JOURNAL OF MULTI-CRITERIA DECISION ANALYSIS, VOL 6, 309–319 (1997)

On the Measurement of Preferences in the Analytic Hierarchy Process

AHTI A. SALO and RAIMO P. HÄMÄLÄINEN*
Systems Analysis Laboratory, Helsinki University of Technology, Otakaari 1 M, 02150 Espoo, Finland

ABSTRACT

In this paper we apply multiattribute value theory as a framework for examining the use of pairwise comparisons in the analytic hierarchy process (AHP). On one hand our analysis indicates that pairwise comparisons should be understood in terms of preference differences between pairs of alternatives. On the other hand it points out undesirable effects caused by the upper bound and the discretization of any given ratio scale. Both these observations apply equally well to the SMART procedure which also uses estimates of weight ratios. Furthermore, we demonstrate that the AHP can be modified so as to produce results similar to those of multiattribute value measurement; we also propose new balanced scales to improve the sensitivity of the AHP ratio scales. Finally we show that the so-called supermatrix technique does not eliminate the rank reversal phenomenon which can be attributed to the normalizations in the AHP. © 1997 John Wiley & Sons, Ltd.

J. Multi-Crit. Decis. Anal. **6**: 309–319 (1997)
No. of Figures: 4. No. of Tables: 2. No. of References: 40.

KEY WORDS: hierarchical weighting; analytic hierarchy process; multiattribute value measurement; ratio scales

1. INTRODUCTION

The AHP (Saaty, 1980) has been very successful in gaining the acceptance of practitioners, possibly owing to the helpfulness of the hierarchical problem representations and the appeal of pairwise comparisons in preference elicitation. At present, popular software products such as Expert Choice and HIPRE 3+ continue to promote the AHP as an easy method of multicriteria decision analysis (e.g. Buede, 1992). The range of reported practical applications is extensive (e.g. Vargas, 1990).

Despite the popularity of the AHP, many authors have expressed concern over certain issues in the AHP methodology. The possibility of rank reversals, for instance, has been regarded as unacceptable. Other much-debated problems include the meaning of pairwise comparisons, the relationship between scores and criteria weights, the properties of the 1–9 ratio scale and the prohibitive complexity of the supermatrix approach (Kamenetzky, 1982; Watson and Freeling, 1982; Belton and Gear, 1983; Belton, 1986; Dyer, 1990; Weber and Borcherding, 1993; Schenkerman, 1994).

This paper addresses each of the above points from the perspective of multiattribute value measurement and, in doing so, aims to provide a constructive summarizing analysis. In particular we demonstrate what restrictions multiattribute value representations place on the use of ratio statements in pairwise comparisons. We also show how the elicitation procedures in the AHP could be carried out so that the results are in accordance with multiattribute value measurement. Thus our results can be seen as a step towards the reconciliation of the two methodologies and the resolution of issues raised earlier in the literature.

At present, alternative methods are often portrayed as rivalling approaches and the emphasis tends to be placed on the differences rather than on the similarities. Against this background there is a continuing need for comparative research which seeks to clarify interrelationships between alternative methods, thus helping practitioners in the choice of well-suited approaches to the problems they are facing. Indeed, we believe that comparative research, together with the possible convergence of methodologies, contributes significantly to the important goal of improving the practice of decision analysis.

*Corresponding author.

CCC 1057–9214/97/060309–11$17.50

Received 20 November 1994
Accepted 21 February 1997

2. ASPECTS OF MEASUREMENT

Multiattribute value measurement is based on the primitive notion of an underlying preference relation. If this relation satisfies a number of conditions, formalized as axioms, then it has a functional representation which attaches a real number to each consequence. The construction of this value representation provides decision support by allowing preference statements to be inferred even for those consequences about which no explicit judgements have been made. The value representation also ensures that the inferred statements comply with the underlying axioms which are regarded as maxims of rational choice behaviour.

More specifically, let $\succeq$ denote the decision maker's (DM's) preference relation defined on the set of n-attribute consequences $\mathbf{x} = (x_1, \ldots, x_n) \in X$. For practical purposes, consequences can be thought of as real or hypothetical alternatives, i.e. the consequence space contains all the relevant alternatives. If the relation $\succeq$ has the required properties (e.g. transitivity, completeness, mutual preferential independence; see Krantz *et al.* (1971) and Keeney and Raiffa (1976)), then there exists a real-valued numerical representation, i.e. a value function on $v(x)$ such that

$$\begin{aligned} &\mathbf{x} \succeq \mathbf{y} \iff \\ &v(\mathbf{x}) = \sum_{i=1}^{n} v_i(x_i) \geqslant \sum_{i=1}^{n} v_i(y_i) = v(\mathbf{y}) \end{aligned} \tag{1}$$

That is, consequence $\mathbf{x}$ is preferred to consequence $\mathbf{y}$ if and only if $v(\mathbf{x})$, the value attached to $\mathbf{x}$, is greater than $v(\mathbf{y})$, the value attached to $\mathbf{y}$. In (1) the value representation is unique up to positive affine transformations; that is, if $v(\mathbf{x}) = \Sigma_{i=1}^{n} v_i(x_i)$ is a model of the DM's preference relation, then for any $\alpha > 0$ and real β the function $v'(\mathbf{x}) = \alpha v(\mathbf{x}) + \beta$ is an equivalent representation of the relation $\succeq$.

Dyer and Sarin (1979) specify conditions for the DM's preference relation $\succeq^*$, defined on differences (or exchanges) in consequence pairs of $X^* = \{\mathbf{xy} | \mathbf{x} \succeq \mathbf{y}\}$, such that the additive representation in (1) has the property

$$\begin{aligned} &\mathbf{x}^1\mathbf{x}^2 \succeq^* \mathbf{y}^1\mathbf{y}^2 \iff \\ &v(\mathbf{x}^1) - v(\mathbf{x}^2) \geqslant v(\mathbf{y}^1) - v(\mathbf{y}^2) \end{aligned}$$

In other words, value difference $v(\mathbf{x}^1) - v(\mathbf{x}^2)$ is greater than value difference $v(\mathbf{y}^1) - v(\mathbf{y}^2)$ if and only if the improvement from alternative $\mathbf{x}^2$ to $\mathbf{x}^1$ is judged to be greater than the improvement from alternative $\mathbf{y}^2$ to $\mathbf{y}^1$. This property gives these *measurable value functions* a strength-of-preference interpretation which allows them to be elicited from comparisons of value improvements in pairs of actual or hypothetical alternatives (e.g. Farquhar and Keller, 1989).

Once a suitable range of achievement levels $[x_i^\circ, x_i^*]$ has been defined for each attribute, it is customary to normalize the value function representation so that the values $v(\mathbf{x}^\circ) = v(x_1^\circ, \ldots, x_n^\circ) = 0$ and $v(\mathbf{x}^*) = v(x_1^*, \ldots, x_n^*) = 1$ are assigned to the worst and best conceivable consequences respectively. This range needs to be wide enough to cover the alternatives' achievement levels (or 'performances') on the ith attribute. By normalizing the component value functions onto the [0,1] range, the additive representation can be written as

$$\begin{aligned} v(\mathbf{x}) &= \sum_{i=1}^{n} v_i(x_i) \\ &= \sum_{i=1}^{n} [v_i(x_i) - v_i(x_i^\circ)] \\ &= \sum_{i=1}^{n} [v_i(x_i^*) - v_i(x_i^\circ)] \frac{v_i(x_i) - v_i(x_i^\circ)}{v_i(x_i^*) - v_i(x_i^\circ)} \\ &= \sum_{i=1}^{n} w_i s_i(x_i) \end{aligned} \tag{2}$$

where $s_i(x_i) = [v_i(x_i) - v_i(x_i^\circ)]/[v_i(x_i^*) - v_i(x_i^\circ)] \in [0, 1]$ is the normalized score of $\mathbf{x}$ on the ith attribute and $w_i = v_i(x_i^*) - v_i(x_i^\circ)$ is the weight of the ith attribute.

Formally, the additive expression (2) resembles the way in which the AHP uses the equation

$$w(x) = \sum_{i=1}^{n} w_i w_i(x) \tag{3}$$

for preference aggregation; here $w(x)$ and w_i stand for the overall weights (or priorities) of alternative x and the ith criterion respectively and $w_i(x)$ is the x-component of the local priority vector at the ith criterion. In view of the apparent similarity of (2) and (3), it is appropriate to ask under what conditions the results of the two approaches coincide.

3. PREFERENCE ELICITATION IN THE AHP

Pairwise comparisons

The questions that the AHP uses to elicit preference information about the alternatives are typically of the form 'Which of the alternatives, Mercedes or Honda, is better with respect to quality and by how much?'. However, Watson and Freeling (1982), Belton (1986) and Dyer (1990), among others, have argued that such value comparisons do not constitute an acceptable procedure of preference elicitation.

In view of the properties of the value representation (1), this is indeed the case, for if $v_i(\cdot)$ is a component value function, then $v_i'(\cdot) = v_i(\cdot) + \beta$ is an equivalent representation of preferences, and yet, if the achievement levels x_i and y_i are not equally preferred, the ratio $[v_i(x_i) + \beta]/[v_i(y_i) + \beta]$ assumes different values depending on the choice of β. That is, the result of the comparison depends on a parameter (β) whose value may be selected arbitrarily.

In contrast, positive value differences can be legitimately measured on a ratio scale because the expression

$$\frac{v_i'(x_i^1) - v_i'(x_i^2)}{v_i'(y_i^1) - v_i'(y_i^2)} = \frac{[\alpha v_i(x_i^1) + \beta] - [\alpha v_i(x_i^2) + \beta]}{[\alpha v_i(y_i^1) + \beta] - [\alpha v_i(y_i^2) + \beta]} = \frac{v_i(x_i^1) - v_i(x_i^2)}{v_i(y_i^1) - v_i(y_i^2)}$$

remains constant for all admissible choices of parameters α and β. Together with (2), this suggests a preference elicitation procedure where the DM is asked to make ratio statements about value differences $v_i(x_i^k) - v_i(x_i^\circ)$ defined by the alternatives' achievement levels x_i^k and criterion-specific reference points x_i°. On each criterion these reference points must be less preferred than the actual alternatives' achievement levels. In other words, the DM should be asked questions such as 'Which of the alternatives, Mercedes or Honda, gives the greater quality improvement over BadQualityCar (i.e. the poor-quality reference car)?' and, assuming that the reply is Mercedes, 'How many times greater is the quality improvement from BadQualityCar to Mercedes than the quality improvement from BadQualityCar to Honda?' (see Watson and Freeling (1982) for an analogous phrasing).

The need for reference points can be easily understood, for example, in the case of measuring distances between cities (see the example in Saaty (1980)). We need to designate one city (i.e. Philadelphia) as a reference point from which the distances are measured. It would be pointless to ask questions such as 'Which city is farther away, London or San Francisco?'. In the context of intangibles it is equally imperative to provide reference points relative to which the pairwise comparisons are performed.

The AHP has two measurement modes, absolute and relative, which differ in that in relative measurement the alternatives' local priorities are normalized so that they add up to one, whereas in absolute measurement no such normalization is applied to the alternatives (Saaty, 1994).

With regard to the different measurement modes, a further advantage of the value difference interpretation is that they need not be treated as inherently different types of information. For example, it has been suggested that the comparison matrix should be built from ratios of absolute measurements whenever such measurements are available (Schoner *et al.*, 1993; Schenkerman, 1994). However, this approach cannot always be recommended because it presumes that (i) the absolute measurements and (ii) the subjective values associated with these measurements are linearly related to each other, even though in reality the relationship may well be non-linear. For example, a 2 week holiday need not be twice as attractive as a 1 week holiday in terms of its duration. Rather than looking at ratios of durations, one should compare the benefit, or value increment, of having a 2 week holiday (as opposed to none) with that obtained from a 1 week holiday (as opposed to none).

Criteria weights

In their discussion of criteria weights, Belton and Gear (1983), Saaty *et al.* (1983), Schoner and Wedley (1989a) and Schoner *et al.* (1993) assert that in the AHP the criteria weights should be proportional to the average (or total) contribution of the alternatives on the respective criteria. In the value difference framework the correctness of this assertion can be demonstrated as follows. Given that the pairwise comparisons among the alternatives $\mathbf{x}^1, \ldots, \mathbf{x}^m$ are correctly linked to the value differences $v_i(x_i^j) - v_i(x_i^\circ)$, $j = 1, \ldots, m$, then the value representation (2) can be rewritten as

$$v(\mathbf{x}^k) = \sum_{i=1}^{n} [v_i(x_i^k) - v_i(x_i^\circ)] = \sum_{i=1}^{n} \sum_{j=1}^{m} [v_i(x_i^j) - v_i(x_i^\circ)] \frac{v_i(x_i^k) - v_i(x_i^\circ)}{\sum_{j=1}^{m} [v_i(x_i^j) - v_i(x_i^\circ)]} \tag{4}$$

A comparison with (3) now shows that w_i, the weight of the ith criterion, should indeed be proportional to the value difference $\Sigma_{j=1}^{m} [v_i(x_i^j) - v_i(x_i^\circ)]$. This logical requirement should be followed in other hierarchical weighting procedures as well (see Edwards and Barron (1994) for related comment about SMART), yet it can be easily forgotten in cases of practical decision support, as the following example demonstrates.

A company is operating in countries C_1 and C_2. It has two investment plans A_1 and A_2 which produce revenues as follows:

	C_1	C_2	Total	
A_1	3	4	7	
A_2	1	8	9	= best
Total	4	12	16	

The manager decides to use a simple weighting model (Figure 1) to compare the plans. In this model the overall revenue is divided by the country criteria. The easy mistake in weighting is to assign equal weights (0.5) to the countries, based on the thinking that profits are equally important irrespective of their origin. This leads to the choice of plan A_1 although the total profits from plan A_2 are higher:

$$w(A_1) = \tfrac{1}{2} \times \tfrac{3}{4} + \tfrac{1}{2} \times \tfrac{4}{12} = \tfrac{13}{24} = \text{best}$$
$$w(A_2) = \tfrac{1}{2} \times \tfrac{1}{4} + \tfrac{1}{2} \times \tfrac{8}{12} = \tfrac{11}{24}$$

The correct weighting of the country criteria should take into account the relative amount of revenues coming from the respective countries (criteria). This leads to the criteria weights $w_1 = \frac{4}{16} = \frac{1}{4}$ and $w_2 = \frac{12}{16} = \frac{3}{4}$, which in turn produce the correct overall weights:

$$w(A_1) = \tfrac{1}{4} \times \tfrac{3}{4} + \tfrac{3}{4} \times \tfrac{4}{12} = \tfrac{7}{16}$$
$$w(A_2) = \tfrac{1}{4} \times \tfrac{1}{4} + \tfrac{3}{4} \times \tfrac{8}{12} = \tfrac{9}{24} = \text{best}$$

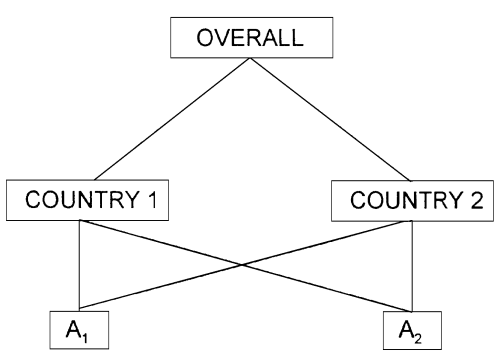


Figure 1. Prioritizing two investment plans A_1 and A_2 with revenues from two different countries

Local priorities can be normalized in other ways as well provided that the interpretation of criteria weights is modified accordingly (Belton and Gear, 1983; Schoner and Wedley, 1989a; Schoner *et al.*, 1993). For instance, if under each criterion the local priority of the kth alternative is set to one, then (2) becomes

$$\sum_{i=1}^{n} v_i(x_i) - v_i(x_i^\circ) = \sum_{i=1}^{n} [v_i(x_i^k) - v_i(x_i^\circ)] \frac{v_i(x_i) - v_i(x_i^\circ)}{v_i(x_i^k) - v_i(x_i^\circ)}$$

which together with (3) shows that the corresponding criteria weights should now be proportional to the value differences $v_i(x_i^k) - v_i(x_i^\circ)$.

Once the local priorities $w_i(x)$ in (3) are available, questions for the assessment of criteria weights can be phrased in terms of any of the alternatives. For example, if the DM estimates that the value difference from the reference point x_1° to the kth alternative under the first criterion is r times larger than the difference from x_2° to the lth alternative under the second criterion, then the ratio of criteria weights can be written as

$$\frac{w_1}{w_2} = \frac{\sum_{j=1}^{m} [v_1(x_1^j) - v_1(x_1^\circ)]}{\sum_{j=1}^{m} [v_2(x_2^j) - v_2(x_2^\circ)]}$$

$$= \frac{[v_1(x_1^k) - v_1(x_1^\circ)]\left\{1 + \sum_{j \neq k} \frac{[v_1(x_1^j) - v_1(x_1^\circ)]}{[v_1(x_1^k) - v_1(x_1^\circ)]}\right\}}{[v_2(x_2^l) - v_2(x_2^\circ)]\left\{1 + \sum_{j \neq l} \frac{[v_2(x_2^j) - v_2(x_2^\circ)]}{[v_2(x_2^l) - v_2(x_2^\circ)]}\right\}}$$

$$= r\frac{1 + \sum_{j \neq k} w_1(j)/w_1(k)}{1 + \sum_{j \neq l} w_2(j)/w_2(l)}$$

The resulting alternative ways of estimating criteria weights can be used to check the consistency of the DM's responses.

On the higher levels of tree-shaped hierarchies the normalized value representation (4) implies that the weight of a criterion must be proportional to the average value improvement that the alternatives bring over the reference points in the entire subhierarchy of which the criterion is the topmost element. In more general AHP hierarchies, which do not have a tree structure, the criteria weights do not have this clear-cut interpretation because they are not uniquely related to the weights of the lowest-level criteria.

With the above interpretation of ratio comparisons the AHP can be expected to give results similar to those of multiattribute value measurement. However, in practice the comparison of AHP against other preference elicitation methodologies is problematic because different procedures of multiattribute value assessment produce divergent results; for instance, the observed violations of procedural invariance imply that there is no single benchmark for comparisons (Weber and Borcherding, 1993). Consequently, future empirical research should compare the performance of the conventional AHP and its variants with a number of multiattribute evaluation techniques. Among other things, such research should investigate how decision makers understand elicitation questions and to what extent ratios of value differences give results that are insensitive to the choice of reference points and in keeping with other preference assessment methods.

4. DISCRETIZATION OF RATIO SCALES

In their empirical comparison of ratio scales, Schoner and Wedley (1989b) requested subjects to judge the proportion of different colours in displays of varying fuzziness using Stevens' magnitude estimation and the 1–9 scale of the AHP. They found that the AHP produced consistently less accurate estimates, especially when the colours covered roughly equal proportions of the display.

As illustrated in Figure 2, a plausible explanation for this finding lies in the uneven dispersion of the local priorities $w = (w_1, w_2)$ that are supported by the ratios $r = \frac{1}{9}, \frac{1}{8}, \ldots, \frac{1}{2}, 1, 2, \ldots, 8, 9$ through the mapping $w_1 = 1/(r+1)$, $w_2 = r/(r+1)$ (see Figure 2). The effect of replacing the ratio 1 by 2, for instance, is 15 times greater than the local priority difference between the ratios 8 and 9. Similar results hold for a larger number of alternatives, although the procedures for resolving inconsistencies in a complete set of pairwise comparisons may moderate the effect.

For a given set of priority vectors the corresponding ratios can be computed from the inverse relationship $r = w/(1 - w)$. In particular, by choosing priority vectors which are equally far apart from each other, we can define so-called *balanced scales*: for instance, the priorities 0.1, 0.15, 0.2, ..., 0.80, 0.85, 0.9, for example, lead to the scale 1, 1.22, 1.50, 1.86, 2.33, 3.00, 4.00, 5.67, 9.00, whereas the scale 1, 1.27, 1.62, 2.09, 2.78,

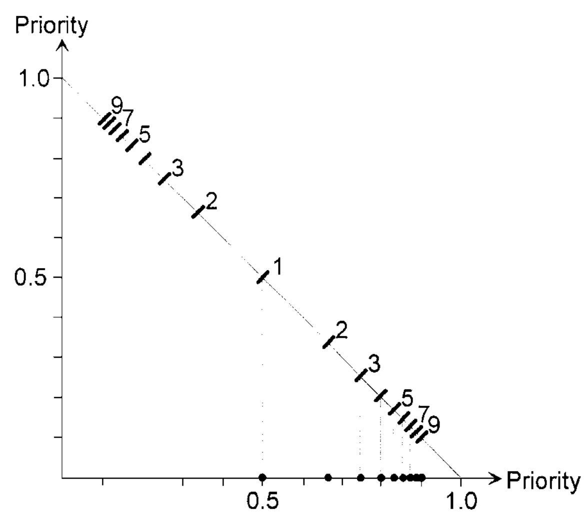
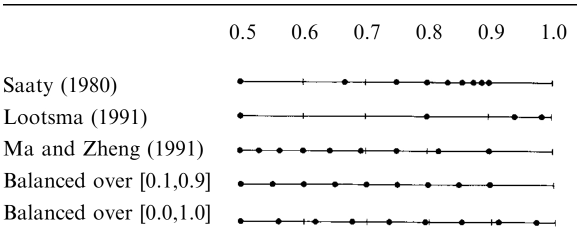


Figure 2. Dispersion of local priorities as a function of ratio estimate

Table I. Priority distributions in discretized ratio scales



3.86, 5.80, 10.3, 33.3 is obtained by covering the unit range [0,1] with 17 priority vectors such that each represents priorities in segments of equal length (see Table I). The 9/9, 9/8, ..., 9/2, 9/1 scale of Ma and Zheng (1991) also gives more uniformly distributed priorities than the 1–9 scale, while the priorities in Lootsma's (1991) geometrically progressing 1, 4, 4^2, ... scale are sparse over most of the priority range.

A shortcoming of the 1–9 scale (or, for that matter, of any scale with a finite upper bound M) is that the upper bound restricts the range of local priority vectors. Specifically, if the $n \times n$ comparison matrix is consistent (so that its largest eigenvalue is n), then the maximum for the first component is achieved at

$$w_1 + \sum_{i=2}^{n} Mw_i = nw_1$$
$$\iff \quad w_1 + M(1 - w_1) = nw_1$$
$$\iff \quad w_1(n + M - 1) = M$$

i.e. no component of a local priority vector can exceed the bound $w_{\max} = M/(n + M - 1)$. The corresponding lower bound $w_{\min} = 1/[M(n-1) + 1]$ follows from the equality

$$w_1 + \sum_{i=2}^{n} \frac{1}{M} w_i = nw_1$$
$$\iff \quad Mw_1 + (1 - w_1) = Mnw_1$$
$$\iff \quad w_1[M(n-1) + 1] = 1$$

Thus the existence of a finite upper bound imposes the fundamental restrictions

$$w_{\max} = \frac{M}{n + M - 1} \tag{5}$$

$$w_{\min} = \frac{1}{M(n-1) + 1} \tag{6}$$

which are illustrated in Figure 3.

It is worth noting that the above restrictions apply to any ratio procedure in which the DM

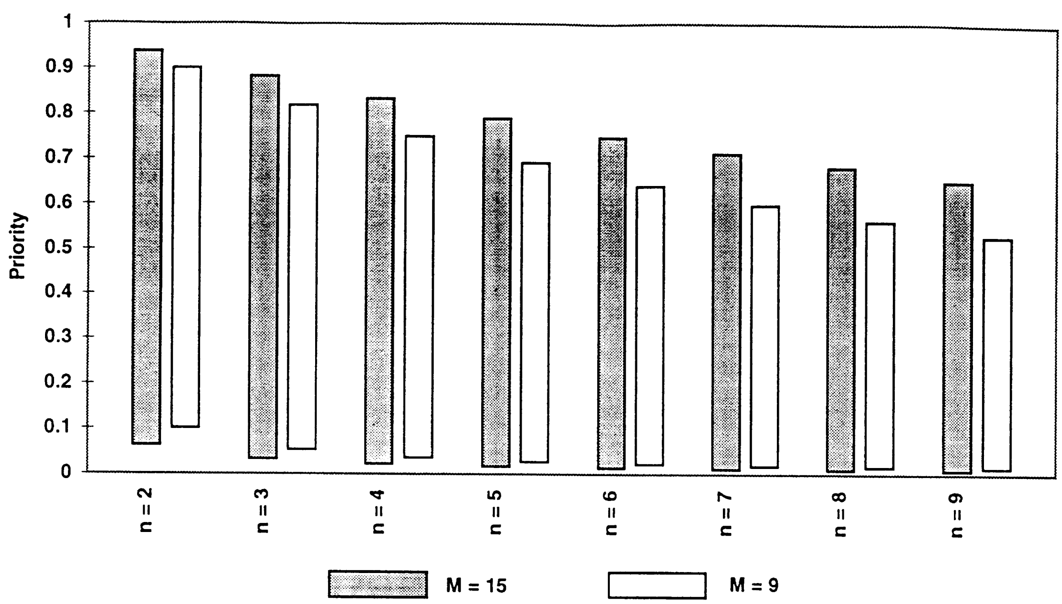


Figure 3. Range of possible priority as a function of maximum ratio (M) and number of subcriteria (n)

(perhaps unknowingly) places an upper bound for the submitted ratios. For example, if the DM uses 10 as the lowest reference score in SMART and, at the same time, implicitly assumes that 100 is the highest possible, he/she effectively puts an upper bound of 10 on the ratio comparisons. In this case the restrictions (5) and (6) hold with M = 10.

To some extent the above restrictions can be relaxed by introducing additional criterion levels so that each criterion has fewer subcriteria. For instance, a single criterion with eight subcriteria may be divided into two intermediate subcriteria ($n_1 = 2$) which in turn have four subcriteria ($n_2 = 4$) each (see Figure 4). In the modified hierarchy the maximum priority for the initial subcriteria, $[M/(n_1 + M - 1)][M/(n_2 + M - 1)]$, is larger than the corresponding upper bound in the single-level hierarchy, i.e. $M/(n_1 n_2 + M - 1)$.

Table II shows how the maximum overall priorities become higher when $n = n_1 n_2$ subcriteria are divided under n_1 intermediate criteria with n_2 subcriteria for each. Exchanging the roles of n_1 and n_2 has no effect on the upper bounds; for instance, the eight subcriteria in Figure 4 could be structured under four (instead of 2) intermediate subcriteria and the upper bound would still remain the same. The behaviour of these theoretical bounds parallels empirical findings according to which larger priority ratios are observed in steeper hierarchies (e.g. Stillwell *et al.*, 1987). However, the influence of the hierarchical structure on the results of weighting procedures still calls for further research (Pöyhönen and Hämäläinen, 1996).

The consistency ratio (CR) in the AHP is computed as the ratio between the consistency index $(\lambda_{\max} - n)/(n-1)$ of the given comparison matrix (with dimension n and principal eigenvalue $\lambda_{\max}$) and the average of similar indices computed for randomly generated reciprocal comparison matrices. Thus by construction the consistency ratio is a meaningful measure only on condition that the same scale has been employed both in the assessment of the actual comparison matrix and in the generation of the random matrices, but this in turn suggests that scale-invariant measures should be preferred to those that depend on a particular discretization of the ratio scale or on a given probability distribution on the points included in the discretization.

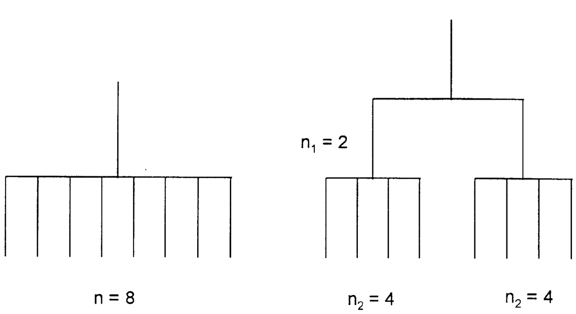


Figure 4. Conversion of a one-level hierarchy into a hierarchy with two intermediate subcriteria

Table II. Effect of decomposition on maximum local priorities for $M = 9$

n	4	6	8	10
Maximum priority	0.75	0.64	0.56	0.50
$n_2(n_1 = 2)$	2	3	4	5
Maximum priority	0.81	0.74	0.68	0.62

One way of deriving scale-invariant consistency measures is to transform the inconsistent replies into a non-empty set of feasible priorities, whereafter the properties of the resulting set can be used for measuring the inconsistency of the original comparison matrix. Specifically, if $a(i, j)$ denotes the element in the ith row and jth column of the comparison matrix, then the corresponding *extended bound* can be defined as $\bar{r}(i, j) = \max_k a(i, k)a(k, j)$ (Salo, 1993). Using these extended bounds, a scale-invariant consistency measure (CM) is obtained through the equation

$$\mathrm{CM} = \frac{2}{n(n-1)} \sum_{i>j} \frac{\bar{r}(i,j) - \underline{r}(i,j)}{[1 + \bar{r}(i,j)][1 + \underline{r}(i,j)]} \qquad (7)$$

where $\underline{r}(i, j)$ stands for the inverse of $\bar{r}(j, i)$. The above measure is in essence an indicator of the size of the *extended region*, i.e. the set of local priorities such that $w_i \leqslant \bar{r}(i, j)w_j$ for all $i, j \in \{1, \ldots, n\}$ (Salo, 1993), which in turn grows as the entries of the comparison matrix become more inconsistent.

The consistency measure (7) is incorporated in the HIPRE 3+ software, which, among other things, allows its users to employ different scales in hierarchical preference elicitation (Hämäläinen and Lauri, 1995).

There are situations in which the 1–9 scale seems to work well, such as in estimating distances from Philadelphia to several other cities in the example of Harker and Vargas (1987). However, the choice of another set of cities could well have favoured some other scale. The ratios to be included in a discretized scale should be selected with regard to entire sets of objects about which ratio comparisons may be made. In a recent experiment we discovered interpersonal differences in the interpretation of verbal statements, which complicates the use of such statements in preference elicitation. In addition, we found that the balanced scale in Table I outperforms the 1–9 scale in capturing the subjects' understanding of verbal expressions. All in all, there is a need for a better understanding of what people really understand by expressions such as 'weakly more important' or 'very strongly more important', especially since verbal statements sometimes produce greater inconsistencies without improving the accuracy of the results (Lund and Palmer, 1986). Related results have been reported in the literature on interpersonal differences in verbal probability assessment (Beyth-Marom, 1982) and in the extensive psychological research on scale problems (Birnbaum, 1978).

Apart from criterion-specific concerns, one also needs to examine to what extent the interpretation of verbal statements depends on the problem context, for if the verbal expressions are understood differently from one criterion to the next, then their automated use in preference elicitation becomes problematic indeed. Here new and interesting approaches are provided by methods which allow ambiguity or incompleteness in preference statements. In these approaches, verbal statements can be mapped into real-valued intervals in order to cover a range of interpretations and preferences (Hämäläinen *et al.*, 1991; Salo and Hämäläinen, 1992a, 1995; Hämäläinen and Pöyhönen, 1996).

5. THE SUPERMATRIX APPROACH

The supermatrix technique (Saaty, 1980, Chap. 8) has been suggested as a remedy to the rank reversals in the AHP (Harker and Vargas, 1987). Unfortunately, there are no clear-cut rules for determining when supermatrices should be used (Dyer, 1990); moreover, the supermatrix technique requires that the DM answers a much larger number of questions. These questions may also be quite complex, e.g. 'Given an alternative and a criterion, which of the two alternatives influences the given criterion more and how much more than another alternative?' (Saaty and Takizawa, 1986). In fact, the complexity of questions such as this may in part explain the scarcity of reported supermatrix applications. Hämäläinen and Seppäläinen (1986) were the first to call the method the analytic network process (ANP) and were indeed able to apply it in a policy problem with interrelationships between two planning horizons.

Despite claims to the contrary, the supermatrix technique does not eliminate rank reversals. For instance, assume that the DM makes the following statements about a new, third alternative in the example of Harker and Vargas (1987, p. 1395):

- w.r.t. C_1, A_3 is two times better than A_1;
- w.r.t. C_2, A_1 is four times better than A_3;
- w.r.t. C_3, A_1 and A_3 are equally preferred;
- w.r.t. A_3, C_2 is seven times more important than C_3;
- w.r.t. A_3, C_1 is two times more important than C_3.

After these statements the supermatrix takes the form

$$W = \begin{array}{c} \\ C_1 \\ C_2 \\ C_3 \\ A_1 \\ A_2 \\ A_3 \end{array} \begin{array}{c} \begin{array}{cccccc} C_1 & C_2 & C_3 & A_1 & A_2 & A_3 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0.3 & 0.3 & 0.2 \\ 0 & 0 & 0 & 0.3 & 0.5 & 0.7 \\ 0 & 0 & 0 & 0.4 & 0.2 & 0.1 \\ 0.25 & 0.364 & 0.444 & 0 & 0 & 0 \\ 0.25 & 0.545 & 0.111 & 0 & 0 & 0 \\ 0.5 & 0.091 & 0.444 & 0 & 0 & 0 \end{pmatrix} \end{array} \tag{8}$$

and straightforward computations now show that these statements change the feedback weights of the alternatives from (0.554, 0.446) to (0.352, 0.360, 0.287); in other words, the introduction of A_3 reverses the ranks of alternatives A_1 and A_2. There is also experimental evidence which suggests that reversals of this kind are too frequent to be classified as artefacts of numerical manipulation (Salo and Hämäläinen, 1992b).

The feedback technique suffers from rank reversals because it fails to ensure that the entries of the supermatrix are correctly linked to value differences between the alternatives. These links can be created by requiring, as in (2), that the local priority of the ith alternative on the kth criterion is related to the value differences by $w_k(i) = [v_k(x_k^i) - v_k(x_k^\circ)]/\Sigma_{j=1}^m [v_k(x_k^j) - v_k(x_k^\circ)]$. At the same time, feedback priorities need to be selected so that correct criteria weights are obtained when they are combined with the alternatives' weights; thus the relative importance of the kth criterion at the ith alternative, $w_i(k)$, should be such that the expression $w_i(k)\Sigma_{j=1}^{m}[v_j(x_j^i)-v_j(x_j^0)]$ is proportional to the value difference $v_k(x_k^i)-v_k(x_k^0)$ or, more specifically,

$$\frac{w_i(k)}{w_i(l)}=\frac{v_k(x_k^i)-v_k(x_k^\circ)}{v_l(x_l^i)-v_l(x_l^\circ)}=\frac{w_k w_k(i)}{w_l w_l(i)} \tag{9}$$

If these compatibility restrictions on the criteria weights and local priorities do not hold, the DM's judgements are not in agreement with an underlying value representation.

A direct implication of the above remarks is that the feedback questions, when correctly posed, are essentially similar to those employed in the assessment of criteria weights; for instance, the question 'Given an alternative (e.g. Mercedes), which criterion (e.g. style or quality) is more important in the overall choice of best car?' (Harker and Vargas, 1987) should in fact be rephrased as 'Which is preferred, the style improvement from BadCostCar to Mercedes or the quality improvement from BadQualityCar to Mercedes?'. Another reason why complete supermatrices seem excessively laborious is that the feedback replies at any one alternative provide estimates for the criteria weights (Schoner and Wedley, 1989a).

6. RANK REVERSALS

It has been argued that the relative measurement mode in the AHP is superior to the methods of value and utility theory because it incorporates preference reversals which are sometimes exhibited by unaided decision making (e.g. Saaty, 1994; Vargas, 1994). From a normative perspective, however, this line of reasoning is debatable: because one of the main objectives of the analysis is to contribute to decision quality, it is not clear why the results of the analysis should retain the questionable features of unaided human choice behaviour.

Instead, one should seek to determine whether the decision maker wishes to receive guidance in which preferences for pairs of alternatives are independent of what other alternatives are contained in the choice set. If this form of independence exists (which means that pairs of alternatives may be evaluated in isolation from the rest of the choice set), the relative measurement mode of the AHP is unsuitable because rank reversals must not occur. In the absence of such independence the preference relation as such is an elaborate construct which to our knowledge has not been characterized in terms of axioms. Even if a satisfactory axiomatization were available, the assessment of the preference representation would be a complex task because comparisons between pairs of alternatives would have to involve other alternatives as well. In contrast, the AHP elicits preferences without reference to other alternatives, wherefore the occasional rank reversals may be side-effects of the normalization procedure rather than credible results of the modelling procedure.

From the descriptive point of view the fact that rank reversals occur in the AHP does not imply that the AHP is good at predicting preference reversals. The appraisal of this type of predictive ability calls for empirical experiments with carefully selected control groups. Also, on condition that the AHP is found a successful predictor of preference reversals and unaided decision making at large, one may question what the aims of an AHP analysis are if unsupported judgements would produce essentially similar outcomes.

In the relative measurement mode of the AHP, any alternative that is not dominated by some convex combination of other alternatives may become the best one as a result of the introduction or deletion of other alternatives. For example, if the jth alternative is non-dominated in the above sense, there exists a set of constants, c_i, $i=1,\ldots,n$, such that the inequalities

$$\sum_{i=1}^{n} c_i w_i(j) \geqslant \sum_{i=1}^{n} c_i w_i(k)$$

hold for all $k \neq j$. Now, if the priorities for the new $(n+1)$th alternative are chosen so that the term $1/[1+w_i(n+1)]$ is proportional to c_i/w_i (where w_i, $i=1,\ldots,n$, denote initial criterion weights), the revised weight of the jth alternative becomes

$$\sum_{i=1}^{n} w_i \frac{w_i(j)}{1+w_i(n+1)} \propto \sum_{i=1}^{n} w_i \frac{c_i}{w_i} w_i(j) = \sum_{i=1}^{n} c_i w_i(j)$$

Since the same proportionality relation holds for the other alternatives as well, it follows that the introduction of the $(n+1)$th alternative causes the jth alternative to become the best one.

Thus the fundamental mathematical reason for the occurrence of rank reversals in the relative measurement mode is that the local priorities at the lowest level of the hierarchy are normalized so that they add up to one. When new alternatives are added or deleted, the local priorities associated with the other alternatives inevitably change and as a result the final ranking of the alternatives may also change. Any practitioner using normalized priorities should be aware of this possibility.

We feel that the normative and descriptive uses of the AHP, or of any decision theory for that matter, must be understood within their appropriate domains. The claim that some decision support technique is both normatively acceptable and descriptively adequate is in itself contradictory. Why would a support methodology be needed which describes how unaided decisions are made and, at the same time, prescribes how these decisions should be made?

7. CONCLUSIONS

Starting from the foundations of multiattribute value measurement, we have demonstrated that the pairwise comparisons in ratio estimation should be interpreted in terms of value differences between pairs of underlying alternatives. This interpretation is general and applies to all methods (including the AHP and SMART) which make use of ratio statements in the elicitation of hierarchical models. In fact, when the questions in the AHP are rephrased according to the value difference interpretation, the AHP can be regarded as a variant of multiattribute value measurement. While it is still unclear to what extent the DM's intuitive responses to the standard AHP questions conform to the value difference interpretation, we feel that AHP practitioners could improve their analyses by stating the pairwise comparison questions accordingly.

The other topics, i.e. the choice of the scale and whether or not to use normalizations, are important issues which should be seen as practical procedural choices whose consequences need to be understood. Although discretized ratio scales such as the 1–9 scale of the AHP can be very helpful in preference elicitation, they are nevertheless problematic as they severely restrict the range and distribution of possible priority vectors. The balanced scales proposed in this paper provide an essential improvement in this matter. Even so, the assumption that verbal expressions can be mapped onto numbers in the same way, no matter who is responding and in what context, must be regarded with due caution. The implication of scale selections must be considered explicitly, especially if the results are to be used in a normative sense. Risks associated with scale selection can be mitigated through software tools which allow the practitioner to compare results based on different scales.

Often hierarchical weighting procedures such as the AHP are used to create an improved problem understanding and to support communication among a group of decision makers with little interest in the details of deriving numerical results. Even in this kind of case the analyst should make every effort to explain techniques. The decision makers need to understand that both the structure of the hierarchy and the criteria weights need to reflect the set of decision alternatives and their differences.

REFERENCES

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

Belton, V. and Gear, T., 'On a short-coming of Saaty's method of analytic hierarchies', *Omega*, **11**, 228–230 (1983).

Beyth-Marom, R., 'How probable is 'probable'?: numerical translation of verbal probability expressions', *J. Forecast.*, **1**, 257–269 (1982).

Birnbaum, M., 'Differences and ratios in psychological measurements', in *Cognitive Theory*, Vol. 3, New York: Wiley, 1978, pp. 33–74.

Buede, D. M., 'Software review. Three packages for AHP: Criterium, Expert Choice and HIPRE 3+', *J. Multi-Crit. Decis. Anal.*, **1**, 119–121 (1992).

Dyer, J. S., 'Remarks on the analytic hierarchy process', *Manag. Sci.*, **36**, 249–258 (1990).

Dyer, J. S. and Sarin, R. K., 'Measurable multiattribute utility functions', *Oper. Res.*, **27**, 810–822 (1979).

Edwards, W., 'How to use multiattribute utility measurement for social decisionmaking', *IEEE Trans. Syst., Man, Cyber.*, **7**, 326–340 (1977).

Edwards, W. and Barron, F. H., 'SMARTS and SMARTER: improved simple methods of multiattribute utility measurement', *Organiz. Behav. Human Decis. Process.*, **60**, 306–325 (1994).

Farquhar, P. C. and Keller, L. R., 'Preference intensity measurement', *Ann. Oper. Res.*, **19**, 205–217 (1989).

Hämäläinen, R. P. and Lauri, H., *HIPRE 3+ Decision Support Software*, Systems Analysis Laboratory, Helsinki University of Technology, 1995 (distributed by EIA Inc.).

Hämäläinen, R. P. and Pöyhönen, M., 'On-line group decision support by preference programming in traffic planning', *Group Decis. Negot.*, **5**, 485–500 (1996).

Hämäläinen, R. P., Salo, A. A. and Pöysti, K., 'Observations about consensus seeking in a multiple criteria environment', *Proc. Twenty-Fifth Hawaii Int. Conf. on System Sciences*, 1991, Vol. IV, pp. 190–198.

Hämäläinen, R. P. and Seppäläinen, T. O., 'The analytic network process in energy planning', *Socio-Econ. Plan. Sci.*, **20**, 399–405 (1986).

Harker, P. T. and Vargas, L. G., 'The theory of ratio scale estimation: Saaty's analytic hierarchy process', *Manag. Sci.*, **33**, 1383–1403 (1987).

Kamenetzky, R. D., 'The relationship between the analytic hierarchy process and the additive value function', *Decis. Sci.*, **13**, 702–713 (1982).

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

Krantz, D. H., Luce, R. D., Suppes, P. and Tversky, A., *Foundations of Measurement*, Vol. I, New York: Academic, 1971.

Lootsma, F. A., 'Scale sensitivity and rank preservation in a multiplicative variant of the AHP and SMART', *Rep. 91-67*, Delft University of Technology, 1991.

Lund, J. R. and Palmer, N., 'Subjective evaluation: linguistic scales in pairwise comparison methods', *Civil Eng. Syst.*, **3**, 182–186 (1986).

Ma, D. and Zheng, X., '9/9–9/1 scale method of AHP', *Proc. 2nd Int. Symp. on the AHP*, Pittsburgh, PA: University of Pittsburgh, 1991, Vol. I, pp. 197–202.

Pöyhönen, M. A. and Hämäläinen, R. P., 'Notes on the weighting biases in value trees', *Res. Rep. A63*, Systems Analysis Laboratory, Helsinki University of Technology, 1996.

Pöyhönen, M. A., Hämäläinen, R. P. and Salo, A. A., 'An experiment on the numerical modelling of verbal ratio statements', *J. Multi-Crit. Decis. Anal.*, **6**, 1–10 (1997).

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

Saaty, T. L., 'Highlights and critical points in the theory and application of the analytic hierarchy process', *Eur. J. Oper. Res.*, **74**, 426–447 (1994).

Saaty, T. L. and Takizawa, M., 'Dependence and independence: from linear hierarchies to nonlinear networks', *Eur. J. Oper. Res.*, **26**, 229–237 (1986).

Saaty, T. L., Vargas, L. G. and Wendell, R. E., 'Assessing attribute weights by ratios', *Omega*, **11**, 9–12 (1983).

Salo, A. A., 'Inconsistency analysis by approximately specified priorities', *Math. Comput. Model.*, **17**, 123–133 (1993).

Salo, A. A. and Hämäläinen, R. P., 'Preference assessment by imprecise ratio statements', *Oper. Res.*, **40**, 1053–1061 (1992a).

Salo, A. A. and Hämäläinen, R. P., 'Rank reversals in the feedback technique of the analytic hierarchy process', *Res. Rep. A45*, Systems Analysis Laboratory, Helsinki University of Technology, 1992b.

Salo, A. A. and Hämäläinen, R. P., 'Preference programming through approximate ratio comparisons', *Eur. J. Oper. Res.*, **82**, 458–475 (1995).

Schenkerman, S., 'Avoiding rank reversal in AHP decision-support models', *Eur. J. Oper. Res.*, **74**, 407–419 (1994).

Schoner, B. and Wedley, W. C., 'Ambiguous criteria weights in AHP: consequences and solutions', *Decis. Sci.*, **20**, 462–475 (1989a).

Schoner, B. and Wedley, W. C., 'Alternative scales in the AHP', in Lockett, A. G. and Islei, G. (eds.), *Improving Decision in Organisations*, LNEMS Vol. 335, Berlin: Springer, 1989b, pp. 345–354.

Schoner, B., Wedley, W. C. and Choo, E. U., 'A unified approach to AHP with linking pins', *Eur. J. Oper. Res.*, **64**, 384–392 (1993).

Stillwell, W. G., von Winterfeldt, D. and John, R. S., 'Comparing hierarchical and nonhierarchical weighting methods for eliciting multiattribute value models', *Manag. Sci.*, **33**, 442–450 (1987).

Vargas, L. G., 'An overview of the analytic hierarchy process and its applications', *Eur. J. Oper. Res.*, **48**, 2–8 (1990).

Vargas, L. G., 'Reply to Schenkerman's avoiding rank reversal in AHP decision support models', *Eur. J. Oper. Res.*, **74**, 420–425 (1994).

Watson, S. R. and Freeling, A. N. S., 'Assessing attribute weights', *Omega*, **10**, 582–585 (1982).

Weber, M. and Borcherding, K., 'Behavioral influences on weight judgments in multiattribute decision making', *Eur. J. Oper. Res.*, **67**, 1–12 (1993).

