April, 1997

On the convergence of multiattribute weighting methods

Mari Pöyhönen and Raimo P. Hämäläinen

Systems Analysis Laboratory
Helsinki University of Technology
P.O.Box 1100, 02015 HUT, Finland

Abstract

Five multiattribute weighting methods were compared in an Internet experiment. This is the first experiment where the subjects created the alternatives and attributes themselves. Each subject used five methods to assess attribute weights, one version of the Analytic Hierarchy Process (AHP), DIRECT weighting, Simple Multiattribute Rating Technique (SMART), SWING weighting and TRADEOFF weighting. They can all be used following the principles of multiattribute value theory. Furthermore, each of them asks decision makers to give numerical estimates of weight ratios although the elicitation questions are different. In earlier studies, however, these methods have yielded different weights. Our results suggest that weights are different because the methods explicitly or implicitly lead the decision makers to choose their responses from a limited set of numbers. The other consequences are that the spread of weights and the inconsistency between the preference statements depend on the number of attributes that a decision maker considers simultaneously.

Key words: Multi-Attribute Value Theory, Behavioral Decision Making, Attribute Weighting, Decision Support.

Today there are many user friendly software products available for the practitioners of multiattribute value analysis. These attractive tools make it possible to do analysis without the help from an experienced decision analyst. At the same time there is growing interest in using the methodology in important real life problems. Thus it is of utmost importance to learn to avoid the behavioral biases in these models. Still, our understanding of the behavioral aspects of multiattribute weighting procedures is insufficient (see e.g. Weber and Borcherding, 1993). This paper builds up this knowledge by means of a novel interactive Internet experiment comparing multiattribute weighting techniques based on the multiattribute value theory.

The methods compared are four versions of the Analytic Hierachy Process (AHP) (Saaty, 1980; 1994; Salo and Hämäläinen, 1997), DIRECT weighting, Simple Multiattribute Rating Technique (SMART) (Edwards, 1977; von Winterfeldt and Edwards, 1986), SWING weighting (von Winterfeldt and Edwards, 1986) and TRADEOFF weighting (Keeney and Raiffa, 1976). All of these methods are based on eliciting estimates of weight ratios. In earlier studies, however, these methods have yielded different weights. For example, the experiment of Schoemaker and Waid (1982) showed that the weights derived with three of these methods were different. Later this result has been confirmed for different combinations of the methods (see e.g. Belton, 1986; Borcherding et al., 1991; Olson et al., 1996). Yet, the origins of the differences in the resulting weights remain unclear and the researchers are still seeking for superior procedures.

Since the early experiments the methods have been refined. SMART has developed into SMARTS and SMARTER (Edwards and Barron, 1994) and there are several versions of the AHP (see e.g. Belton and

Gear, 1983; Lootsma, 1993; Schoner et al., 1993; Olson and Dorai, 1992). In this experiment we use the original AHP with two different ways of questioning and with a new scale (Salo and Hämäläinen, 1997). We consider the AHP as a variant of multiattribute value theory following Salo and Hämäläinen (1997). Thus the weights derived with the AHP are expected to be comparable to weights obtained with the other methods.

The aim of this experiment is to re-examine the properties of the methods in a computer aided study. To increase the reliability of the results we tried to make the decision task as natural and personal as possible. In our experiment the participants were able to structure the problem individually. Thus the subjects weighted different attributes and the number of attributes varied between the subjects. This is a new feature and improvement compared to many of the earlier experiments. This is also the first time a multicriteria decision making experiment is carried out by using the Internet environment with an interactive interface. The Internet provides an efficient way of approaching subjects from different cultural and professional backgrounds. We discuss some aspects in the use of the Internet in Appedix A.

1 Attribute weighting

Multiattribute value analysis derives an overall value score for each alternative. This is composed of the ratings of the alternatives with respect to each attribute and of the weights of attributes. If the attributes are mutually preferentially independent one can use an additive value function to aggregate the component values (Keeney and Raiffa, 1976). The value of an alternative x is then

$$v(x) = \sum_{i=1}^{n} w_i v_i (x_i) \,,$$

where x_i is the consequence of an alternative x for an attribute i ($i = 1 , \dots , n$), $v_i(x_i)$ is the rating of the consequence, and w_i is the weight of an attribute i. Weights are normalized to sum up to one. Component value functions $v_i(\cdot)$ get values between 0 and 1. Weights w_i indicate the relative importance of an improvement in one attribute from its worst level to its best level compared with changes in other attributes. There are many techniques to elicit attribute weights, for a review see Stewart (1992) and Weber and Borcherding (1993). The methods studied in this experiment and the way they are used are described below.

In DIRECT WEIGHTING the decision maker gives numbers to describe the attribute weights. The decision maker is asked, for example, to divide 100 points among the attributes. In this experiment we did not give a total number of points to be divided but the subjects were asked to give any numbers they liked to reflect the weights.

With SMART the weights are elicited in two steps (Edwards, 1977; von Winterfeldt and Edwards, 1986):

1. Rank the importance of the changes in the attributes from the worst attribute levels to the best levels.

2. Make ratio estimates of the relative importance of each attribute relative to the one ranked lowest in importance.

Step two usually, and also in this experiment, begins by assigning 10 points to the least important attribute. The relative importances of the other attributes are evaluated by giving them points from 10 upwards. Edwards and Barron (1994) listed shortcomings in this original procedure. They stressed that the importance of attributes should clearly be related to the attribute ranges. They also presented a new version, SMARTER, that only uses the rank of attributes to derive weights.

In SWING (von Winterfeldt and Edwards, 1986) the decision maker is asked to consider a situation where he is stuck with a hypothetical alternative that has all the attributes at their worst levels. First the decision maker is asked to move one attribute to its best level: "Select an attribute that you first would like to change to the best level and assign 100 points to this most important attribute". Next the decision maker is asked to choose an attribute change from the worst to the best level which he considers to be the second most desirable improvement and to assign points less than 100 to that attribute change. This procedure is continued with all the remaining attributes.

In the TRADEOFF procedure (Keeney and Raiffa, 1976) the decision maker compares two hypothetical alternatives that differ in two attributes only. The other attributes are kept on the same, fixed levels. Let *x* and *y* denote these two alternatives and let the indexes 1 and 2 refer to the attributes. The decision maker is asked to consider two hypothetical alternatives with the attribute pairs (x_1, x_2) and (y_1, y_2) and to adjust one of the attributes until the alternatives become equally preferred. In order to choose which attribute should be moved one has to know the rank of attributes or which one of the hypothetical alternatives is preferred. This is required to avoid a situation where a decision maker would like to adjust the attribute outside the range defined for it in the beginning. The indifference statement gives an equation

$$w_1 v_1(x_1) + w_2 v_2(x_2) = w_1 v_1(y_1) + w_2 v_2(y_2),$$

where w_1 and w_2 are the unknown attribute weights. The *n-1* indifference statements, known values $v_i(\cdot)$, and the normalization condition yield *n* equations that are used to solve the *n* weights. It is noteworthy that the calculation of the TRADEOFF weights requires that the component values $v_i(\cdot)$ are known for the whole attribute range. More details of the TRADEOFF weighting can be found from Fischer (1995).

The weighting of attributes in the AHP (Saaty, 1980, 1994; Salo and Hämäläinen, 1997) is based on estimates of weight ratios similarly to SMART and SWING. The decision maker is asked to compare the importance of two attributes at a time: "Which one of these two attributes is more important, and how much more important?" Decision makers give weight ratios to indicate the strength of their preferences by using integers from one to nine. Each integer is also associated with a verbal expression. Variants of the AHP studied here are obtained by changing the evaluation scales (see e.g. Lootsma, 1993; Salo and Hämäläinen, 1997). In the AHP the weight ratios are asked for all pairs of attributes. Usually the weights are derived by the principal eigenvector of the comparison matrix. Other estimation methods, such as the logarithmic least squares method, can be used to derive the weights as well.

2 Questions of interest in the experiment

Many earlier studies have shown that the methods yield different attribute weights (for a review see Weber and Borcherding 1993). For example, Schoemaker and Waid (1982) analyzed the differences between a regression based method, the AHP, TRADEOFF, and the DIRECT weighting in an experiment where all the subjects had the same task with four attributes. The differences in attribute weights were described by the average attribute weights and by the average difference between maximum and minimum weights. DIRECT weighting was an outlier: the spread of the direct weights were smaller than the spread of weights with the other methods. Borcherding et al. (1991) measured the convergence between SMART, SWING, TRADEOFF and pricing out techniques by the correlation between the weights of attributes. Weights derived with the pricing out technique did not correlate with any of the other techniques. The highest correlation was found between SMART and SWING.

In this experiment we want to repeat some of the previous analysis in a new framework where the decision task is not fixed to be the same for all the subjects. The basic assumption is that all the weighting methods yield similar weights. In this case the practitioners do not need to worry about which method to choose. Our first hypothesis is that a decision maker should get the same rank for the alternatives with all the weighting methods (Hypothesis 1, Exhibit 1). It is also studied whether the alternative ranked first is the same with all the methods. However, there may be differences in attribute weights even if the rank of alternatives remains the same. For example, Schoemaker and Waid (1982) found that all the methods in their experiment indicated the rank of alternatives similarly although the attribute weights were different. This experiment compares methods that are all based on multiattribute value theory and on estimates of weight ratios. Thus our basic hypothesis is that the weights derived with different methods should be the same (Hypothesis 2, Exhibit 1).

Exhibit 1: *Hypotheses of the experiment*

H1: All weighting methods give the same rank for the alternatives.

H2: AHP, DIRECT, SMART, SWING and TRADEOFF yield same weights

H3: The spread of attribute weights is same for all the methods, but depends on the number of attributes considered simultaneously.

H4: The spread of the AHP weights and inconsistency between the AHP comparisons depend on the evaluation scale used.

H5: The explicit presentation of attribute ranges affects weights.

H6: The inconsistency between statements depends on the number of attributes.

The spread of weights, for example the maximum ratio of weights or the difference between maximum and minimum weights, is used to describe the properties of weight sets (Schoemaker and Waid, 1982; Fischer, 1995). If the weights are the same then the spread of weights is also the same. However, the spread of weights may depend on the number of attributes the decision maker considers simultaneously (Hypothesis 3, Exhibit 1). The resulting weights are also restricted by the upper and lower limits of the weight ratios (Salo and Hämäläinen, 1997). The AHP explicitly uses an evaluation scale with a restricted maximum for the weight ratios and with fixed numbers. Decision makers may also unintentionally restrict their use of numbers with any of the methods.

With the AHP we study how the change of the evaluation scale affects the weights (Hypothesis 4, Exhibit 1). Schoemaker and Waid (1982) found that the AHP produced a larger spread of weights than the other weighting methods. Belton (1986) compared SMART weights and the AHP weights. A result was that the weights produced by the AHP were unevenly dispersed over the possible weight range whereas direct weights covered the range evenly. Schoner and Wedley (1989) found that the AHP led to less accurate results than a procedure similar to SMART in an experiment where the real weights of the objects were known. These observations can be explained to originate from the evaluation scale used in the AHP. Salo and Hämäläinen (1997) suggested new balanced scales derived from the idea of evenly dispersed weights. The experiment of Pöyhönen et al. (1997) showed that the balanced scale (see Exhibit 2) increases the accuracy of the results, decreases the spread of weights and inconsistency of the comparisons compared to the 1-to-9 scale. Both scales are used in this experiment. Later Budescu et al. (1996) and Vrolijk and Huizingh (1996) have also concluded in their experiments that the properties of the AHP weights, for example the spread of weights, depend on the selection of the evaluation scale.

Exhibit 2*: The* AHP *scales.*

	Scale	
Verbal statement	1-to-9	Balanced
Equally important	1	1.00
-	2	1.22
Slightly more important	3	1.50
-	4	1.86
Strongly more important	5	2.33
-	6	3.00
Very strongly more important	7	4.00
-	8	5.67
Extremely more important	9	9.00

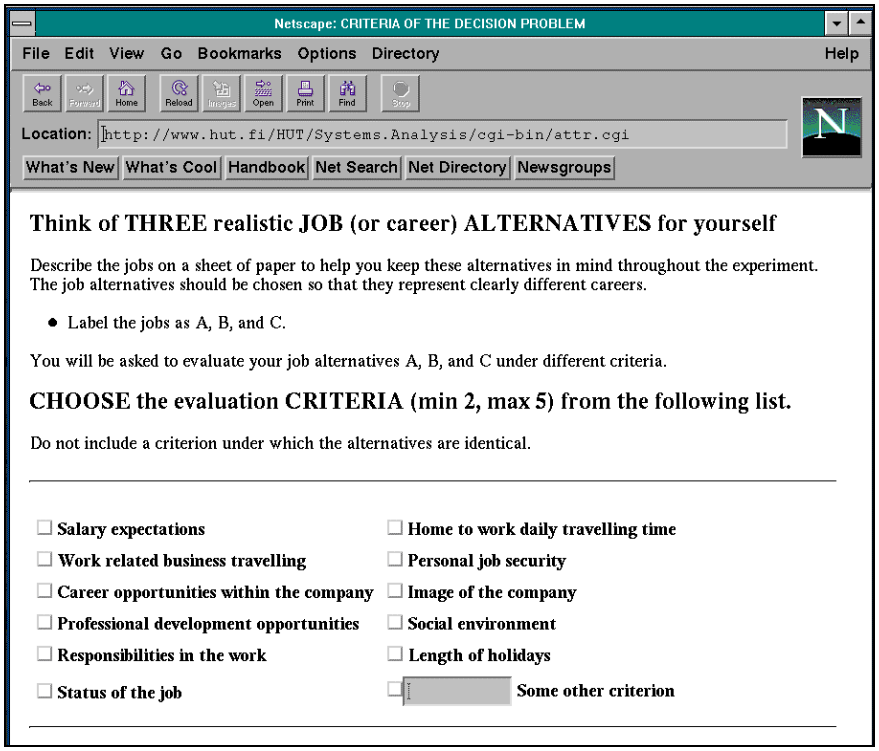
The decision makers may not always be able to state their responses accurately if they choose their responses from a limited set of numbers. For example, in the AHP they only use integers from one to nine and in SMART it can easily happen that one only uses multiples of 10 as the least important attribute is given a reference score of 10. This may be one reason for the so called range insensitivity suggesting that decision makers do not adjust the weights properly as the attribute range varies. Von Nitzsch and Weber (1993) used SMART and a regression based method and found that with SMART the decision makers were particularly range insensitive. Fischer (1995) did a similar study to further examine the range effect and found that DIRECT weighting was range insensitive while SWING and TRADEOFF weights better reflected the changes in the ranges of attributes. In our experiment the attribute ranges will be presented clearly with all the other methods except with the AHP where the attribute ranges are either presented or not. We will study how this affects the weights (Hypothesis 5, Exhibit 1).

Winterfeldt and Edwards (1986, p. 283) state that "Inconsistencies (between weight ratios) inevitably arise unless a respondent recognizes and numerically establishes consistency." In most of the previous studies, however, inconsistency analysis has not been carried out. In this experiment we will study inconsistencies with the AHP and with the TRADEOFF technique. We want to point out that so far the inconsistencies in weight ratio statements with TRADEOFF have received very little attention. Borcherding et al. (1991) reported that in their experiment decision makers found it difficult to give consistent TRADEOFF statements. Regardless of this, TRADEOFF weights have been used to validate the weights derived with other methods (von Nitzsch and Weber, 1993). We think that the dependency between inconsistencies of the statements and the number of attributes compared simultaneously needs further attention (Hypothesis 6, Exhibit 1).

3 The experiment

3.1 Experimental procedure

In a decision making experiment the participants should be interested in the test task and feel that they own the problem. This is likely to help to get reliable responses describing the decision makers' true preferences. Problems of public interest, such as environmental decisions, could be used but they tend to require a lot of country specific information. In an international Internet experiment the test task needs to be of general interest in different cultures. We chose the evaluation of job alternatives as our test problem. The participants defined the alternatives and selected the attributes (called criteria throughout the experiment) personally. The subjects were first asked to think of three job or career alternatives that they found relevant in their own situation. Then they selected two to five attributes from a shortlist or defined an attribute of their own (see Exhibit 3).



Location: http://www.hut.fi/HUT/Systems.Analysis/cgi-bin/attr.cgi

Think of THREE realistic JOB (or career) ALTERNATIVES for yourself

Describe the jobs on a sheet of paper to help you keep these alternatives in mind throughout the experiment. The job alternatives should be chosen so that they represent clearly different careers.

- Label the jobs as A, B, and C.

You will be asked to evaluate your job alternatives A, B, and C under different criteria.

CHOOSE the evaluation CRITERIA (min 2, max 5) from the following list.

Do not include a criterion under which the alternatives are identical.

□ **Salary expectations**
□ **Work related business travelling**
□ **Career opportunities within the company**
□ **Professional development opportunities**
□ **Responsibilities in the work**
□ **Status of the job**
□ **Home to work daily travelling time**
□ **Personal job security**
□ **Image of the company**
□ **Social environment**
□ **Length of holidays**
□ ______ **Some other criterion**

Exhibit 3: *The interface for selecting attributes in the job evaluation task.*

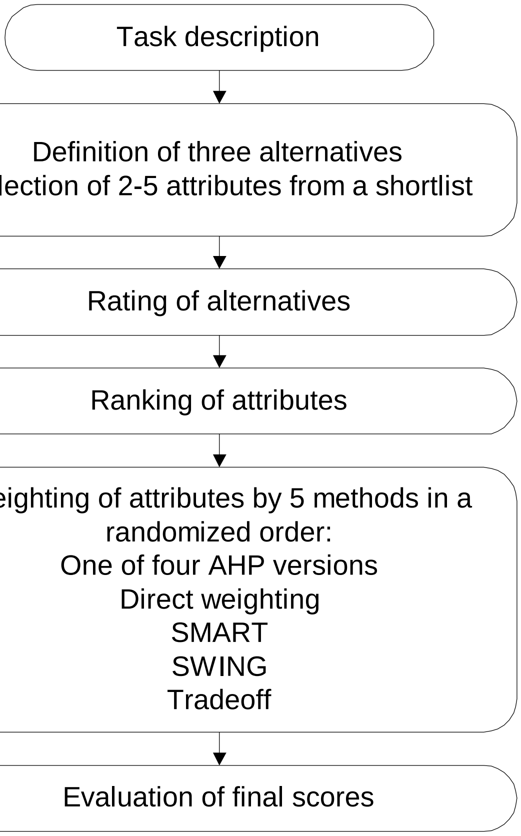

Task description
Definition of three alternatives
Selection of 2-5 attributes from a shortlist
Rating of alternatives
Ranking of attributes
Weighting of attributes by 5 methods in a randomized order:
One of four AHP versions
Direct weighting
SMART
SWING
Tradeoff
Evaluation of final scores

Exhibit 4*: The steps in the experiment*

Next the subjects rated the alternatives with respect to the attributes that they had chosen. The worst alternative was assigned a score of 0 and the best 100. Some attributes, like salary, have a natural measurement scale. For these attributes the subjects rated their alternatives on this scale. The given outcomes were converted into value scores through linear value functions. The assumption of linear value functions affects the overall scores of alternatives. However, the shape of the value function affects the attribute weights in the TRADEOFF weighting only (Fischer, 1995).

The TRADEOFF procedure is problematic with an experimental task like this one because the method assumes that the attributes are measured on a continuous scale. In this experiment we assumed that the subjects are able to give the ratings directly also for hypothetical consequences and that the subjects are able to consider also discrete attributes, like "personal job security," on a continuous scale. These assumptions, however, may be too demanding and thus the TRADEOFF weights should be viewed keeping these problems in mind.

For each attribute the attribute ranges were defined to be the ranges from the worst alternative to the best. These ranges were included in the questions for each method. Before the attribute weighting the subjects ranked the importance of changes from the worst attribute levels to the best levels. These rankings were further needed to formulate the questions. The rank of attributes, for example, defined which attribute a subject was asked to adjust with the TRADEOFF questioning.

We are interested in two experimental factors: different weighting methods and the number of attributes. Each subject had either 2, 3, 4, or 5 attributes and thus the effects of the number of attributes are studied between subjects. Differences between weighting methods are studied within subjects as each participant assessed weights by using the AHP, DIRECT weighting, SMART, SWING, and TRADEOFF in a randomized order. There are four versions of the AHP questions in order to study the effects of the scales and the attribute ranges (See Exhibit 5). Each participant completed only one of these AHP versions. Thus, the comparisons between the versions are done between subjects. Both with the AHP and TRADEOFF a complete set of pairwise comparisons was asked. This gives $n(n-1)/2$ weight ratios to estimate the n weights. Both with the AHP and TRADEOFF the weights were estimated by the principal eigenvector of the comparison matrix. The participants were not able to go back to revise their earlier statements once they completed one weighting method.

***Exhibit 5**: The names used for the AHP versions*

	Scale	
Attribute ranges	1-to-9	Balanced
are not presented	AHP1	AHP2
are presented	AHP3	AHP4

At the end the overall scores of the alternatives and the contributions by each attribute were shown as bar graphs without identifying the related methods. The subjects evaluated how well the shown scores reflected their own opinions on a scale from 1 to 5 (1 = ☺, 5 = ☹).

3.2 Subjects

The subjects were recruited by publicizing the experiment through Internet's mailing lists and by sending information to academic institutions directly. Most participants were students whose instructor had asked them to do this task. The majority of them were male students with a background in either business and economics, engineering, mathematics, or information sciences (See Exhibit 6). The subjects were not required to answer all the questions and part of data were lost due to technical problems. Thus the number of subjects varies between the tables.

Exhibit 6: *Statistics of test subjects*

Total number of subjects:			**407**		
Language			**Country**		
English	166	41%	Canada	18	4%
Other	206	51%	Finland	136	33%
Unknown	35	9%	New Zealand	8	2%
			Portugal	11	3%
Age			United Kingdom	7	2%
			United States	164	40%
15-20	9	2%	Other countries with less than 5	36	9%
21-25	254	62%	Unknown	27	7%
26-30	50	12%			
31-40	37	9%	**Background**		
>41	21	5%			
Unknown	36	9%	Business and Economics	120	29%
			Engineering	105	26%
Sex			Information sciences	58	14%
			Mathematics	54	13%
Female	82	20%	Psychology	10	2%
Male	278	68%	Other	27	7%
Unknown	47	12%	Unknown	33	8%

4 Results

4.1 Overall scores of alternatives

The subjects were equally satisfied with the overall scores produced by different methods (see Exhibit 7). They were shown both the overall scores of the alternatives and the contributions by each attribute. The similarity of the order of the alternatives within a subject was studied by focusing on the alternative ranked first. The alternative with the highest overall score is the same with all the methods for 281 of 407 subjects (69%). The percentages in Exhibits 8a and 8b show how often the alternative with the highest score is the same with the two methods. 227 of 407 subjects (56%) have exactly the same rank for alternatives with all the methods. The similarity of the order of the alternatives within a subject was also measured by Spearman's rank correlation. The correlations are high between AHP, SMART, SWING, DIRECT and TRADEOFF compared to earlier studies (Exhibits 8a and 8b), especially when the AHP is used with the balanced scale (AHP2 and AHP4). These results partly support Hypothesis H1 but one should keep in mind that the order of the alternatives is not the same for all of the subjects.

Exhibit 7: *Averages of subjects' evaluations how well the overall scores reflected their own opinions (1 = ☺, 5 = ☹).*

Direct	2.43	(264)	AHP1	2.56	(48)
Smart	2.52	(251)	AHP3	2.67	(64)
Swing	2.54	(249)	AHP2	2.41	(69)
Tradeoff	2.76	(254)	AHP4	2.75	(60)

Number of subjects in parentheses

Exhibit 8a: *Average rank correlation for the alternatives. Percentages of the subjects whose alternative with the highest score is the same with both methods.*

	Smart		Swing		Tradeoff	
Direct	0.87	84%	0.89	86%	0.87	84%
Smart			0.85	81%	0.81	80%
Swing					0.85	81%

Number of subjects 407

Exhibit 8b: *Average rank correlation for the alternatives with different AHP versions. Percentages of the subjects whose alternative with the highest score is the same wiht both methods.*

	AHP1		AHP3		AHP2		AHP4	
	(108)		(93)		(113)		(92)	
Direct	0.82	82%	0.76	76%	0.89	88%	0.86	87%
Smart	0.79	78%	0.76	75%	0.89	88%	0.84	83%
Swing	0.79	76%	0.77	80%	0.88	85%	0.86	89%
Tradeoff	0.77	78%	0.79	80%	0.87	88%	0.80	82%

Number of subjects in parentheses

We also calculated the overall scores of the alternatives with SMARTER weights. They are based on the centroid method of Solymosi and Dombi (1986) so that the weight of an attribute ranked to be *i*th is

$$w_i = \frac{1}{n}\sum_{k=i}^{n}\frac{1}{k},$$

where *n* is the number of attributes. Olson and Dorai (1992) presented how these weights are used with the AHP. The first alternative is the same with SMARTER and DIRECT, SMART, SWING, and TRADEOFF for 78%, 76%, 77%, and 77% of subjects, respectively.The average correlations of the rank of alternatives between SMARTER and DIRECT, SMART, SWING, and TRADEOFF are 0.78, 0.77, 0.78, and 0.76, respectively. The results are similar between SMARTER and different versions of the AHP. Here correlations are lower than in the experiment of Srivastava et al. (1995) where rank correlations of alternatives were calculated between SMARTER, DIRECT, and SWING weighting.

4.2 Similarity of weights

There are no ideal measures to study whether two sets of weights are close to each other. The normalized attribute weights are used in earlier studies but they depend on the number of attributes normalized simultaneously (Pöyhönen and Hämäläinen, 1997). Normalized weights can be used only if the subjects have same attributes and if the number of attributes does not change across subjects. The correlation between the weights is one measure to study the convergent validity of methods within individual decision maker (Schoemaker and Waid, 1982; Borcherding et al., 1991). However, there may be a perfect linear dependency between weights also in the case when the weights are different and thus correlation alone is not enough to describe differences between weights. For example, the correlation between weight sets (0.03, 0.27, 0.70) and (0.31, 0.33, 0.36) is 1 although the weights are clearly different. In this study, we decided to use the responses from the subjects directly, i.e. the weight ratios. With the AHP and TRADEOFF the weight ratios given by the participants are used instead of the estimated weights to eliminate the effects of possible inconsistencies between the statements.

Each participant gave the weight ratio of the first and the second most important attribute (see Exhibit 9). DIRECT, SWING, and TRADEOFF yield similar weight ratios and one cannot reject Hypothesis H2 for these methods. The AHP and SMART differ from the other methods. With the AHP (all versions) and with

SMART the subjects gave higher weight ratios. Hypothesis H4 is supported as the averages are smaller with AHP2 and AHP4 that used the balanced scale compared to AHP1 and AHP3. The statistical analysis was done by within subjects ANOVA within each group of subjects using different versions of the AHP (The normality assumptions are not met in every case and parts of the analysis here and later were also repeated with non-parametric tests.)

Exhibit 9: *Average weight ratios of the first and the second most important attribute.*

	Subjects doing comparisons with			
	AHP1	AHP3	AHP2	AHP4
	(97)	(93)	(112)	(92)
AHP	3.76	3.93	2.17	2.44
Direct	1.45	1.31	1.33	1.62
SMART	1.56	1.65	1.46	1.83
SWING	1.45	1.23	1.35	1.43
Tradeoff	1.36	1.49	1.37	1.56

Number of subjects in parentheses

This analysis shows clear statistical convergence between DIRECT, SWING, and TRADEOFF although the weight ratios are not exactly the same. The analysis was repeated for the weight ratio between the first and the third most important attribute and for the weight ratio between the second and the thrid most important attribute (for those subjects who had more than two attributes). The results are the same in these cases: DIRECT, SWING, and TRADEOFF yield similar single weight ratios while AHP and SMART yield higher weight ratios. It was also verified that these single weight ratios do not depend on the number of attributes compared simultaneously.

4.3 Spread of weights

The spread of weights is measured by the maximum weight ratio (see Exhibit 10). This clearly depends on the number of attributes and thus the latter part of Hypothesis H3 is supported. Comparisons between the methods show that again DIRECT, SWING, and TRADEOFF yield similar maximum ratios while SMART and the AHP produce larger spread of weights. We are not able to repeat the result of Fischer (1995) that TRADEOFF yields higher spread of weights than SWING. There are differences in maximum weight ratios between the AHP versions. The maximum weight ratios are similar for AHP1 and AHP3, both using the 1-to-9 scale, and again for AHP2 and AHP4 with the balanced scale. This supports Hypothesis H4.

Exhibit 10*: Averages of maximum weight ratios.*

Number of attributes	Direct	SMART	SWING	Tradeoff		AHP1		AHP3		AHP2		AHP4	
n = 2	2.25	2.79	2.32	1.67	(25)	4.13	(8)	3.00	(2)	2.16	(8)	3.15	(7)
n = 3	3.08	5.45	2.75	2.82	(109)	6.04	(24)	6.19	(27)	3.60	(35)	3.51	(23)
n = 4	3.84	6.09	3.47	4.43	(112)	6.76	(25)	7.50	(30)	6.10	(28)	5.01	(29)
n = 5	3.93	6.19	3.54	4.78	(149)	7.08	(40)	7.38	(34)	5.61	(41)	6.46	(33)

number of subjects in the parentheses

Statistical analysis for Hypothesis 3 was done by ANOVA separately for each group of subjects using one AHP version. Within each such group the method was a within-subjects variable and the number of attributes a between-subjects variable. The AHP versions were compared with between subjects ANOVA for separate groups of subjects. Weight ratios over 20 were removed from the analysis as outliers. With the AHP it was not possible to give weight ratios over 9.

There are no differences between the versions of the AHP where attribute ranges are either explicitly presented or not. Thus Hypothesis H5 is not supported. In this experiment we are not able answer the question whether decision makers do indeed take into account the attribute ranges in assessing weights. It may be that the participants considered the attribute ranges with the AHP also when they were not shown because the ranges were initially asked and explicitly dealt with in connection with the other methods. However, it may also be the case that the subjects ignored the shown attribute ranges with all the methods.

4.4 Use of numbers to describe preferences

Our results show that the properties of the AHP weights strongly depend on the evaluation scale used. With the other methods the subjects seemed to use a fixed set of numbers in their comparisons as well. The dependency between the spread of weights and the number of attributes may originate from the tendency to use multiples of tens in the evaluations. In SMART 175 subjects (44%) used only numbers that were multiples of ten. In SWING this was even more common as 247 (62%) subjects used only even tens. In this way the spread of weights depends on the number of attributes. Decision maker starts, for example, with 100 and 80 when he compares two attributes and gives 100, 80, and 50 if there are three attributes. In this way the maximum weight ratio increases from 1.25 to 2. This also explains why the weight ratio between the first and the second most important attribute does not depend on the number of attributes. The tendency to use integer numbers also appears with the DIRECT weighting where 138 subjects (35%) used multiples of ten and 131 subjects (33%) used integers between 1 and 10 only. In DIRECT weighting the presentation order of the methods probably affects these results as the subjects were likely to use the same numbers that they were using with the methods preceeding DIRECT weighting. One possible explanation for these findings is that decision makers do not use numbers to describe the strength of their preferences. Instead, it can be that the numbers are used to describe the rank of the attributes only (Birnbaum, 1982).

Futher support for this behavior can be found from the AHP responses. The numbers used in the evaluation scale are weights ratios and thus the frequencies how often each verbal expression is used should change when the numbers are changed. However, each verbal statement is used equally often in groups having different numerical scales (see Exhibit 11). This suggests that the subjects used the presented evaluation scales without considering the meaning of the actual numbers.

Exhibit 11: *The use of verbal statements with both scales*

	Scale	
Verbal statement	1-to-9	Balanced
Equally important	7%	6%
-	7%	6%
Slightly more important	19%	14%
-	15%	11%
Strongly more important	22%	18%
-	8%	14%
Very strongly more important	13%	19%
-	3%	5%
Extremely more important	6%	7%

The ratio 9 is the upper limit in the AHP scales. However, it is most interesting that the subjects adopted a similar upper limit of 10 in their comparisons with the other methods as well. Both with SMART and SWING the range of the numbers used was from 10 to 100. Only 16 subjects (4%) used numbers over 100 in SMART and 9 subjects (2%) used numbers below 10 in SWING. Some subjects spread the numbers exactly over the range from 10 to 100 with SMART (70 subjects, 18%) and SWING (27 subjects, 7%). It is difficult to say whether this tendency to limit the maximum ratio to 10 is truly a behavioral phenomenon or a consequence of experimental design. The subjects defined the attributes themselves and they probably did a realistic selection where extremely high weight ratios would not occur.

4.5 Inconsistencies in the preference statements

With the AHP and TRADEOFF the subjects were asked to give more weight ratio statements than is actually needed to calculate the weights. The subjects were not, however, given any feedback on the consistency of their statements. We used a scale independent consistency measure, C.M. (see Salo and Hämäläinen, 1997; Salo, 1993) to study the inconsistencies of the statements. This index approximates the range of possible weights resulting from the preference statements. The value 0 refers to perfect consistency and 1 is the highest level of inconsistency meaning that weights can get any values between 0 and 1 due to the inconsistency of the preference statements.

With both methods the inconsistency between the statements increases as the number of attributes increases and thus Hypothesis H6 is supported (see Exhibit 12). The TRADEOFF statements are as inconsistent as AHP2 and AHP4 judgments where subjects used the balanced scale. All subjects who were perfectly consistent with TRADEOFF gave statements indicating that all the attributes are equally important (36 of 373). Some subjects (35 of 373) decreased the adjustable attribute to zero level in which case one of the hypothetical alternatives got an overall score of zero. For example, one states that an alternative having the *highest* possible salary and *worst* possible career opportunities is equally preferred to an alternative having salary and career opportunities *both on their worst* possible levels. Such a statement indicates that an attribute, in the example salary, has a zero weight. Borcherding et al. (1991) found the same phenomenon with the TRADEOFF statements.

There are also differences between the different AHP versions. Inconsistency is higher with AHP1 and AHP3 than with AHP2 and AHP4 indicating that the nine point integer scale leads to more inconsistent comparison matrices than the balanced scale. This supports Hypothesis H4. This result is also explained by the phenomenon that decision makers focus on the verbal statements and use them to describe their preferences while the 1-to-9 scale fails to capture the correct numerical counterparts for the words (Pöyhönen et al., 1997).

Exhibit 12*: Average inconsistency of the comparisons for the different methods (Perfect consistency = 0)*

Number of attributes	AHP1		AHP3		AHP2		AHP4		Tradeoff	
n = 3	0.17	(23)	0.21	(23)	0.19	(34)	0.10	(23)	0.15	(103)
n = 4	0.28	(25)	0.32	(26)	0.24	(27)	0.15	(28)	0.22	(108)
n = 5	0.39	(40)	0.34	(31)	0.33	(41)	0.22	(33)	0.29	(127)

Number of subjects in parentheses

In summary we feel that with every method redundant preference statements should be asked and thus get more reliable estimates for the weights. Human judgments are bound to errors and biases, not all of which are directly related to the elicitation procedure. Inconsistencies are found with every weighting method. Weighting techniques that allow decision makers to directly give imprecise preference statements may be one way to help in practical preference elicitation (Salo and Hämäläinen, 1992).

Exhibit 13: *Summary of the results.*

H1: All weighting methods give the same rank for the alternatives	Over half of the subjects have the same order for alternatives with all the methods and 69% get the same first alternative with all the methods.

H2: AHP, DIRECT, SMART, SWING and TRADEOFF yield same weights.	DIRECT, SWING, and TRADEOFF weights do not differ from each other. With SMART and AHP subjects gave higher weight ratios.
H3: The spread of attribute weights is same for all methods, but depends on the number of attributes.	DIRECT, SWING, and TRADEOFF do not differ from each other. SMART and AHP are different from the other methods. Yes, spread of weights depends on the number of attributes.
H4: The spread of the AHP weights and inconsistency between AHP comparisons depend on the evaluation scale used .	Yes. The 1-to-9 scale leads to higher spread of weights and increase the inconsistency between preference statements compared with the balanced scale.
H5: The explicit presentation of attribute ranges affects weights	Not supported. One cannot see differences depending on whether attribute ranges are presented or not.
H6: The inconsistency between statements depends on the number of attributes.	Yes. Inconsistency between statements increases as the number of attributes becomes higher.

5 Conclusion

Our main result is that weights differ because decision makers choose their responses from a limited set of numbers. This can happen with all the methods. Decision makers use explicitly defined evaluation scales like in the AHP or they implicitly create evaluation scales which consist of a limited set of numbers only. This happens easily with methods that start the weight elicitation with even numbers like 100 (SWING) or 10 (SMART). The consequences are that the spread of weights and the inconsistencies between the preference statements become dependent on the number of attributes present in the comparison. Future research should focus more on this type of response scale effect instead of trying to find the superior procedures.

There are no fundamental differences between the weighting methods because they have the same theoretical background. Based on our results we think that the practitioners can choose whatever method they like for weight elicitation. They should, however, recognize the response scale effects that lead to different weights when the weighting method is changed. More than one method should be used to check out possible inconsistencies and to increase the understanding on how the weights are interpreted. This topic is increasingly important nowadays as the softwares allow one to combine easily different methods. This means that practitioners do not need to use the methods in a puristic way and they are able to redefine the methods to be more suitable for a particular decision making situation. The strict boundaries between different methods are already passed history.

Appendix A: Use of the Internet in experimental research

As far as we know, this was the first time when the Internet was used to collect experimental data in the field of multicriteria decision making. The new technology gave us experiences that we would have missed in a classroom and thus we want to discuss some of these phenomena. We do not go into technical details because the technology has improved greatly during last years and our solutions are definitely already old-fashioned. Furthermore, we do not go into details how to design the interface for an experiment like this.

A new way to do experimental research

Our system recorded the responses on-line and updated the research statistics almost automatically after a new respondent left our pages. In this way we were able to monitor the data, change experimental design during the experiment, give feedback to teachers around the world who asked their students to participate (each teacher got next day a file containing the responses of his or her own students), and get feedback from researcher around the world. From a researcher point of view the last remark is the most important one. The other researchers are themselves able to validate the experimental setup, participate the experiment, and give feedback already during the early phases of the research. In the beginning of the experiment it was extremely helpful to do this kind of ”adaptive experimental design” instead of traditional pilot testing and to correct the mistakes in experimental setup before starting the main data collection. The comments that we got back during this phase were extremely valuable.

It was very interesting to follow how the statictics evolved. When the number of respondents was small, less than 50, almost each new response changed some of the conclusions in the statistical sense. The fact that small sample size does not give reliable results is of course well known, but yet it was very educating to follow this evolution with your own data. The collection of responses ”on-line” may also lead to a temptation to stop data collection when the results are what a researcher hopes for. In our experiment we continued the data collection until the results were very stable. After first 150 respondents the final results did not change any more.

Quality of responses

People are very impatient when they surf through web pages and we did not even try to get an average user of the Internet to visit our page. We decided to focus on a target group that is usually used in the experiments of this type, students, and to mail information of the experiment to the mailing lists that are read by the researchers in this area. However, if this experiment would start again we would use more the possibilities offered by the Internet and send information of the experiment to a larger population over the borders of disciplines.

Most of the student subjects completed the experiment alone and we are not able to control what they did during the experiment. On the average a respondent spent 30 minutes with the experiment. We do not know what proportion of that time was used to answer questions and we do not know what type of problems subjects had. We know that about 10% of all the people who started our experiment left the pages before the end. We got feedback that some parts of the experiment were difficult and the subjects wanted to get examples (especially with the TRADEOFF questions). This was an issue that we considered for a long time. It is clear that good instructions are needed when a subject uses the Internet and is not able to get immediate answer to his or her questions. On the other hand, there should not be very much text because people get easily tired when reading from the screen and the length of the experiment is very critical in order to get responses of good quality. Furthermore, examples affect the responses greatly. We did not want to get results that would look like our examples. We solved this problem by adding checkings that prompted back instructions and notes if the participants made mistakes. Nevertheless, one drawback of the use of the Internet is that it is difficult to give a participant sufficient instructions.

Technical problems

A lot of extra work was needed to set up the experiment compared to normal class room experiments. One aspect is the security. The files of the experiment were located at one of the central computers of the Helsinki University of Technology and the security regulations forced us to do a lot of additional programming work to prevent unwanted attempts to misuse the experiment. Variety of the browsers caused problems because we had to design the experiment to be browser independent. The problems of the computers in other universities and slow connections caused unexpected problems and loss of data. We estimated that due to technical problems 5% of data was lost. The amount of additional work that was needed to set up this experiment was about 6 months of programming (one person doing full time work). Although the technology has developed during last years and this time would now be shorter, this work is still an addition to the work needed to design the experiment and longer than what it takes to run classroom experiments.

Would we do it again?

There are pros and cons in using the Internet. The biggest drawback is that the connection to participants is lost and researchers do not know what their subjects are doing. This may affect the quality of data but we believe that at least in our experiment the results are as reliable as if they were collected in a class room with computers. The use of the Internet also requires extra time and resources to set up. On the other hand, the biggest advantage is that researchers are able to openly present their experimental setup and get comments during the early phases of their research. As a summary we conclude that yes, we would do this experiment again in the Internet. It was thrilling to follow the flow of data to you own computer from the different corners of the world.

References

Belton, V. and Gear, T., “On a Short-coming of Saaty’s Method of Analytic Hierarchies,“ *OMEGA*, 3 (1983), 228-230.

Belton, V., “A Comparison of the Analytic Hierarchy Process and a Simple Multi - Attribute Value Function,“ *European Journal of Operational Research*, 26 (1986), 7-21.

Birnbaum, M. H., “Problems with So-Called "Direct" Scaling,“ *Selected Sensory Methods: Problems and Approaches to Hedonics, ASTM STP 773*, J. T. Kuznicki, R. A. Johnson, and A. F. Rutkiewic, Eds., American Society for Testing and Materials, (1982), 34-48.

Borcherding, K., Eppel, T. and von Winterfeldt, D., “Comparison of Weighting Judgments in Multiattribute Utility Measurement,“ *Management Science*, 37 (1991), 1603-1619.

Budescu, D., Crouch, B., and Morera, O., "A Multi-Criteria Comparison of Response Scales and Scaling Methods in the AHP," *Proceedings of the Fourth International Symposium on the Analytic Hierarchy Process*, Simon Fraser University, Burnaby, Canada, (1996).

Edwards, W., “How to Use Multiattribute Utility Measurement for Social Decision Making,“ *IEEE Transactions on Systems, Man and Cybernetics*, SMC-7 (1977), 326-340.

Edwards, W. and Barron, F.H., “SMARTS and SMARTER: Improved Simple Methods for Multiattribute Utility Measurement“, *Organizational Behavior and Human Decision Processes*, 60 (1994), 306-325.

Fischer, G.W., "Range Sensitivity of Attribute Weights in Multiattribute Value Models," *Organizational Behavior and Human Decision Processes*, 62 (1995), 252-266.

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

Lootsma, F.A., "Scale Sensitivity in the Multiplicative AHP and SMART," *Journal of Multi-Criteria Decision Analysis*, 2 (1993), 87-110.

Olson, D.L. and Dorai, V.K., "Implementation of the Centroid Method of Solymosi and Dombi," *European Journal of Operational Research,* 60 (1992), 117-129.

Olson, D.L., Moshkovich, H.M, Schellenberg, R., and Mechitov, A.I., "Consistency and Accuracy in Decision Aids: Experiments with Four Multiattribute Systems," *Decision Sciences*, 26 (1996), 723-748.

Pöyhönen, M., Hämäläinen, R.P. and Salo, A.A., “An Experiment on the Numerical Modeling of Verbal Ratio Statements,“ *Journal of Multi-Criteria Decision Analysis,* 6 (1997), 1-10.

Pöyhönen, M. and Hämäläinen, R.P., “Notes on the Weighting Biases in Value Trees,“ forthcoming in *Journal of Behavioral Decision Making* (1997).

Saaty,T.L., *The Analytic Hierarchy Process*, McGraw-Hill, New York, (1980).

Saaty, T.L., “Highlights and Critical Points in the Theory and Application of the Analytic Hierarchy Process,“ *European Journal of Operational Research*, 74 (1994), 426-447.

Salo, A.A. and Hämäläinen, R.P., “Preference Assessment by Imprecise Ratio Statements,“ *Operations Research*, 40 (1992), 1053-1061.

Salo, A.A. and Hämäläinen, R.P., “On the Measurement of Preferences in the Analytic Hierarchy Process,“ (and comments by V. Belton, E. Choo, T. Donegan, T. Gear, T. Saaty, B. Schoner, A.Stam, M. Weber, B. Wedley) forthcoming in *Journal of Multi-Criteria Decision Analysis,* (1997).

Salo, A.A., “Inconsistency Analysis by Approximately Specified Priorities,“ *Mathematical and Computer Modelling*, 17 (1993), 123-133.

Schoemaker, P.J. and Waid, C.C., “An Experimental Comparison of Different Approaches to Determining Weights in Additive Value Models,“ *Management Science,* 28 (1982), 182-196.

Schoner, B. and Wedley, W.C., “Alternative Scales in AHP,“ in A.G.Lockett and G.Islei (Eds.), *Improving Decision in Organisations*, Lecture Notes in Economics and Mathematical Systems 335, Springer-Verlag, Berlin (1989), 345-354.

Schoner, B. and Wedley, W.C., and Choo, E.U., "A Unified Approach to AHP with Linking Pins," *European Journal of Operational Research*, 64 (1993), 384-392.

Solymosi, T. and Dombi, J., "A Method for Determining the Weights of Criteria: The Centralized Weights," *European Journal of Operational Research*, 26 (1986), 35-41.

Srivastava, J., Connolly, T., and Beach, L. R., "Do Ranks Suffice? A Comparison of Alternative Weighting Approaches in Value Elicitation," *Organizational Behavior and Human Decision Processes*, 63 (1995), 112-116.

Stillwell, W.G., von Winterfeldt, D. and John, R.S., “Comparing Hierarchical and Nonhierarchical Weighting Methods for Eliciting Multiattribute Value Models,“ *Management Science,* 33 (1987), 442-450.

Stewart, T.,”A Critical Survey on the Status of Multiple Criteria Decision Making Theory and Practice," *OMEGA*, 20 (1992), 569-586.

Weber, M. and Borcherding, K., “ Behavioral Influences on Weight Judgments in Multiattribute Decision Making,“ *European Journal of Operational Research*, 67 (1993), 1-12.

von Nitzsch, R. and Weber, M., “The Effect of Attribute Ranges on Weights in Multiattribute Utility Measurements,“ *Management Science*, 39 (1993), 937-943.

von Winterfeldt, D. and Edwards, W., *Decision Analysis and Behavioral Research*, Cambridge University press, Cambridge, MA, (1986).

Vrolijk, H. and Huizingh, E. "An Empirical Evaluation of the Scale Sensitivity in the AHP: An Assessment of Managerial Implications," *Proceedings of the Fourth International Symposium on the Analytic Hierarchy Process,* Simon Fraser University, Burnaby, Canada, 1996.