ISAHP 1996, Vancouver, Canada, July 12-15, 1996

A MULTI-CRITERIA COMPARISON OF RESPONSE SCALES AND SCALING METHODS IN THE AHP[1]

David V. Budescu, Bradley D. Crouch and Osvaldo F. Morera
Department of Psychology
The University of Illinois at Urbana-Champaign
dbudescu@s.psych.uiuc.edu

Abstract: Two methodological byproducts of the recent intensive work within the Analytical Hierarchy Process (AHP) framework are (1) the development of new (and refinement of some old) scaling techniques for the extraction of the priority weights from a matrix of direct ratio judgments and (2) the derivation of alternative scales for the pairwise comparisons of the AHP. This paper compares a number of these scales in combination with some of the methods used to extract priority weights for three distinct types of pairwise comparisons in the AHP. The three types of matrices consist of *between-attributes* comparisons, *within-attribute* comparisons and *global* comparisons (Jensen, 1984a). The methods and scales are compared in terms of several criteria including the solution's goodness of fit, the variance and entropy of the priority weights and the number of order violations. Results indicate that the nature of the scale has systematic and significant effects of various characteristics of the solution, under all scaling methods. In particular, increasing the spacing of the scale points tends to increase the differentiation among the priority weights, but has a negative effect on the solution's goodness of fit.

Introduction

The Analytic Hierarchy Process (AHP) is a popular multi-criteria decision methodology which was developed and discussed extensively by Saaty (e.g. 1977, 1980, 1986, 1990a). Since its introduction, the AHP has been applied in a wide variety of domains and disciplines (see Zahedi, 1986a; Saaty & Vargas, 1982, 1991 for representative lists). A typical AHP analysis consists of four interrelated stages:

(1) The decision problem is structured as a dominance hierarchy.
(2) Data are collected through a process of pairwise comparisons among all elements at a specific level of the hierarchy with respect to single, well-defined, criteria from higher levels of the hierarchy.
(3) Priority weights are extracted from each set of comparisons obtained at stage (2) through an appropriate scaling/estimation procedure.
(4) The various weights derived at stage (3) are combined, using a particular aggregation model, to yield an overall weight for each alternative.

Consider, for example, a student who is looking for a campus apartment and has to choose among one of $n = 5$ alternatives. At the top level of the hierarchy lies the goal of "Selecting the best apartment". In structuring the problem, the student decides to focus on $p = 5$ sub-criteria at the second level (e.g. cost, size, location, reputation of the landlord and amenities). This hierarchy is displayed Figure 1.

[1] Bradley D. Crouch's work was supported by a National Science Foundation Graduate Fellowship.

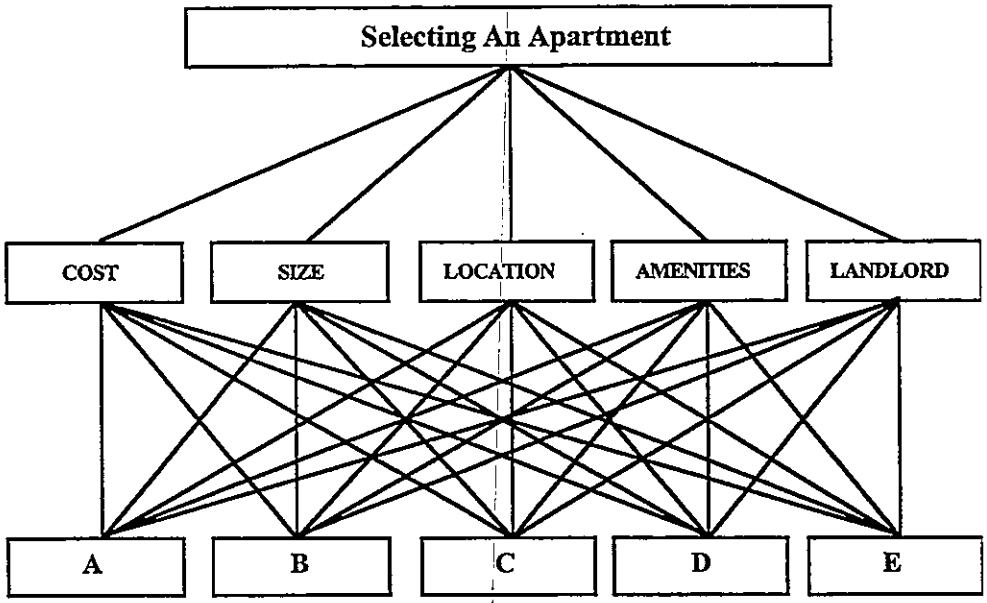


FIGURE 1. A hierarchy for apartment selection

At the second stage of the process the student compares, in pairwise fashion, all attributes (we will refer to these comparisons as *between-attributes)* as well as all apartments on each of the attributes (we label these *within-attribute* comparisons). At the third stage, each of these sets of comparisons is analyzed and priority weights are extracted. Let w_i be the weight assigned to the i_{th} criterion (i=1..p) at level 2, and v_{ij} be the estimated weight of the j_{th} apartment (j = 1..n) on criterion i. Finally, these estimates are combined to yield V_j, the inferred overall weight of apartment j. This is defined as a simple weighted sum of the criterion specific priorities (over all criteria), i.e.:

$$V_j = \sum_{i=1}^{p} w_i v_{ij} \qquad (1)$$

A simplified approach to this decision problem (Jensen, 1984a) requires the Decision Maker (DM) to compare all the apartments in a pairwise fashion, *considering all the criteria simultaneously* (we will refer to these comparisons as *global*). In other words, one level of the typical AHP hierarchy is eliminated. Figure 2 depicts this simplified version of the AHP for the same decision.

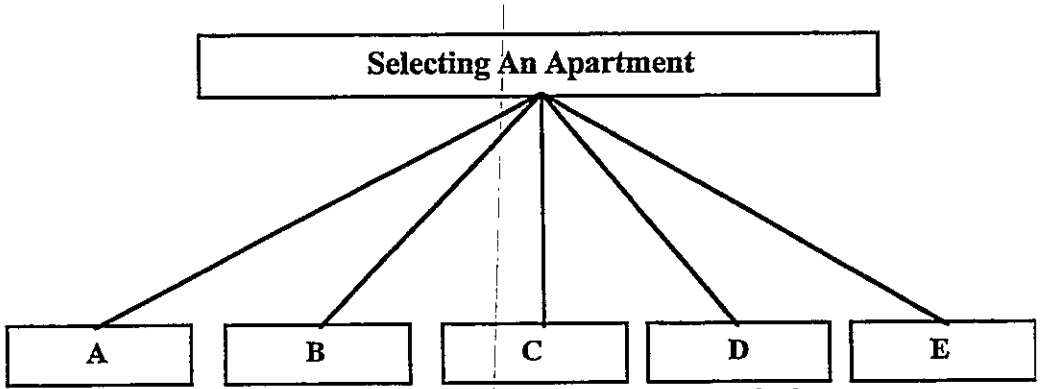


FIGURE 2. A global hierarchy for apartment selection

Two methodological byproducts of the work with AHP are: (1) The development of new (and refinement of some old) scaling and estimation techniques to be used in stage 3 of the AHP in extracting priority weights from a matrix of direct ratio judgments, and (2) The derivation of alternative scales to be used in the process of pairwise comparisons (*within-* and *between-* attributes) in the AHP. This paper provides a comparison of a number of these scales in combination with the most popular methods used to extract priority weights, by applying all scales and solutions to a set of judgments obtained as part of decision making study, in a

meaningful context. To motivate this work, we point out that the choices made with regard to the nature of the specific scale used in the elicitation of judgments and the scaling methodology can affect the nature of the eventual decisions (see Lootsma, 1993; Olson, Fliedner & Currie, 1995 for recent examples).

Scaling a matrix of ratio judgments: An Overview

The ratio judgments can be represented in a square matrix R (with elements r_{ij}), of order n, with the following properties (for all $i,j = 1..n$):

(i) Positivity: $r_{ij} > 0$ *(ii)* Reciprocity: $r_{ij} = 1/r_{ji}$, *(iii)* Reflexivity: $r_{ii} = 1$.

Assume now a simple model relating these overt judgments to an unobservable vector of values, $\mathbf{w} = (w_1,..w_n)$, which represent the priority weights of the n stimuli:

$$r_{ij} = w_i / w_j . \tag{2}$$

Under this model, the matrix R is fully determined by the vector w. Thus, R is of unit rank and, for any triple of distinct stimuli ($i,j,k=1...n$), we expect:

(iv) Consistency: $r_{ij} \bullet r_{jk} = r_{ik}$ (3)

It follows that R has a single positive eigenvalue, $\lambda=n$, and n-1 zero eigenvalues (Saaty, 1977, 1986).

The estimation problem is to infer the set of weights, w, underlying a given matrix of empirical judgments, R. Under this model any column of R is a solution to this *algebraic* problem. Solutions are unique up to multiplication by a positive constant, so it is customary to normalize the weights by imposing the constraint $\Sigma w_i=1$. Recall however, that R is obtained from a sequence of fallible human judgments and, therefore, conditions *(i)-(iv)* may not be satisfied. Positivity is enforced by the nature of the scale used. Typically, judgments are elicited for $n(n-1)/2$ pairs of distinct stimuli and the rest of the entries are calculated such that conditions *(ii)-(iii)* are satisfied. However, this does not guarantee consistency. In such cases, the researcher is faced with the *statistical* problem of estimating the weights from a redundant (and, possibly, inconsistent) set of judgments (Weber & Borcherding, 1993). Much of the work in this area has been devoted to the comparison of various estimation methods and identification of "the best" among them. Partial reviews of the methods and some of their properties are found in Budescu (1984), Cook & Kress (1992), Golany & Kress, (1993) and Saaty & Vargas (1984).

Saaty (1977, 1980) suggested estimating w by $\mathbf{w}^{(r)}$, the right eigenvector of R corresponding to its largest eigenvalue, i.e. by solving the eigenvalue problem $\mathbf{R}\mathbf{w}^{(r)} = \lambda\mathbf{w}^{(r)}$. He also provided a simple numerical algorithm for solving this problem (we refer to this as the EV solution). The inconsistency of the solution is measured by the index:

$$\mu = (\lambda - n) / (n - 1), \tag{4}$$

which vanishes for perfectly consistent matrices. Saaty and Vargas (1984) present a loss function which is optimized by this solution. Interestingly, Gulliksen (1959, 1975) applied a very similar procedure for ratio scaling. Cogger and Yu (1985) proposed a modification to the EV method, which is easier to compute but lacks the intuitive appeal and the optimal properties of the EV.

Jensen (1984b) described a Direct Least Squares (DLS) procedure, but it is seldom used because it yields multiple solutions and the computation is difficult (Golany & Kress, 1993). However, Chu, Kalaba & Spingarn (1979) proposed a unique and computationally feasible Weighted Least Squares (WLS) solution, $\mathbf{w}^{(w)}$, which minimizes the quantity:

$$S^{(w)} = \Sigma\Sigma \ (r_{ij} w^{(w)}_i - w^{(w)}_j)^2. \tag{5}$$

Barzilai, Cook & Golany (1987), Crawford & Williams (1985), Crawford (1987), deJong (1984) and Lootsma (1993) favor the Logarithmic Least Squares (LLS) method. The solution $\mathbf{w}^{(g)}$, which minimizes the quantity:

$$S^{(g)} = \Sigma\Sigma \left[\ln (r_{ij}) - \ln (w^{(g)}_i / w^{(g)}_j) \right]^2, \tag{6}$$

can be shown to consist, simply, of the geometric means of the rows of R (it appears that Torgerson, 1958, was the first to establish this result). Due its simplicity, Saaty and Kearns (1985) suggest using geometric means as an approximation to the EV solution. Because of its intuitive appeal and the ease of its computation, $\mathbf{w}^{(g)}$ has become the most popular method, next to $\mathbf{w}^{(r)}$.

Finally, Cook and Kress (1982) offer a compelling axiomatization of this problem that favors a Logarithmic Least Absolute Values (LLAV) solution. Unfortunately, LLAV has multiple (and difficult to calculate) solutions.

Comparisons of the various methods (e.g. Budescu, Zwick & Rapoport, 1986; Golany & Kress, 1993; Saaty 1990b; Saaty & Vargas, 1984; Takeda, Cogger & Yu, 1987; 1984; Zahedi, 1986b) found that in most cases there is good agreement between the various solutions, and neither is uniformly superior. Whenever discrepancies between the solutions emerge, they reflect the interaction between the different loss functions employed and specific features of the data. For the purpose of this paper, we will focus on the three following solutions, yielding unique solutions: EV, WLS and LLS solutions (denoted $\mathbf{w}^{(r)}$, $\mathbf{w}^{(w)}$, and $\mathbf{w}^{(g)}$, respectively).

Ratio judgments and the nature of the scale

Of course, all the methods described above apply to any matrix R satisfying requirements *(i)-(iii)*. However, Saaty (1977, 1980) proposed, for theoretical and practical reasons, to restrict the ratio judgments to the integers 1-9 and their reciprocals (Donegan, Dodd and McMaster (1992) refer to these 17 values as "the Saaty set"). To facilitate their use, Saaty suggested using specific verbal labels in conjunction with some of these values. These labels are supposed to convey the relative importance of the superior element within each pair, relative to the second element, and are to be applied for all comparisons. In his original paper, Saaty (1977) proposed the labels *1: equal, 3: weak, 5: essential (or strong), 7: demonstrated, 9: absolute*[2], but others have used slightly different labels (see Pöyhönen, Hämäläinen & Salo, 1996 and Donegan, Dodd & McMaster, 1992 for examples).

In recent years, researchers have tried to determine whether Saaty's set is appropriate in terms of its size and the spacing of its values for all comparisons and a number of alternative scales that have been proposed. The earliest allusion to this possibility is due to Harker (1987), who showed that the eigenvalue method can be extended to the case where the observed judgments are power transforms of the Saaty set. He hinted at the possibility of employing various non-linear scales but stopped short of endorsing a specific scale, or proposing a general choice criterion.

Lootsma (1993) proposed a scale based on the assumption that human judgments follow a geometric progression with a fixed factor. The scale values in this system (called Ratio Estimation in Magnitudes or deci-Bels to Rate Alternatives which are Non-Dominated, or REMBRANDT for short) can be expressed as:

$$r^{(l)}_{ij} = exp\left[\tau \left(r_{ij} - 1 \right) \right] \qquad \text{if } r_{ij} \geq 1, \tag{7a}$$

$$r^{(l)}_{ij} = exp\left[-\tau \left(r_{ij} - 1 \right) \right] \qquad \text{if } r_{ij} < 1 \tag{7b}$$

where r_{ij} are the values in the Saaty set and τ is the progression factor. Lootsma recommends $\tau_1 = ln(2) \approx 0.7$ and $\tau_1 = ln(\sqrt{2}) \approx 0.35$ as "natural" factors for alternatives and criteria respectively. The most obvious problem associated with the use of a (relatively) small and restricted set of values is the inherent inconsistency with the mathematical form, and the implications of the model (Eqs. 3 and 4). For example, of the 969 distinct triples that can be generated within the Saaty set, only 45 (4.6%) satisfy Eq. 4. If all indifference judgments, i.e. $r_{ij}=1$,

[2]The even numbers (used to define intermediate levels) are typically not labeled.

are eliminated, this number is further reduced to 20 (2.1%)! It follows that a fully consistent judge would experience serious difficulties in expressing his priorities. If, for example, he thinks that both r_{ij} and $r_{jk} = 8$, we would expect him to judge option i to be 64 times better than k, but this calls for values outside the restricted set. As a partial remedy for this problem, Donegan et al. (1992) and Dodd, Donegan & McMaster (1995) proposed a Modified AHP (MAHP) in which the response scale is fairly linear in the middle of the scale and distinctly non-linear near the end points. Of the various possible stretching functions they selected, quite arbitrarily, $tanh^{-1}$, the inverse hyperbolic tangent. The new scale is defined by:

$$r^{(2)}_{ij} = exp\ \{\ tanh^{-1}\ [\ (r_{ij} - 1) / (H - 1)\ ]\ \} \qquad \text{if } r_{ij} \geq 1 \tag{8a}$$

$$r^{(2)}_{ij} = exp\ \{\ -tanh^{-1}\ [\ (r_{ij}^{-1} - 1) / (H - 1)\ ]\ \} \qquad \text{if } r_{ij} < 1, \tag{8b}$$

where H is the "Horizon" parameter. An horizon of 8, for example, yields $H=(1 + 14\sqrt{3}) \approx 9.083$. This choice guarantees that the most extreme ratio (corresponding to $r_{ik}=9$) is obtained as a product of two ratios corresponding to $r_{ij} = r_{jk} = 8$, thus solving the problem described above. Alternatively, an horizon of 7 implies, $H=(1 + 6\sqrt{2}) \approx 9.485$, guaranteeing that the most extreme ratio (corresponding to $r_{ik}=9$) is obtained as a product of two ratios corresponding to $r_{ij}=r_{jk}=7$.

Table 1 displays values prescribed by the various scales described here for the 1-9 range. To illustrate the differences between the scales, we also list three informative statistics for each of them: The variance of the scale values, $V(s_i)$; the inter-quartile range of the values, $IQR(s_i)$; and the variance of the "gaps" (the distances between adjacent scale values), $V(s_i - s_{i-1})$. The latter is an index of the departure from linearity of the scale (note that it is 0 for Saaty's linear scale).

TABLE 1. Seven alternative nine-point scales for the AHP and some summary statistics

Saaty $\alpha = 1.0$	Power $\alpha = 0.5$	Power $\alpha = 2.0$	REMBRANDT $\tau \approx 0.7$	REMBRANDT $\tau \approx 0.35$	MAHP 7-Based	MAHP 8-Based
1	1.00	1	1	1.00	1.00	1.00
2	1.41	4	2	1.41	1.13	1.13
3	1.73	9	4	2.00	1.27	1.29
4	2.00	16	8	2.83	1.45	1.47
5	2.24	25	16	4.00	1.67	1.72
6	2.45	36	32	5.66	1.97	2.06
7	2.65	49	64	8.00	2.42	2.60
8	2.83	64	128	11.31	3.23	3.73
9	3.00	81	256	16.00	5.83	13.93
Variance of scale values: $V(s_i) = [\ \sum s_i^2 - (\sum s_i)^2 / n\ ] / (n-1)$						
7.5	0.45	788	7,296	26.00	3.33	16.88
Interquartile range of scale values: $IQR(s_i)$						
4	0.92	40	60	6.00	1.15	1.31
Variance of gaps: $V(s_i - s_{i-1})$						
0.0	0.01	24	1,960	2.23	0.70	12.14

Each of these scales is a continuous and monotonic transformation of the Saaty set, and each has some desirable properties. However, they all invoke, more or less, arbitrary assumptions regarding the functional form of the key transformation and/or its parameters. Clearly, neither is uniformly superior to the others on normative grounds. Also, in the absence of comprehensive comparative studies of the various scales (see Olson, et al., 1995 for a recent exception) it is impossible to determine which scale fits better most cases in an empirical sense.

An empirical multi-criteria comparison of various response scales

Previous empirical work has focused primarily on comparisons between scaling methods (EV, WLS, LLS, etc.) and has devoted only limited attention to the comparison of the various scales. In particular, there are no studies comparing systematically the effects of using different scales (MAHP, REMBRANDT, Power, etc.) in combination with the various solution concepts. Our study will address this issue. We will apply three scaling methods (EV, LLS and WLS) to analyze a set of matrices represented in seven distinct alternative scales (Saaty's 1-9 scale, power transformations with $\alpha = 0.5$ and 2, REMBRANDT with $\tau = 0.35$ and 0.7, and MAHP with a horizon of 7 and 8).

The matrices consists of actual judgments obtained from judges in real decision contexts. We will analyze three types of matrices consisting of *between-attributes*, *within-attribute* and *global* comparisons, in the same domain. This is an intriguing comparison. From a mathematical point of view, the various matrices are identical, but in practice DMs may treat them differently. There are at least two justifications for this speculation. The first is "structural": Typically, the DM has more control over the selection of the relevant attributes than over the selection of the alternatives. This would suggest more homogeneity among attributes than alternatives. This appears to be the implicit assumption underlying Lootsma's (1993) recommendation to use different scaling factors for the *within-attribute* and *between-attributes* judgments. The second justification is "cognitive". It is conceivable that DMs invoke different psychological processes when comparing concrete alternatives versus more abstract attributes (Payne, Bettman & Johnson, 1993).

Whenever several solutions to the same problem are compared the question of the most appropriate criterion arises. The problem is even more acute in this case when the solutions being compared vary along two dimensions. As mentioned in the introduction, various methods were designed to optimize different criteria. Obviously, none of the solution specific criteria, such as Saaty's μ, can be used to compare all the solutions, although these indices are meaningful in the context of a specific method. For example, it is informative to compare the $S^{(2)}$ of the various LLS solutions, but it makes little sense to calculate this measure for the EV or LLS solution.

A natural and compelling criterion of comparison is the solution's *external validity*, i.e. its ability to capture and reproduce accurately the DM's "true" priorities. This compelling universal criterion requires direct access to one's preferences. Unfortunately, this information is rarely available, so most comparisons rely on alternative (proxy) criteria. Golany & Kress (1993) pointed out, and illustrated, the importance of considering multiple criteria in order to obtain a complete and comprehensive understanding of the properties of the various solutions.

In the current study we will adopt a similar approach, and will compare the alternative solutions and scales along the following four criteria:

(1) The variance of the derived weights:

$$S^2(w_i) = [\sum w_i^2 - 1/n]/(n-1). \qquad (9)$$

(2) The relative entropy of the weights (Noble & Sanchez, 1993):

$$RE(w_i) = 1 - \sum w_i \ ln\ (w_i) / ln\ (n). \qquad (10)$$

Note that $0 \leq RE \leq 1$ such that $RE = 0$ when all the weights are equal, and RE approaches 1 as one weight approaches 1, while the other n-1 vanish.

The number of order reversals: The number of cases in which one alternative *is judged superior* to another, but *assigned a lower weight* by the solution. We distinguish between two types of reversals:

(3) Strong Order Reversals (SOR) is a count of all cases in which $r_{ik} < r_{jk}$ but $w_j < w_i$, or $r_{ik} > r_{jk}$ but $w_j > w_i$.

(4) Weak Order Reversals (WOR) is a count of all cases in which $r_{ik} \neq r_{jk}$ but $w_j = w_i$ or $r_{ik} = r_{jk}$ but $w_j \neq w_i$.

> Golany & Kress (1993) analyzed a global index of ordinal consistency, TOR, which is equivalent to (SOR + WOR / 2).

One would expect a "good" solution to be highly informative, differentiate well between the n entities, and to have very few order reversals (if any).

Method

Subjects: The present study involved 29 students from an introductory psychology class at the University of Illinois at Urbana-Champaign that completed the study in partial fulfillment of the course requirements.

Procedure: All subjects were asked to evaluate five hypothetical apartments that varied with respect to five attributes: rent, size, proximity to campus, amenities and landlord reputation. The information concerning the apartments was presented to subjects in a summary packet. All comparisons were made on a nine-point scale.

The stimuli were constructed such that no one apartment clearly dominated any of the other four. In other words, each apartment was the best in the set on one attribute, second in the set on another attribute, third on one attribute, and so on. The apartments were constructed in this manner so that no one apartment could be obviously considered to be the "best" apartment in the global comparisons.

Fifteen subjects assessed apartments using the global version of the AHP (Jensen, 1983), and 14 subjects used the regular (decomposed) procedure. In the latter group the *between-attributes* comparisons preceded the five *within-attribute* sets of judgments. Experimental sessions lasted no more than one hour.

Results

The results for each of the criteria considered were analyzed and summarized in the framework of a multi-factor Analysis of Variance (ANOVA) with the following three factors:

(1) Scale type: Saaty's 1-9, MAHP with a horizon of 7 and 8, REMBRANDT with $\tau_1 = ln\ (2) \approx 0.7$ and $\tau_2 = \ln \sqrt{2} \approx 0.35$ and power transformations with $\alpha = 0.5$ and $\alpha = 2.0$ (Ofcourse, Saaty's linear scale is also a member of this family of transformations with $\alpha = 1$).

(2) Solution method: EV, LLS and WLS.

(3) Type of matrix (global, between, and within). Note that all three types of matrices are of size $n = 5$, so the results can be easily compared. A slight problem is introduced by the fact that subjects generated five within matrices (one for each attribute), but only one between or global matrix. To simplify comparisons across types of matrices, in all subsequent analyses the results of the within attribute judgments are based on average values, taken across the five attributes, for each of the analysis criteria (such as RI, SOR, etc.).

Somewhat surprisingly, we found no significant differences between the results for the three types of matrices. Since this pattern holds for all the criteria considered, all subsequent figures and tables present results averaged across the three types of matrices.

Measures of goodness of fit: Table 2 presents the mean goodness of fit for each of the three solutions and scales compared. Recall that each of these indices is computed in a different metric and, therefore, comparisons between solutions are not possible. Note, however, that for each solution method the scales are ranked identically in terms of the goodness of their implied solutions. The seven scales can be clustered into three coarse but distinct groups: At one end we observe excellent levels of fit for the two MAHP scales (7 and 8-

based) and the "square root" scale; At the other extreme we observe extremely low levels of fit for the "squared" scale and the REMBRANDT scale with $\tau_1 = ln\ (2) \approx 0.7$. Saaty's linear scale and the REMBRANDT scale with $\tau_2 = \ln \sqrt{2} \approx 0.35$ are in between (but much closer to the desirable end). Note that this clustering corresponds with the grouping of the seven scales in terms of their IQR(s_i) in Table 1.

TABLE 2. Mean (and SD) of goodness of fit by response scale and scaling method

Response Scale	EV	LLS	WLS
MAHP (8-Based)	.05 (0.06)	.34 (0.36)	.10 (0.12)
Power (α=0.5)	.07 (0.04)	.43 (0.24)	.13 (0.09)
MAHP (7-Based)	.10 (0.15)	.58 (0.79)	.16 (0.22)
REMBRANDT (τ=*ln* $\sqrt{2}$)	.28 (0.23)	1.55 (1.11)	.16 (0.22)
Saaty	.31 (0.20)	1.72 (0.94)	.79 (1.00)
REMBRANDT (τ=*ln* 2)	1.76 (2.21)	6.19 (4.45)	18.20 (98.07)
Power (α=2.0)	1.87 (1.70)	6.89 (3.76)	16.41 (54.01)
Mean	.63 (1.29)	2.53 (3.45)	5.21 (42.60)

Variance and entropy of scale values: Tables 3 and 4 present the variance and entropy of the priority weights for the 21 solutions compared. The results for these two criteria are highly similar, so they can be discussed jointly. There are several noticeable features in both tables. With only a few minor and insignificant exceptions, the seven scales are ordered identically for the three solutions, and the three solutions are ordered identically for each of the seven scales. Clearly, the EV and LLS solutions are practically identical and they yield slightly more homogeneous (and less informative) weights than the WLS solution. The seven scales are clustered in the same three classes identified above: The two "stretched" scales (squared and REMBRANDT scale with $\tau_1 = ln\ (2) \approx 0.7$ are characterized by the highest levels of differentiation between weights, the three "shrunken" scales (the two MAHP scales and the square root scale) yield the most homogeneous weights and, as before, the central cluster consists of the linear scale and the REMBRANDT scale with $\tau_2 = \ln \sqrt{2} \approx 0.35$. Finally, note that the WLS solution is considerably more sensitive to the nature of the scale used than the other two.

TABLE 3. Mean (and SD) of variance of weights by response scale and scaling method

Response Scale	EV	LLS	WLS	Mean
MAHP (8-Based)	.01 (0.01)	.01 (.01)	.01 (0.01)	.01 (.01)
Power (α=0.5)	.01 (.003)	.01 (.003)	.01 (.004)	.01 (.003)
MAHP (7-Based)	.01 (0.01)	.01 (0.01)	.01 (0.02)	.01 (0.02)
Saaty	.03 (0.01)	.03 (0.01)	.04 (0.02)	.03 (0.01)
REMBRANDT (τ=*ln* $\sqrt{2}$)	.03 (0.01)	.03 (0.02)	.04 (0.04)	.03 (0.02)
REMBRANDT (τ=*ln* 2)	.07 (0.02)	.07 (0.03)	.09 (0.04)	.08 (0.04)
Power (α=2.0)	.07 (0.03)	.07 (0.03)	.10 (0.04)	.08 (0.03)
Mean	.03 (0.03)	.03 (0.03)	.04 (0.04)	

TABLE 4. Mean (and SD) entropy of weights by response scale and scaling method

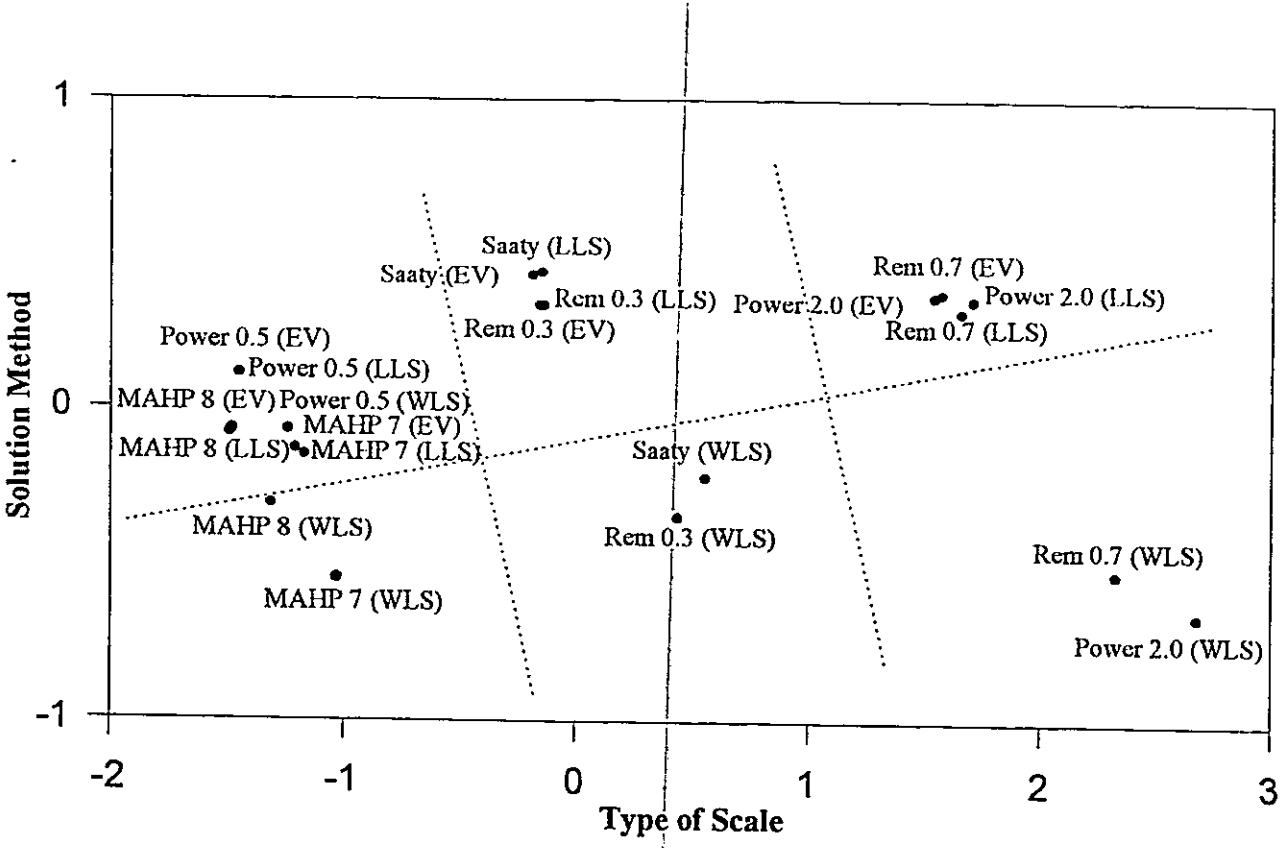
Response Scale	EV	LLS	WLS	Mean
MAHP (8-Based)	.09 (0.07)	.09 (0.07)	.10 (0.09)	.09 (0.08)
Power (α=0.5)	.10 (0.03)	.10 (0.03)	.11 (0.04)	.10 (0.03)
MAHP (7-Based)	.13 (0.12)	.13 (0.12)	.15 (0.17)	.14 (0.14)
Saaty	.32 (0.10)	.33 (0.10)	.41 (0.14)	.35 (0.12)
REMBRANDT (τ=*ln* $\sqrt{2}$)	.32 (0.14)	.33 (0.15)	.41 (0.20)	.35 (0.17)
REMBRANDT (τ=*ln* 2)	.76 (0.29)	.77 (0.29)	.97 (0.30)	.83 (0.31)
Power (α=2.0)	.75 (0.22)	.79 (0.22)	1.04 (0.28)	.86 (0.27)
Mean	.35 (0.32)	.36 (0.33)	.45 (0.42)	.39 (0.36)

Order reversals: In the next analysis we focus on the ordinal properties of the data and the solutions. Table 5 displays the number of strong order reversals. In evaluating these results, keep in mind that the maximal number of reversals is a function of the matrix size. In general, $0 \leq$ WOR, SOR $\leq n(n\text{-}1)/2$, and in our case $0 \leq$ WOR, SOR ≤ 10. The most salient feature of the Table 5 is that WLS induces almost twice as many order reversals as LLS and EV for most, and across all, scales examined. Consider now the rate of order violations observed for the different scales when using EV and/or LLS solutions. In general the most accurate and faithful representation of the original ordering of the apartments, is obtained under the family of power transformations ($3\% \leq$ SOR $\leq 3.35\%$), followed by the family of exponential transformations (i.e. the REMBRANDT scales, where $4.33\% \leq$ SOR $\leq 4.84\%$), and the worst results are obtained for MAHP ($5.67\% \leq$ SOR $\leq 6.47\%$). We do not present a similar table of WOR because, with only two exceptions (the REMBRANDT scales analyzed by LLS), all combinations of scales and solutions yield, practically, identical rates of order violations (between 0.46 and 0.47, i.e. slightly under 5%).

TABLE 5. Mean (and SD) number of strong order reversals by response scale and scaling method

Response Scale	EV	LLS	WLS	Mean
Power (α=0.5)	.30 (0.61)	.34 (.67)	.50 (0.76)	.38 (.68)
Saaty	.30 (0.61)	.32 (.61)	.60 (0.80)	.40 (.70)
Power (α=2.0)	.43 (0.61)	.34 (.67)	1.03 (1.07)	.56 (.87)
REMBRANDT (τ=*ln*2)	.57 (0.65)	.46 (.68)	.97 (1.07)	.62 (.85)
MAHP (8-Based)	.65 (0.76)	.57 (.73)	.78 (0.93)	.64 (.81)
MAHP (7-Based)	.43 (0.81)	.65 (.78)	.82 (1.00)	.70 (.87)
REMBRANDT (τ=*ln*$\sqrt{2}$)	.43 (0.65)	.48 (.64)	1.20 (1.24)	.70 (.95)
Mean	.43 (0.68)	.45 (.69)	.84 (1.01)	.57 (.83)

Similarity of solution: We calculated the (Euclidean) distances between all 21 solutions. To examine the pattern of distances, we performed a Metric Multidimensional Scaling of this matrix (e.g. Davison, 1983; Schiffman, Reynolds & Young, 1981). This analysis represents the various solutions as points in a two-dimensional space. The configuration of points is displayed in Figure 3. This two dimensional representation fits the data almost perfectly (Stress = 0.02). To facilitate the interpretation of the resulting configuration we added a few separation lines. The first (horizontal) dimension reflects the *stretching* of the response scale and the two vertical lines separate between the three clusters described above with the three "shrunken scales" at the left end and the two "stretched" scales at opposite end. The second dimension reflects the *method of solution* used. The horizontal line separates between the WLS solutions (below the line) and the EV and LLS solutions. Note that, in most cases, these two solutions for a given type of scale are "nearest neighbors" and, for all practical purposes, indistinguishable.



Summary

The present study is unique in several ways. It is the first systematic comparison of a large number of nine-point scales, in conjunction with several solutions. Unlike most methodological studies in this domain, which tend to compare solutions of artificially simulated matrices, we analyzed actual judgments obtained in the course of a controlled experiment. The obvious drawback of this fact is that our results may reflect, to some degree, the specific features of the decision problem (the number and nature of attributes and apartments presented, peculiarities of the sample, etc.). On the other hand, this methodology has allowed us, for the first time, to contrast various characteristics of solutions extracted from different types of matrices without invoking any specific assumptions. Obviously, our conclusions need to be replicated and validated with a much larger, and more diverse, set of matrices. With this cautionary note in mind, we turn now to a summary of our main results.

The first, and most important, generalization from all our analyses is that the choice of scale is of crucial importance. Our results indicate that all characteristics of the solutions (goodness of fit, information, number of order violations, etc.) were affected by the nature of the scale and, in some cases, the effects were quite dramatic. Also note that the effects generalized across the three methods of solution, so one cannot bypass the problem of choosing a scale by invoking some "scale invariant" estimation method. Furthermore, note that in all our analyses, the effects due to the choice of scale were more extreme than differences associated with the choice of the solution method.

The second conclusion is that the effects of the scale choice are systematic and can be predicted, with some level of accuracy, from the nature of the scale. Recall that in all our analyses the seven scales clustered in the same three groups which corresponded, roughly, to the degree to which the 1-9 scale was stretched/shrunken. This is most clearly illustrated in the MDS of the 21 solutions. An interesting, and unexpected, observation is that this clustering is better predicted by the scales' IQR rather than their variances. To understand this pattern we re-examined the original matrices and tabulated the frequency of use of the various points on the scale. It turns out that, on the average, the extreme points of the nine-point scale are underused by comparison with those in its center. Specifically, only 9.5% of the judgments are "1" or "2" and 13.6% are "8" or "9" while the central five values (3 - 7) represent 76.9% of the total judgments[3]. Thus, it is not surprising that our results are better predicted by measures of scatter that underplay the effects of the extreme values.

[3]This result casts some doubts on the validity of the standard practice of assuming in simulation studies that responses are uniformly distributed across all scale values.

Interestingly, the transformations of the scales have opposite effects on the goodness of fit/consistency of the solutions and the information/variance of the priority weights. In general, "stretching" the scale by increasing the distance between its points tends to decrease the goodness of fit but, at the same time, it increases the degree of differentiation between the priority weights (obviously, "condensing" the scale, by bringing its points closer together, has the opposite effect). In our opinion, it is important that users of the AHP be sensitive to this tradeoff. Sometimes, attempts to maximize the consistency of the judgments may induce a very conservative use of the scale, i.e. using only a limited set of values and/or treating them as more homogeneous than originally intended. Such a process may yield highly consistent, but totally uninformative solutions. A better strategy would be to strive for solutions with maximal information, and acceptable levels of consistency (e.g. a CR < 0.1, as suggested by Saaty, 1977).

A surprising result in our analyses is the absence of systematic differences between the three types of matrices. Two aspects of our study cause us to interpret this result very cautiously and reserve judgment regarding its generalizability. We designed our study such that all judgment matrices be of the same order, and all judgments be performed in a relatively short time. However, in many decisions the number of comparisons is larger and, typically, the number of alternatives is larger than the number of attributes considered. We suspect that the way in which DMs use the scale is affected by the number of entities that are being compared. Our second reservation is motivated by one aspect of the analysis. Recall, that in order to simplify comparisons between the three types of matrices we averaged the within-attribute results across all attributes. Closer inspection of the responses for the attribute-specific judgments revealed some substantial differences across attributes. For example, when subjects compared the alternatives along the *most important* attribute they used the high end of the scale (8 and 9) in 29.1% of the judgments and the lower end (1 and 2) in 23.6% of the cases. By contrast, when the alternatives are compared along the *least important* attribute, the two ends of the scale were used in 13.6% and 38.2% of the cases, respectively. In other words, there seems to be a propensity to use extreme values for important attributes and to favor indifference judgments for the non-important ones. It is possible that by combining together all the attributes some real effects were obscured. We plan to follow up this point with more powerful and sensitive analyses in future studies.

Although our primary interest was in a comparison of the scales, we can't ignore the implications of our results to the long controversy regarding the choice of the appropriate method of scaling. Consistent with many previous studies (e.g. Golany & Kress, 1993), we found that in the vast majority of cases the EV and LLS are, practically, indistinguishable. In our opinion, they are superior to WLS in at least two important senses. First, they are more likely to preserve the ordinal properties of the original judgment (see also Golany & Kress, 1993, on this point). Second, and more in line with the focus of our work, WLS was shown to be more sensitive to nature of the scale used, for all the criteria considered.

References

Barzilai, J., Cook, W., & Golany, B. (1987). Consistent weights for judgment matrices of the relative importance of alternatives. *Operations Research Letters, 6,* 131-134.

Budescu, D.V. (1984). Scaling binary comparison matrices: A comment on Narasimhan's proposal. *Fuzzy Sets and Systems, 14,* 187-192.

Budescu, D.V., Zwick, R., & Rapoport, A. (1986). A comparison of the eigenvalue method and the geometric mean procedure for ratio scaling. *Applied Psychological Measurement, 10,* 69-78.

Chu, A.T.W., Kalaba, R.E., & Spingarn, K. (1979). A comparison of two methods for determining the weights of belonging to fuzzy sets. *Journal of Optimization Theory and Applications, 27,* 531-538.

Cogger, K.O., & Yu, P.L. (1985). Eigenweight vectors and least distance approximation for revealed preference in pairwise weight ratios. *Journal of Optimization Theory and Applications, 46,* 483-491.

Cook, W.D. & Kress, M. (1992). *Ordinal Information and Preference Structures: Decision Models and Applications.* Englewood Cliffs, NJ: Prentice Hall.

Crawford, G.B. (1987). The geometric mean procedure for estimating the scale of a judgment matrix. *Mathematical Modding, 9,* 327-334.

Crawford, G., & Williams, C. (1985). A note on the analysis of subjective judgment matrices. *Journal of Mathematical Psychology, 29,* 387-405.

deJong, P. (1984). A statistical approach to Saaty's scaling method for priorities. *Journal of Mathematical Psychology, 28,* 467-478.

Davison, M.L. (1983). *Multidimensional Scaling.* New York: Wiley.

Dodd, F.J., Donegan, H.A., & McMaster, T.B.M. (1995). Scale horizons in analytic hierarchies. *Journal of Multi-Criteria Decision Analysis, 4,* 177-188.

Donegan, H.A., Dodd, F.J., & McMaster, T.B.(1992). A new approach to AHP decision making. *The Statistician, 41*, 295-302.

Golany, B., & Kress, M. (1993). A multicriteria evaluation of methods for obtaining weights from ratio-scale matrices. *European Journal of Operational Research, 69*, 210-220.

Gulliksen, H. (1959). Mathematical solutions for psychological problems. *American Scientist, 47*, 178-201.

Gulliksen, H. (1975). Characteristic roots and vectors indicating agreement of data with different scaling laws. *The Indian Journal of Statistics, 37*, 363-384.

Harker, P.T. (1987). Alternative modes of questioning in the analytic hierarchy process. *Mathematical Modelling, 9*, 353-360.

Jensen, R.E. (1984a). Aggregation (Comparison) schema for eigenvector scaling of criteria priorities in hierarchical structures. *Multivariate Behavioral Research, 18*, 63-84.

Jensen, R.E. (1984b). An alternative scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology, 28*, 317-332.

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

Noble, E.E., & Sanchez, P.P. (1993). A note on the information content of a consistent pairwise comparison judgment matrix of an AHP decision maker. *Theory and Decision, 34*, 99-108.

Olson, D.L., Fliedner, G., & Currie, K. (195). Comparison of the REMBRANDT system with analytic hierarchy process. *European Journal of Operations Research, 82*, 522-539.

Payne, J.W., Bettman, J.R., & Johnson, E.J. (1993). *The Adaptive Decision Maker*. Cambridge: Cambridge University Press.

Pöyhönen, M.A., Hämäläinen, R.P., & Salo, A.A. (1996). An experiment on the numerical modeling of verbal ratio statements. *Journal of Multi-Criteria Decision Analysis*, in press.

Saaty, T.L. (1977). A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology, 15*, 234-281.

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

Saaty, T.L. (1986). Axiomatic foundation of the Analytic Hierarchy process. *Management Science, 32*, 841-855.

Saaty, T.L. (1990a). An exposition of the AHP in reply to the paper "Remarks on the Analytic Hierarchy Process", *Management Science, 36*, 259-268.

Saaty, T.L. (1990b). Eigenvector and logarithmic least squares. *European Journal of Operations Research, 48*, 156-160.

Saaty, T.L. (1993). What is relative measurement? The ratio scale phantom. *Mathematical Computer Modelling, 17*, 1-12.

Saaty, T. L. & Kearns, K. P. (1985). *Analytical Planning*. New York: Pergamon Press

Saaty, T. L. & Vargas, L. G. (1982). *The logic of priorities: Applications in business, energy, health and transportation*. Boston: Kluwer Nijhoff Publishing.

Saaty, T.L., & Vargas, L.G. (1984). Comparison of eigenvalue, logarithmic least squares and least squares methods in estimating ratios. *Mathematical Modelling, 5*, 309-324.

Saaty, T. L. & Vargas, L. G. (1991). *Prediction, Projection and Forecasting*. Boston: Kluwer Publishing.

Schiffman, S.S., Reynolds, M.L., & Young, F.W. (1981). *Introduction to multidimensional scaling: Theory methods and applications*. New York: Academic Press.

Takeda, E., Cogger, K.O., & Yu, P.L. (1987). Estimating criterion weights using eigenvectors: A comparative study. *European Journal of Operations Research, 39*, 360-369.

Torgerson, W.S. (1958). *Theory and methods of scaling*. New York: Wiley.

Weber, M., & Borcherding, K. (1993). Behavioral influences on weight judgments in multiattribute decision making. *European Journal of Operational Research, 67*, 1-12.

Zahedi, F. (1986a). The analytic hierarchy process – A survey of the method and its applications. *Interfaces, 16*, 96-108.

Zahedi, F. (1986b). A simulation study of estimation methods in the Analytical Hierarchy Process. *Socio-Economic Planning Sciences, 20*, 347-354.