for ordinal judgments) may lead to the wrong solutions. Those ordinal preferences among pairs of real alternatives that subjects wre able to express through ZAPROS were more stable and showed higher consistency when compared with the preference results of LOGICAL DECISION and DECAID than those items that were incomparable when using ZAPROS.

DISCUSSION

We have presented the results of an experiment where subjects were to compare five multiattribute alternatives using two decision aids (decision support systems) LOGICAL DECISION and DECAID. These two systems were selected because both are based on the theory of multiattribute utility theory, and both use a multiattribute additive value function. Both were easy to implement, but varied slightly in the way in which they elicited attribute weights and estimates of attribute values. The most popular elicitation procedures (direct graphical estimation vs tradeoffs with "mid-points") were used in these two systems.

In real decision making cases difficulty can vary. In simple cases there is an alternative in the initial set that is far superior to the others. There would be an obvious correct solution, and this solution would be selected with almost any multiattribute procedure. In experiments described in Larichev *et al.* (1993), four different methods were used. Seventy percent of the subjects selected the same best alternative (although there was a far lower proportion, 20%, agreeing on the second alternative). In the experiment presented in this paper we tried to develop a difficult decision context. As a result, we obtained very high instability in the selection of the best alternative (and even more instability in ranking of alternatives).

TABLE 5

***F* Test Results for Data Dimilarity in the Form of Pairwise Comparisons for the Entire Group**

	Systems			
	DC–LD	DC–Z	LD–Z	*F*-crit.
F for attribute weights	0.707	4.79	4.967	3.899 alpha = 5%
F for real alternatives	0.0298	4.735	5.375	3.899 alpha = 5%

There is an opinion that sensitivity analysis can be used to compensate to some extent for instability of quantitative analysis. Sensitivity analysis is useful if there is a good scenario for it. Both systems used in this study provided the user with the ability to change estimations of all parameters and to see the impact of these changes. However, the analysis of differences in several parameters simultaneously is complicated, fragmentary, and does not help the user when different solutions have widely varying attainment levels. Therefore effective, comprehensive, and useful sensitivity analysis is quite difficult. Useful systems should provide comfortable means for analysis of results, and we feel that the appropriate language should be more "qualitative" in nature, as qualitative discussion is more natural for people.

These doubts and our experiences lead us to the conclusion that the exactness of results must not always be the primary goal of analysis. How exact is exact enough? We feel that the answer to this question must be related to the decision maker's ability to give exact judgments. It is well known that the exactness of physical measurements depends on the exactness of the instrument used. We think that the same is true for human measurements. When it is difficult to assume that the decision maker is able to give exact valid numerical estimations of different parameters, it is better to carry out the analysis using ordinal (and often verbal) judgments with the appropriate logical analysis of possible inconsistencies. The "inexactness' of the result in some cases (e.g., when it is not possible to select the best alternative based upon information given) will indicate that the decision maker is not able to make the final choice based on the evidence given. This result may be a stimulus for reformulation or modification of the task. In general, the decision maker can reformulate the problem by aggregating or disaggregating some attributes, gaining additional information, and so on. Sometimes it is reasonable to conclude that insufficient information is available for a decision. However, the attempt to achieve the "exactness' of the result through some artificial means of substituting valid ordinal judgments by some numerical rescaling may lead to the wrong decision. Thus, such attempts may only give the impression of successful task solution instead of more appropriately analyzing the task in greater depth.

Our results show that ordinal relations between task parameters are much more stable than those based on quantitative scaling. Results from LOGICAL DECISION and DECAID are much less coincident with each other than with data obtained through ZAPROS. At the same time, errors were identified in ordinal judgments, even in response to simple questions. That is why decision support techniques must pay more attention to the testing of the consistency of the judgments entered, because there usually is no means of checking the validity of the results of the analysis due to the subjective nature of multiattribute decision making.

CONCLUSION

Our aim was not to find the best method, but to analyze the difference in applying two decision support systems based on the same theoretical framework (MAUT) for cases where hard choices were present. ZAPROS was used as a basis for evaluating the performance of the other two methods, because ZAPROS relies on ranking (which is more reliable than rating according to von Winterfeldt and Edwards (1986)). Furthermore, ZAPROS uses swing preference information for evaluating ranking of attribute importance and uses only ordinal comparisons of holistic multiattribute alternatives differing on two criteria, thus simplifying the comparison process. ZAPROS checks decision maker preference statements for transitivity and, when intransitivity is identified, explains the case to the decision maker. Thus ZAPROS provides consistent and stable information on pairwise comparisons, providing a stable basis for evaluation.

Our study showed significant differences in the resulting choice, although differences in numerical values on weights and values were not significant. We were not able to conclude that one method was better than the other on the basis of objective information. We were able to conclude on the basis of subjective preferences about the appropriateness and dependencies of results on different forms of information elicitation.

Attempts to solve decision tasks through more "exact" (quantitative) judgments about value function parameters in real tasks may lead to erroneous results, sometimes merely due to small biases in quantitative data. In these circumstances, it may be more reasonable to simply use ordinal judgments, which usually are more easily verified. If this qualitative information does not resolve the decision choice between available alternatives, the problem can appropriately be dealt with by reformulation of the criteria space (Berkeley, Humphreys, Larichev, & Moskovch, 1990) or gathering of additional information.

APPENDIX 1: CRITERIA FOR JOB EVALUATION

Criterion 1: SALARY
 Salary is $30,000
 Salary is $35,000
 Salary is $40,000
Criterion 2: JOB LOCATION
 Job location is very attractive
 Job location is adequate
 Job location is not attractive
Criterion 3: JOB POSITION
 Job position is almost ideal
 Job position is good enough
 Job position is not appropriate
Criterion 4: PROSPECTS
 Prospects for training and promotion are good
 Prospects for training and promotion are moderate
 There are almost no prospects for training and promotion.

APPENDIX 2: LIST L OF ALTERNATIVES

Alternative	Salary	Location	Position	Prospects	Vector
I1	$35,000	Very attractive	Almost ideal	Good	2111
I2	$40,000	Adequate	Almost ideal	Good	1211
I3	$40,000	Very attractive	Good enough	Good	1121
I4	$40,000	Very attractive	Almost ideal	Moderate	1112
I5	$30,000	Very attractive	Almost ideal	Good	3111
I6	$40,000	Unattractive	Almost ideal	Good	1311
I7	$40,000	Very attractive	Not appropriate	Good	1131
I8	$40,000	Very attractive	Almost ideal	Almost none	1113

APPENDIX 3: LIST OF INITIAL ALTERNATIVES

Firm	Salary	Location	Position	Prospects
a1	$30,000	Very attractive	Good enough	Moderate
a2	#35,000	Unattractive	Almost ideal	Moderate
a3	$40,000	Adequate	Good enough	Almost none
a4	$35,000	Adequate	Not appropriate	Good
a5	$40,000	Unattractive	Good enough	Moderate

APPENDIX 4: QUESTIONNAIRE

Name________________________________

Question 1. How easy was it to use the system?

	DECAID	LOGICAL DECISION	ZAPROS
1. Very easy to use	________	________	________
2. Some minor difficulties, but easy	________	________	________
3. Some difficulties, but usable	________	________	________
4. Inconvenient, workable with great effort	________	________	________
5. Unusable	________	________	________

Question 2. How satisfied were you with the recommendations given by the system?

	DECAID	LOGICAL DECISION	ZAPROS
1. I fully agree with the system's result	________	________	________
2. I think the system's result is mostly accurate	________	________	________
3. I doubt the system's result is accurate	________	________	________
4. I think the system is inaccurate	________	________	________

Question 3. How understandable was the system output?

	DECAID	LOGICAL DECISION	ZAPROS
1. System output was very easy to understand	________	________	________
2. System output a little hard to understand	________	________	________
3. I don't have any idea how the system made its conclusion	________	________	________

Question 4. How quick was the system?

	DECAID	LOGICAL DECISION	ZAPROS
1. Very quick	________	________	________
2. Quick enough	________	________	________
3. Not quick, but reasonable	________	________	________
4. Took a long time	________	________	________
5. Unreasonable	________	________	________

Question 5. Was the system useful to you?

	DECAID	LOGICAL DECISION	ZAPROS
1. By using the system, I understood a lot more	________	________	________
2. By using the system, I understood a little more	________	________	________
3. The system did not improve my understanding of the decision	________	________	________

Question 6. Would you use the system for a real choice?

	DECAID	LOGICAL DECISION	ZAPROS
1. I think this system would help in real decisions	________	________	________
2. The system might help in real decisions	________	________	________
3. The system would not be useful in real decision making	________	________	________

REFERENCES

Belton, V. (1986). A comparison of the analytic hierarchy process and a simple multiattribute value function. *European Journal of Operational Research* **26,** 7–21.

Berkeley, D., Humphreys, P., Lairchev, O. I., & Moshkovich, H. M. (1990). Modelling and supporting the process of choice between alternatives: The focus of ASTRIDA. In H. G. Sol and J. Vecsenyi, (Eds), *Environments for supporting decision processes.* Amsterdam: North-Holland.

Buede, D. M., & Choisser, R. W. (1992). Providing an analytic structure for key system design choices. *Journal of Multi-Criteria Decision Analysis* **1**(1), 17–27.

Edwards, W. (1983). Human cognitive capabilities, representativeness and ground rules for research in decision making. In Humphreys, P. C., Svenson, O., and Vari, A. (Eds.), *Analysing and Aiding Decision Processes.* Amsterdam: North-Holland.

Edwards, W. (1992). *Utility Theories: Measurements and Applications.* Kluwer Academic, Dordrecht.

Forman, E. (1992). *Expert Choice.* Pittsburgh, PA: Decision Support Software.

Humphreys, P., & McFadden, W. (1980).Experiences with MAUD: Aiding decision structuring versus bootstrapping the decision maker. *Acta Psychologica* **45,** 51–89.

Keeney, R. L. (1992). *Value-Focused Thinking.* Cambridge, MA: Harvard Univ. Press.

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

Larichev, O. I. (1992). Cognitive validity in design of decision-aiding techniques. *Journal of Multi-Criteria Decision Analysis* **1**(3), 127–138.

Larichev, O. I., & Moshkovich, H. M. (1991). *ZAPROS: A method and system for ordering multiattribute alternatives on the base of a decision-maker's preferences,* preprint. All-Union Research Institute for Systems Studies, Moscow.

Larichev, O. I., & Moshkovich, H. fM. (1994). ZAPROS-LM—A method and system for rank-ordering of multiattribute alternatives. *European Journal of Operational Research,* in press.

Larichev, O. I., Moshkovich, H. M., Mechitov, A. I., & Olson, D. L. (1993). Experiments comparing qualitative approaches to rank ordering of multiattribute alternatives. *Journal of Multi-Criteria Decision Analysis* **2**:1, 5–26.

Montgomery, H. (1977). A study of intransitive preferences using a think aloud procedure. In Jungerman, H., and de Zeeuw, G. (Eds.), *Decision Making and Change in Human Affairs.* Riedel: Dordrecht.

Montgomery, H., Garling, T., Linberg, E., & Selart, M. (1990). Preference judgment and choice: Is the prominent effect due to information integration or information evaluation? In Borcherding, K., Larichev, O. I., and Messick, D. M. (Eds.), *Contemporary Issues in Decision Making.* Amsterdam: North Holland.

Received: April 14, 1994

Nikiforov, A., Rebrik, S., & Sheptalova, L. P. (1984). Experimental investigation of stability of preferences in some tasks of decision making. *VNIISI Transactions* **9,** VNIISI, Moscow. [In Russian]

Olson, D. (1992). Review of empirical studies in multiobjective mathematical programming: Subject learning and response to nonlinear utility. *Decision Sciences* **23**(1), 1–20.

Payne, J. W. (1976). Task complexity and contingent processing in decision making: An information search and protocol analysis. *Organizational Behavior and Human Processing* **16,** 366–387.

Pitz, G. F. (1987). *DECAID Computer Program.* Carbondale, IL: Univ. of Southern Illinois.

Rohrmann, B. (1986). Evaluating the usefulness of decision aids: A methodological perspective. In Brehmer, B., Jungermann, H., Lourens, P. and Sevon (Eds.), *New Directions in Research on Decision Making.* Amsterdam: North-Holland.

Russo, J. E., & Rosen, L. D. (1975). An eye fixation analysis of multiattribute choice. *Memory and Cognition* **3,** 267–276.

Schenkerman, S. (1991). Use and abuse of weights in multiple objective decision support models. *Decision Sciences* **22,** 369–378.

Schoemaker P. J. H., & Waid C. C. (1982). An experimental comparison of different approaches to determining weights in additive utility models. *Management Science* **28:**2, pp. 182–196.

Smith, G. R., & Speiser, F. (1991). *Logical Decision: Multi-Measure Decision Analysis Software.* Golden, CO: PDQ Printing.

Timmermans, D. (1991). *Decision aids for bounded rationalists (an evaluation study of multiattribute decision support in individual and group settings).* Dissertation, Groningen University.

Winterfeldt, D. von, & Edwards, W. (1986). *Decision Analysis and Behavioral Research.* UK: Cambridge University Press.