ELSEVIER

Decision Support Systems 22 (1998) 73–84

Decision Support Systems

Data influences the result more than preferences: Some lessons from implementation of multiattribute techniques in a real decision task

Helen M. Moshkovich [a], Robert E. Schellenberger [b], David L. Olson [c,*]

[a] University of West Alabama, Livingston, AL 35470, USA
[b] Department of Decision Sciences, East Carolina University, Greenville, NC, USA
[c] Department of Business Analysis, Texas A & M University, College Station, TX 77843-4217, USA

Abstract

A multiattribute selection task requiring identification of a subset of the best fifteen or so applicants for a faculty position is analyzed with two techniques—SMART and ZAPROS. SMART provides a cardinal measure of each alternative that is easily used to identify the top candidates. ZAPROS relies on ordinal, verbal input from the decision makers, but provides only partial order of alternatives, meaning that the specific number of applicants identified is not guaranteed to be the number desired. Four decision makers took part in the selection task. Essential differences between results across these SMART and ZAPROS were found for all four subjects engaged in this task. Further analysis showed that alternative scores on attributes were found to influence the results more than attribute weights. Verbal scales and judgment used in ZAPROS were considered by the participants to be of more meaning and better enabled understanding of the similarities and differences in preferences and positions. ZAPROS was considered useful in the first stage of the task, as the basis of elaborating group policy in establishing relative importance of attributes and for establishing relative performance of alternatives on attributes. © 1998 Elsevier Science B.V.

Keywords: MCDM selection methods; SMART; ZAPROS; Preference

1. Introduction

Most business decisions involve a desire to fulfil multiple objectives. To deal with such situations the concepts of multiattribute decision making have been articulated. However, a difficulty with multiattribute methods is the absence of any objective model capable of dealing with many criteria. Thus, multiattribute methods need to use information (or judgments) from decision makers to reveal his (or her) preferences for characteristics of different alternatives.

One of the most popular approaches in this field is that of multiattribute utility theory (MAUT) which is often substituted by multiattribute value theory for practical tasks under certainty [14,19,27]. The traditional approach consists of the following steps: a list of alternatives is formed; a list of attributes, characterizing aspects important to the decision maker in discriminating among these alternatives is developed; the decision maker determines weights for attributes and scores for each alternative upon each

* Corresponding author. Tel.: + 1-409-845-2254.

PII S0167-9236(97)00024-9

attribute reflecting his or her preferences; some scalar function (usually additive) is used to combine these scores and weights into an overall value estimate for each alternative. After that all alternatives are rank-ordered according to these estimates and the decision task may be considered accomplished.

Multiattribute decision making is a rather straight-forward process if the decision maker can unambiguously establish numeric values for the importance of each attribute and can unambiguously establish the value of each attribute for each alternative. As the work of Slovic et al. [23], Kahnemann et al. [13], Keeney and Raiffa [14] and others recognize, the process of eliciting necessary information for such decisions is one of the major challenges facing the field. That is why many studies were devoted to questions of the stability, consistency and correctness of judgments used in decision analysis.

There are a number of methods and procedures for eliciting attribute weights and scores. Some of these methods are based on sound theory, while others use simplified heuristic approaches. Experiments [4,22,29] show that different techniques for deriving weights may lead to different results. There is evidence that in many modelling situations, varying specific weights assessed for separate attributes often does not change the selection of the most preferred alternative, thus leading to the conclusion that the model with equal weights works sufficiently for many purposes [6,7,9,28]. However, the situation may not be the same for real decision tasks when differences between alternatives are small. Stillwell et al. [24] reported that these methods are inconsistent when dominated alternatives (i.e., those with lower values for all attributes) are eliminated. In such situations slight differences in weights can lead to reversals in the ranking of alternatives [17,18].

Although discussion about the appropriateness and effectiveness of different multiattribute decision methods is very active in current publications [1,17,19,20,25–27], not much evidence exists concerning application of different decision techniques for solution of the same task in a real task environment. Usually studies apply one method and analyze the others as possible alternatives [11,21].

This paper reports on the applications of two multiattribute decision making techniques to the selection of new faculty members at an AACSB Southern School of Business. Selection of faculty in the current environment has several peculiarities:

1. the number of alternatives for selection is rather large (possibly over 100 applicants for a position);
2. attributes, characterizing each applicant, are elaborated by the selection committee and are essentially qualitative;
3. the aim of the first step is not to select the best applicant, but rather to select a small group (up to 15 applicants) for further investigation;
4. each applicant is evaluated on attributes by members of the selection committee.

The usual procedure (a form of SMART, proposed by Edwards [8] is used for evaluation of applicants. The set of attributes is formed. Each member of the selection committee evaluates each applicant over these attributes using a 10-point scale (10 is assigned to the best level of attainment). These evaluations are averaged for each applicant. The committee agrees on the relative importance of each attribute and an appropriate percentage is applied to each. These weights and values are used to generate scores for each applicant, which in turn are used for selection of the appropriate group of applicants.

Several questions arise about this procedure:

1. are average scores the best approach to use in selection of applicants?
2. do weights reflect real or imagined preferences of committee members?
3. are there essential differences in estimation of applicants over attributes among committee members, and if so, what are the reasons?
4. what influences the selection more: applied weights, or applied attribute values?
5. is a 10-point scale very comfortable and meaningful to committee members?

All of these questions led to the decision to apply an alternative technique to the same task in order to gain some insight about the procedure and its meaning. ZAPROS [16] was used because: (1) it is based on the existence of an additive value function, as is SMART;(2) attribute importance is not directly elicited, but information is obtained about rank ordering of attribute importance; (3) ZAPROS works with verbal scales and preference is elicited through pairwise comparisons of specially constructed alternatives (differing in value on only two attributes). Preferences are expressed in verbal ordinal form (more preferable, equally preferable, less preferable); (4) ZAPROS can work with large numbers of alternatives; (5) ZAPROS is able to select a subset of better alternatives of approximately the desired size.

While based on the same theoretical model as SMART, ZAPROS uses a very different elicitation procedure, and provides some insights about the preference structure of the decision maker.

The analysis shows that there are essential differences between results across these two methods for all four subjects engaged in this task. It is interesting that attribute importance (presented in the form of weights or rank ordering of attributes on importance) influences the result (rank ordering of alternatives) less than the value scores assigned to applicants by members of the selection committee. Verbal scales and judgment were considered by the participants to be of more meaning and better enabled understanding of the similarities and differences in preferences and positions.

1.1. Task

There were two positions open in the department. One position was in Systems Analysis and Database Systems. The second position was in Production/Operations Management. There were 48 valid applications for the first position, and 45 valid applications for the second. Two faculty members from each of the two disciplines were appointed to the selection committee. The first group (raters 1 and 2) elaborated six attributes, and the second group (raters 3 and 4) used five attributes (attributes are presented in Table 1). For the ZAPROS application, three possible levels for each attribute were used: (1) above average; (2) average; (3) below average.

First, each group conducted the SMART analysis.

1. the rank order of attributes by importance was agreed upon:
   for the first group: C4 = C5 > C1 = C2 = C6 > C3
   for the second group: C1 = C2 = C5 > C3 = C4
2. weights were assigned in accordance with this rank order:
   for the first group: C4 = C5 = 23%; C1 = C2 = C6 = 15%; C3 = 9%
   for the second group: C1 = C2 = C5 = 24%; C3 = C4 = 14%
3. each applicant was evaluated on a 10-point scale.

After the SMART analysis, ZAPROS was used. Before presenting the results of the ZAPROS analysis, we will describe the method in greater detail.

2. Main ideas of ZAPROS

The main idea of ZAPROS is to ask the decision maker to make trade-offs for each pair of attributes and for each pair of possible attributes in ordinal form. To conduct this task, we need to ask a decision maker questions of the type:

What do you prefer: to have above average level on attribute 'Ability to teach our students' and average level on attribute 'The ability to teach Systems Analysis and Database', or above average level on 'The ability to teach Systems Analysis and Database' and

Table 1
List of attributes with possible levels

Attributes for first position	Attributes for second position
1. Ability to teach our students	Ability to teach our students
2. Ability to teach systems analysis and database systems	Maturity and experience in POM
3. Evaluation of completed research and scholarship	Ability for leadership of POM program
4. Potential for publication	Evaluation of completed research and scholarship
5. Potential for leadership in research	Potential for journal publication
6. Match of research interests	
Levels for ZAPROS	
1. Above average	
2. Average	
3. Below average	

Attributes	Alternative A	Alternative B
Ability to teach our students	**Below Average**	**Above Average**
Ability to teach systems analysis & DBS	Above Average	**Below** Average
Evaluation of completed research and scholarship	Above Average	Above Average
Potential for publication	Above Average	Above Average
Potential for leadership in research	Above Average	Above Average
Match of research interests	Above Average	Above Average

POSSIBLE ANSWERS

1. Alternative A is more preferable than Alternative B
2. Alternatives A and B are equally preferable
3. Alternative B is more preferable than Alternative A

Fig. 1. Making trade-offs for applicants resumes' evaluation task.

average level on attribute 'The ability to teach our students'?

These questions have to be asked for each pair of attributes and for each level on an attribute scale. The same information can be obtained in a much more natural way in the form of comparison of two hypothetical alternatives, differing only in attainment on two attributes. This form of ordinal judgment was investigated and found to be quite valid for prefer-

Attributes	Alternative A	Alternative B	Alternative C
Ability to teach our students	Above average	Above average	Above average
Ability to teach systems analysis and DBS	Above average	Above average	Above average
Evaluation of completed research and scholarship	Above average	Above average	Above average
Potential for publication	**Above average**	**Below average**	**Above Average**
Potential leadership in research	**Above average**	**Above average**	**Below Average**
Match of research interests	Below average	Above average	Above average

Earlier you said that Alternative A is equal to Alternative B; Alternative B is equal to Alternative C. This leads to Alternative A being equal to Alternative C.

Now you say that Alternative A is more preferable than Alternative C.

What comparison(s) would you like to change: AB, BC, AC?

Fig. 2. An example of contradictory answers during trade-offs.

ence elicitation [15]. An example of such a question is presented in Fig. 1.

As a result of this approach we can construct a set of hypothetical alternatives L with the feature that the best attribute levels are obtained on all but one attribute. The number of such alternatives is not large: $N = \sum_{q=1}^{Q} n_q - 1$, where Q is the number of attributes, and n_q is the number of possible levels on a specific attribute scale. The task is to fill in the results of pairwise comparisons of alternatives from L to a matrix of size $N \times N$.

2.1. Inconsistent responses

Implementation of this or any other elicitation process involves the potential for inconsistent responses. Some mechanism to deal with such inconsistency is required. The simple way to compare the three possible answers allows one to check new judgments for consistency on the basis of transitivity. If A was said to be preferred to B and B was said to be preferred to C, then A must be preferred to C (an example of such a case is presented in Fig. 2).

The assumed transitivity of preferences and rank orderings of attribute levels make it possible to construct an effective procedure for pairwise comparisons that essentially reduces the number of required comparisons from $N \times (N-1)/2$ [see Ref. [17]].

2.2. Use of pairwise comparisons

As a result of the interview it is possible to rank order alternatives from L. Each alternative from L has only one attribute level differing from the best one. Therefore, we are able to rank order all levels on all attributes. This rank ordering is called the Joint Ordinal Scale (JOS). We are now able to compare levels on different attributes (see example JOS in Fig. 3).

The construction of the joint ordinal scale leads to a simple and easy to understand rule for comparison of multiattribute alternatives. Let us denote the rank in the joint ordinal scale of level j on attribute q as $r(x_{iq})$. This means that the hypothetical alternative from list L with the best levels on all attributes except q and with x_{jq} on the qth attribute has this rank in the order of alternatives from list L. A smaller rank identifies a better attribute level. The ideal alternative would have best levels of each attribute equal to 1.

Rank	Joint Ordinal Scale	Vector
1	Ability to teach our students is above average Ability to teach systems analysis and DBS is above average Research and scholarship completed are above average Potential for publication is above average Potential for research leadership is above average Match of research interests is above average	111111
2	Match of research interests is average	111112
3	Ability to teach systems analysis and DBS is average	121111
4	Research and scholarship completed are average	112111
5	Ability to teach our students is average Potential for publication is average Match of research interests is below average	211111 111211 111113
6	Ability to teach systems analysis and DBS is below average	131111
7	Potential for research leadership is average	111121
8	Research and scholarship completed are below average	113111
9	Potential for research leadership is below average	111131
10	Potential for publication is below average Ability to teach our students is below average	111311 311111

Fig. 3. Results of trade-offs for the applicant selection task.

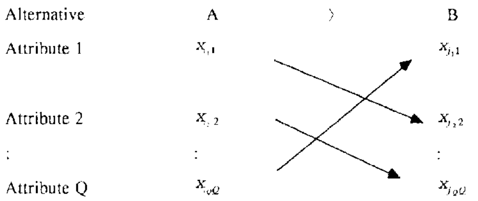

Fig. 4. Comparison of alternatives on the basis of pairwise comparison.

Let us now consider two alternatives, A and B, for comparison. $A = (x_{i(1)1}, x_{i(2)2}, \ldots, x_{i(Q)Q})$ and $B = (x_{j(1)1}, x_{j(2)2}, \ldots, x_{j(Q)Q})$. We can now substitute each attribute level in A and B with its rank in the joint ordinal scale: $A = (r(x_{i(1)1}), r(x_{i(2)2}), \ldots, r(x_{i(Q)Q}))$; $B = r(x_{j(1)1}), r(x_{j(2)2}), \ldots, r(x_{j(Q)Q}))$.

The following rule for comparison of alternatives A and B may be formulated: Rule—Alternative A is not less preferable than alternative B if for all $x_{i(q)q}$ $(q = 1,2,\ldots,Q)$ there is $x_{j(p(q))p(q)}$ such that $r(x_{i(q)q}) \leq r(x_{j(p(q))p(q)})$ and when q is not equal to q_1 than $p(q)$ is not equal $p(q_1)$. When we compare two alternatives, we try to check if A is not less preferable than B. If this is so, and at least one of the inequalities is strict, than A is preferred to B. If we have all equalities, A is equal in preference to B. If the rule does not state that A is not less preferable than B, then we try to determine if B is not less preferable than A. If this is not the case, A and B are incomparable. The applicability of the rule for preferentially independent attributes is proven in Larichev and Moshkovich [16].

The idea of the compensation procedure is shown in Fig. 4.

As a result we can obtain a matrix of pairwise comparisons of alternatives, but without the requirement of quantitative estimation of attribute levels. From this matrix a partial rank ordering of alternatives is constructed. This partial rank ordering can be used for selection of a group of better alternatives of the desired size. However, a partial order does not guarantee a specific number of alternatives to be ranked in the top levels. Yet this approach gives a logical means to use ordinal judgments for pairwise comparisons of alternatives. The results are stable to all variations in weights and values that do not violate elicited ordinal properties.

3. Analysis of results

Four faculty members participated in the evaluation and selection processes to determine a 'short list' of the better applicants for further investigation. Raters 1 and 2 (MIS) worked with the six attributes described in Table 1, while raters 3 and 4 used the five attributes for POM. Applicants were evaluated by each rater on elaborated attributes using a 10-point scale for the SMART procedure, and the three point verbal scale for ZAPROS. A group of 48 applicants were evaluated for the MIS position by raters 1 and 2. A group of 45 applicants were evaluated for the POM position by raters 3 and 4.

First, discrepancies in evaluations of applicants on separate attributes for both forms were investigated. To analyze the similarity in these evaluations for two raters, statistical analysis was applied. The hypothesis of equal means for both raters was tested. Because it is logical to present data in pairs (for each applicant evaluated by two raters) and as the assumption of normality may not be valid for our data, the nonparametric Wilcoxon signed rank test was used [12]. Results are given in Table 2.

These results show that for both pairs of raters, data using a 3-point scale (ZAPROS) were more stable than data from the SMART procedure (using a 10-point scale). These results were significant for five attributes (out of six) for raters 1 and 2 for the 3-point scale, and only for three attributes (out of six) using the 10-point scale. The results from raters 3 and 4 were similar (although on the whole, similar-

Table 2
Analysis of differences in applicant evaluations across raters

Attribute number	Raters 1 and 2		Raters 3 and 4	
	3-point scale	10-point scale	3-point scale	10-point scale
1	$p = 0.9930$	$p = 0.2383$	$p = 0.2471$	$p = 0.0768$
2	$p = 0.0202$	$p = 0.7498$	$p = 0.5883$	$p = 0.5776$
3	$p = 0.6169$	$p = 0.1236$	$p = 0.0186$	$p = 0.0557$
4	$p = 0.8677$	$p = 0.1968$	$p = 0.6909$	$p = 0.2134$
5	$p = 0.8247$	$p = 0.0340$	$p = 0.4566$	$p = 0.6643$
6	$p = 0.9274$	$p = 0.0571$		
	$n = 48$		$n = 45$	

Null hypothesis was that applicants were similar. Therefore, a high p indicates raters had similar ratings for applicants.

ity in evaluating applicants was lower than in the case of raters 1 and 2). For the 3-point scale the results were significant for all attributes except the third. For the 10-point scale significant results were identified for attributes 2, 4, and 5. This could be expected considering the fewer number of possibilities on the 3-point scale. But as can be seen, there are cases when data for the 10-point scale are more similar than the analogous data on the 3-point scale (see, e.g., the second attribute for raters 1 and 2, and the fifth attribute for raters 3 and 4). Differences are greater for the 3-point scale. Additional analysis of the data showed essentially different approaches of raters 1 and 2 with respect to attribute two. The first attribute, 'Ability to teach our students,' deals with the applicant's ability to teach the level of students at the particular university. The attribute, 'Ability to teach systems analysis and database management systems' deals with the applicant's level of knowledge. The first rater treated these attributes as equivalent, while the second rater had a stricter standard for the second attribute relative to the first. Analysis of the data showed that rater 2 marked more applicants as average for the 10-point scale, while rater 1 spread his ratings more evenly between above average, average, and below average levels. This resulted in a large number of negative differences in evaluations of applications by the two raters using the 3-point scale. In the majority of those cases where these two raters had different ratings, the second rater gave lower ratings. At the same time this unidirectional tendency did not have as much impact when the 10-point scale was used. The Wilcoxon similarity test is based on the equality of the number of positive and negative differences in pairs of data. As a result the 3-point scale was measured as being much less similar for both raters than was found when the 10-point scale was used. The results for the 3-point scale revealed essential differences in the attitude of raters towards the attribute, while the 10-point scale hid these dissimilarities.

According to the opinions of the raters it was easier for them to express their evaluations using definite words (as in ZAPROS). Subjects commented that they agreed that a particular applicant was average on a particular attribute, but for one rater average was 5 to 7, while for the second rater average was from 4 to 6. This noted variance in scoring supports the observations of others that people often can agree on verbal definitions for an object, but have greater difficulty in numerical estimation of the same concept [5,10].

It is therefore necessary to state that there were essential differences in evaluation of applicants over attributes, especially using the 10-point scale. We will next look at subsets of better alternatives identified by the different raters, and evaluate their procedures.

3.1. *Results on the selection of a group of better alternatives*

Following the traditional approach, attribute weights were developed collectively beforehand. Thus to investigate differences among raters it was only possible to use individual rater evaluations with the same attribute weights for rank ordering of alternatives. The committee was given the task of selecting 15 applicants for further processing. The data on the best 15 applicants are given in Table 3. If there was a major difference in ranking across two raters, that rater with the higher ranking was given the opportunity to argue his (or her) case.

Table 3
First fifteen most preferable alternatives for different raters

Rank	Raters					
	1	2	Average	3	4	Average
1	29	44	29[a]	15	26	26[a]
2	14	29	45[a]	33	31	15[a]
3	45	45	14[a]	26	32	33[a]
4	35	2	35[a]	12	36	36[a]
5	20	14	20[a]	20	15	32[a]
6	47	35	47[a]	19	33	12[a]
7	21	26	5[a]	13	23	13[a]
8	4	28	4	36	3	3[a]
9	3	8	21	24	1	20[a]
10	16	42	3	25	45	31
11	5	20	16	35	40	45[a]
12	36	47	28	34	14	14[a]
13	19	38	8	3	12	23
14	41	5	41	32	20	43
15	7	27	38	14	43	24

[a] Indicates that raters in that set included that alternative (candidate) among the top fifteen candidates rated.
Raters 1 and 2 were MIS faculty; raters 3 and 4 were POM faculty.

We see that raters 1 and 2 agreed on only seven applicants as belonging to the group of top fifteen (applicants 5, 14, 20, 29, 35, 45, and 47). All of these seven were in the top fifteen on the basis of average evaluations on attributes. As far as other elements of the actual choice are concerned, they represent five additional applicants from the list of rater 1 and only three from the list of rater 2. Data for raters 3 and 4 were somewhat different. They shared eleven applicants on their lists of fifteen. Of the other four, one came from the list of rater 3 and three from the list of rater 4.

Which is better, to simply average evaluations and use these for alternative scores, or to rank order alternatives for each rater individually to study the results and form the final choice considering this additional information? The first approach is easy to use and to defend. The second approach (as stated by the raters) gives you more information and allows you to make selections in a more responsible way. Further, in our case all raters agreed that the number of applicants in the group actually chosen could range from eight to fifteen, thus providing the opportunity to enlarge the final selected group in accordance with some principle (e.g., first select applicants presented in choices of both raters; if the group is too small, add an applicant from pools of each rater with the highest rank in this group, but not present in the final group, and so on). Such a procedure could help to form the selected group of applicants that were considered by at least one rater to be quite good.

This was followed by application of ZAPROS. Each rater formed a preferable group of about fifteen elements (ZAPROS provides a partial order, and therefore does not guarantee a complete rank order of all alternatives). Results for all raters are presented in Fig. 5.

For the first rater this group contains 13 applicants. The second rater has 12, the third rater 13, and the fourth rater 12. Analogous to the results from SMART, raters 1 and 2 had only six applicants mutually appearing on their lists. Raters 3 and 4 had nine mutual applicants on their final groups.

Comparison of results for the same raters across the two methods show that groups on the whole are quite similar. The first rater had 13 applicants on the preferred list from ZAPROS. Of these 13, 11 were

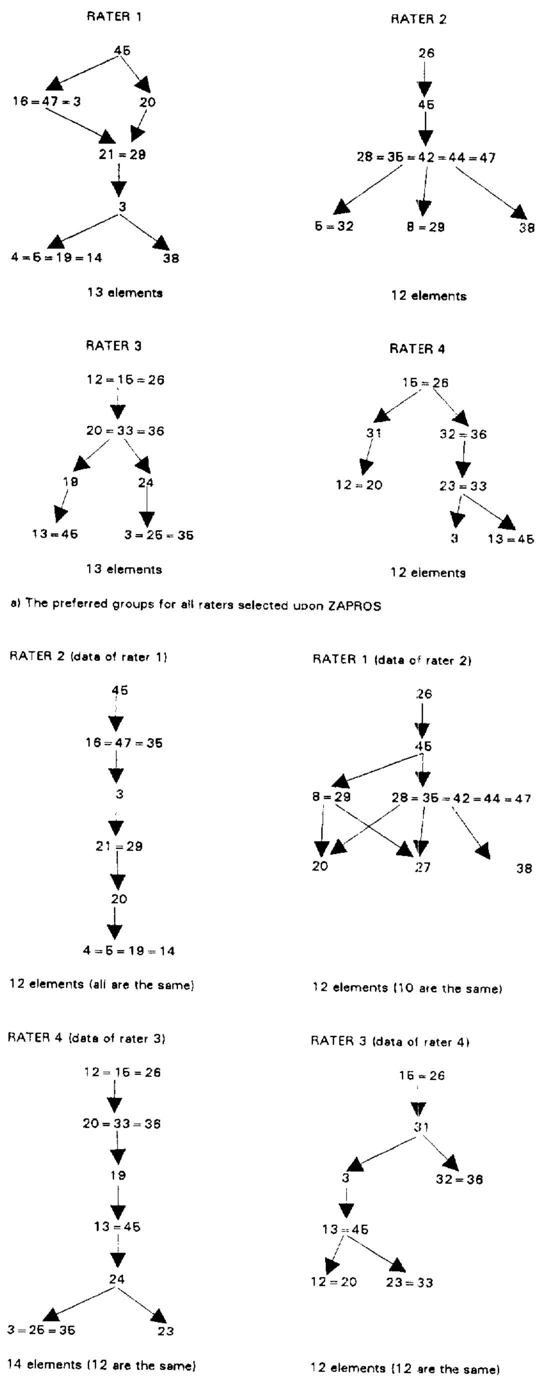

Fig. 5. The stable preferred group for rater.

Table 4
Rank ordering of attributes on importance

Raters	Rank orderings of attributes	Attribute weights					
		C1	C2	C3	C4	C5	C6
Rater 1	C1 = C4 > C5 > C3 > C2 > C6	0.25	0.1	0.15	0.25	0.2	0.05
Rater 2	C1 = C3 = C4 > C2 = C5 = C6	0.2	0.13	0.2	0.2	0.14	0.13
Initial mutually agreed rating	C4 = C5 > C1 = C2 = C6 > C3	0.15	0.15	0.09	0.23	0.23	0.15
Rater 3	C1 > C3 > C4 > C2 > C5	0.3	0.15	0.25	0.2	0.1	
Rater 4	C4 > C2 > C1 = C3 = C5	0.15	0.25	0.15	0.3	0.15	
Initial mutually agreed rating	C1 = C2 = C5 > C3 = C4	0.24	0.24	0.14	0.14	0.24	

Attribute weights for individual raters were inferred from the rank orderings obtained from ZAPROS. Initial mutually agreed upon ratings were from SMART analysis.

also on the list of 15 obtained using SMART. For rater 2, with 12 applicants on the preferred list, 11 of which were on the list of 15 obtained using SMART. Rater 3 had a list of 13 applicants from ZAPROS, of which 11 were on the list of 15 obtained from SMART. For rater 4, 10 of the 12 applicants obtained from ZAPROS that had been on the list of 15 obtained from SMART. This implies stability of results across methods.

On the other hand, statistical analysis with respect to the rank ordering of alternatives identified significant differences in results from each rater across methods, as well as between raters using the same method. This result supports analogous findings in experiments with students using different multiattribute techniques for rank ordering of the same alternatives [17,18].

3.2 *Analysis of attribute importance and attribute values*

Further analysis examined differences in preferences expressed by raters across the two methods, and to find out the level of influence on the result of preferences and of alternative evaluations over attributes. First, using information obtained from ZAPROS, the underlying structure of attribute importance (presented in the joint ordinal scale) was identified. Rank orderings of attributes for each rater are given in Table 4. Although different weights can be used to provide the same rank order of attributes, raters revealed their preferences in the presented weights. Investigations by Barron and Barrett [2] showed that different schemas of acquiring weights on the basis of their rank ordering provided very stable results in selecting the best alternative from a predefined set of alternatives.

We see that there are essential differences in rater attitudes towards attributes. More interesting is the fact that there are also significant differences between these ratings and those obtained from SMART. This supports the observation that people can think of different things when they are asked about attribute importance, although attribute weights are used in a definite way requiring meaning [3,17,18].

To evaluate the extent these differences influenced the results, the following analysis was conducted. The joint ordinal scale constructed for one rater was used on a set of alternatives evaluated by the other rater. Results are presented in Fig. 5.

Table 5
Rank correlations (Spearman) of different alternatives' rankings

Pair of rank orderings	Spearman correlation	
	ZAPROS	Traditional
Different preferences and different data		
Rater 1 vs. rater 2	0.0304	0.0147
Rater 3 vs. rater 4	0.2835	0.0451
Different preferences but the same data		
Rater 1 vs. rater 2 with data of rater 1	0.9438	0.989
Rater 2 vs. rater 1 with data of rater 2	0.7493	0.9745
Rater 3 vs. rater 4 with data of rater 3	0.9503	0.9841
Rater 4 vs. rater 3 with data of rater 4	0.8475	0.9558
The same preferences but different data		
Rater 1 vs. rater 1 with data of rater 2	0.1995	0.0312
Rater 2 vs. rater 2 with data of rater 1	0.0133	0.0128
Rater 3 vs. rater 3 with data of rater 4	0.1076	−0.0699
Rater 4 vs. rater 4 with data of rater 3	0.3394	0.0496

Weights of rater 1 and data of rater 1
29 → 45 → 35 → 14 → 20 → 47 → 21 → 4 → 16 → 3 → 5 → 19 → 36 → 43 → 41

Weights of rater 2 and data of rater 2
44 → 2 → 29 → 45 → 26 → 35 → 28 → 14 → 8 → 20 → 5 → 42 → 47 → 38 → 27

Weights of rater 1 and data of rater 2
29 → 2 → 44 → 45 → 14 → 26 → 35 → 8 → 28 → 38 → 5 → 42 → 47 → 20 → 27

Weights of rater 2 and data of rater 1
29 → 45 → 35 → 20 → 14 → 47 → 4 → 16 → 21 → 3 → 5 → 43 → 41 → 19 → 36

Weights of rater 3 and data of rater 3
15 → 33 → 26 → 12 → 20 → 19 → 13 → 36 → 24 → 25 → 35 → 45 → 3 → 32 → 14

Weights of rater 4 and data of rater 4
26 → 31 → 32 = 36 → 15 → 33 → 23 → 3 → 13 → 45 → 40 → 14 → 12 → 20 → 43

Weights of rater 3 and data of rater 4
31 = 32 = 36 → 26 → 33 → 23 → 40 → 15 → 3 → 13 → 45 → 20 → 12 → 14 → 43

Weights of rater 4 and data of rater 3
15 = 33 → 20 → 12 = 26 → 19 → 36 → 13 → 24 → 35 → 3 → 25 → 45 → 32 → 14

Fig. 6. Rank orderings of alternatives with different preference structures and data.

These results show that the ranking of alternatives are similar for different joint ordinal scales used for the equally evaluated alternatives for both pairs of raters. Statistical analysis of this similarity is given in Table 5.

This supports the idea that differences in evaluation of alternatives over attributes influences the final ranking of alternatives more than the applied preference structure. This was demonstrated using ZAPROS, which used 3-point scales (and in this case differences among raters in evaluation of alternatives over attributes was smaller than in the 10-point case). Therefore, additional analysis was conducted.

Raters were asked to assign attribute weights individually in accordance with the rank ordering of attributes obtained through ZAPROS. These results were given in Table 4. These weights were used to calculate alternative scores for each rater, and in addition, using the other rater's evaluation of alternatives over attributes. Results of rank correlation between these rankings (Spearman correlation coefficients) are presented in the last column of Table 5. The rankings obtained are given in Fig. 6.

These results match the previous results in this study, supporting the idea that the structure of alternatives influences the result to a larger extent than does the applied preference model.

3.3. Discussion

Implementation of different procedures for evaluation of multiattribute alternatives is widespread in real practice. Usually people prefer to use simple and easily understandable procedures (as SMART) for such cases. When there are several decision makers, data on attribute importance and alternative scores over these attributes are averaged to reach a final decision.

The results obtained in this real case study using real decision makers support previous findings based on experimental work.

(1) Different procedures used for elicitation of attribute importance may lead to different results even for the same decision maker. Table 4 gave results demonstrating differences in attribute weights from the two methods used.

(2) Different multiattribute procedures used for rank ordering of the same alternatives may lead to different results even for the same decision maker.

SMART rankings by rater are given in Table 3. ZAPROS rankings by rater are given in Fig. 5. The first choice by ZAPROS was ranked 3rd for rater 1, and 7th by rater 2. The partial order of ZAPROS gave three top rated alternatives for rater 3, ranked 1, 3 and 4 by the SMART procedure. For rater 4, the two top rated alternatives from ZAPROS were ranked 1 and 5 by the SMART procedure. Thus, there is obvious correlation, but differences in selection of the top choice across methods (more obvious for raters 1 and 2).

(3) Our experience indicates that differences in alternative scores over attributes may influence the result to a larger extent than do differences in attribute weights.

We see these results providing some insight into group decision making: (1) committee members usually agree on some averaged presentation of their preferences to reach some average decision, but find it easier to obtain a compromise position when their differences with other committee members were identified for them; (2) evaluation of alternatives over attributes using the stabler verbalized scales (as in ZAPROS) led to greater similarity of results across committee members; (3) differences in attribute weights, as well as the method used, influence the results less than evaluation of alternative scores on attributes; (4) rank ordering of alternatives is highly influenced by the parameters of the task, but the contents of a larger group of better alternatives may be rather stable across different environments.

4. Conclusions

Two MAUT techniques were compared in the task of identifying a 'short list' of applicants for further consideration for academic positions. Technique SMART provides a cardinal measure of each alternative that is easily used to identify the top candidates. ZAPROS, relying on ordinal, verbal input from the decision makers, was found to provide better insight into individual differences in preference structures.

A number of process features were inferred by the group.

The committee members found that a less quantitative method was better for revealing inner preference structures. This information was then used to develop a group strategy for the evaluation process.

Evaluation of alternatives against attributes was considered important by the group. These scorings were found to influence the final selections a great deal, even more than the weights on attributes.

During this experience it was noticed that attribute weights and scores of alternatives over these attributes are not independent. If an attribute is considered to be very important, raters tended to give lower scores over these attributes. On the other hand, if attributes were considered minor, raters tended to give high scores to all alternatives. This feature violates the fundamental assumptions of MAUT, and should be investigated in future research.

References

[1] R.J. Aldag, D.J. Power, An empirical assessment of computer-assisted decision analysis, Dec. Sci. 17 (3) (1986) 572–588.

[2] F.H. Barron, B.E. Barrett, Decision quality using ranked attribute weights, Manage. Sci. 42 (11) (1996) 1515–1523.

[3] V. Belton, A comparison of the analytic hierarchy process and a simple multiattribute value function, Eur. J. Oper. Res. 26 (1986) 7–21.

[4] K. Borcherding, T. Eppel, D. von Winterfeldt, Comparison of weighting judgments in multiattribute utility measurement, Manage. Sci. 37 (12) (1991) 1603–1619.

[5] D.V. Budescu, S. Weinberg, T.S. Wallsten, Decisions based on numerically and verbally expressed uncertainties, J. Exp. Psych.: Human Percep. Perfor. 14 (1988) 281–294.

[6] R.M. Dawes, Rational Choice in an Uncertain World, Harcourt Brace Jovanovich, San Diego, 1988.

[7] R.M. Dawes, B. Corrigan, Linear models in decision making, Psych. Bull. 81 (1974) 95–106.

[8] W. Edwards, How to use multiattribute utility measurement for social decisionmaking, IEEE Trans. Sys. Man, Cybernet. SMC-7 (1977) 326–340.

[9] H. Einhorn, R. Hogarth, Unit weighting schemes for decision making, Org. Behav. Human Dec. Proc. 13 (1975) 171–192.

[10] I. Erev, B.L. Cohen, Verbal versus numeric probabilities: efficiency, biases, and the preference paradox, Org. Behav. Human Dec. Proc. 45 (1990) 1–18.

[11] A. Goicoechea, E.Z. Stakhiv, F. Li, Experimental evaluation of multiple criteria decision models for application to water resources planning, Water Resour. Bull. 28 (1) (1992) 89–102.

[12] M. Hollander, D.A. Wolfe, Nonparametric Statistical Methods, Wiley, New York, 1973.

[13] D. Kahneman, P. Slovic, A. Tversky (Eds.), Judgment Under Uncertainty: Heuristics and Biases, Cambridge Univ. Press, Cambridge, 1982.

[14] R. Keeney, H. Raiffa, Decisions with Multiple Objectives: Preference and Value Tradeoffs, Wiley, New York, 1976.

[15] O.I. Larichev, Cognitive validity in design of decision-aiding techniques, J. Multi-Crit. Dec. Anal. 1 (1992) 127–138.

[16] O.I. Larichev, H.M. Moshkovich, ZAPROS-LM-A method and system for rank-ordering of multiattribute alternatives, Eur. J. Oper. Res. 82 (1995) 503–521.

[17] O.I. Larichev, H.M. Moshkovich, A.I. Mechitov, D.L. Olson, Experiments comparing qualitative approaches to rank ordering of multiattribute alternatives, J. Multi-Crit. Dec. Anal. 2 (1) (1993) 5–26.

[18] O.I. Larichev, D.L. Olson, H.M. Moshkovich, A.I. Mechitov, Numeric vs. cardinal measurements in multiattribute decision making: (How exact is enough?), Org. Behav. Human Dec. Proc. 64 (1) (1995) 9–21.

[19] D.L. Olson, Review of empirical studies in multiobjective mathematical programming: Subject learning and response to nonlinear utility, Dec. Sci. 23 (1) (1992) 1–20.

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

[21] B. Roy, D. Bouyssou, Comparison of two decision-aid models applied to a nuclear plant siting example, Eur. J. Oper. Res. 25 (1986) 200–215.

[22] P.J.H. Shoemaker, C.C. Waid, An experimental comparison of different approaches to determining weights in additive utility models, Manage. Sci. 28 (2) (1982) 182–196.

[23] P. Slovic, B. Fishhoff, S. Lichtenstein, Behavioral Decision Theory, Annu. Psych. Rev. 28 (1977) 1–29.

[24] W.G. Stillwell, D.A. Seaver, E. Edwards, A comparison of weight approximation techniques in multiattribute utility decision making, Org. Behav. Human Dec. Proc. 28 (1981) 62–77.

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

[26] E. Triantaphyllou, S.H. Mann. An examination of the effectiveness of multidimensional decision-making methods: A decision making paradox, Dec. Supp. Sys. 5 (1989) 303–312.

[27] D. von Winterfeldt, W. Edwards, Decision Analysis and Behavioral Research, Cambridge Univ. Press, Cambridge, 1986.

[28] H. Wainer, Estimating coefficients in linear models: It don't make no nevermind, Psych. Bull. 85 (1976) 213–217.

[29] M. Weber, F. Eisenfuhr, D. von Winterfeldt, The effects of splitting attributes on weights in multiattribute utility measurement, Manage. Sci. 34 (1988) 431–445.

Helen M. Moshkovich received her M.S. in MIS (1973) from the Moscow University of Management. She received her Ph.D in Management Science (1984) from the Institute for System Studies of the USSR Academy of Sciences. She is currently Visiting Associate Professor in the School of Business at the University of West Alabama. Dr. Moshkovich's primary research interests include multicriteria decision making, human decision making behavior, and intelligent decision support systems. She has been involved in projects such as R&D planning, oil technology selection, and strategic decision support. She has published over 40 articles in leading Russian journals, as well as in Organizational Behavior and Human Decision Processes, Acta Psychologica, and Decision Sciences. She is coauthor of three books.

Robert E. Schellenberger is Professor of Decision Sciences at East Carolina University. He served as Chair of the Department of Management at Southern Illinois University, Director of Program Evaluation at the Babcock School of Management at Wake Forest University, Chair of the Department of Management at Temple University, and Chair of the Decision Sciences Department at East Carolina University. He is author of ten books and monographs as well as of numerous journal articles.

David L. Olson received his Ph.D. in Business Administration (1981) from the University of Nebraska—Lincoln. He is currently the Business Analysis Excellence Fellow and Professor in the Department of Business Analysis and Research at Texas A&M University. Dr. Olson's primary research interest is in multicriteria decision making. He has published over forty refereed articles in journals such as Organizational Behavior and Human Decision Processes, Decision Sciences, European Journal of Operational Research, and other journals. He is author of one book, and coauthor of four other books.