AN EMPIRICAL EVALUATION OF THE SCALE SENSITIVITY IN THE AHP: AN ASSESSMENT OF MANAGERIAL IMPLICATIONS

Hans C.J. Vrolijk and Eelko K.R.E. Huizingh
University of Groningen, Department of Business Administration and Management Sciences
P.O. Box 800, 9700 AV Groningen, The Netherlands
e-mail: vrolijk@eco.rug.nl / huizingh@eco.rug.nl

Abstract: The Analytic Hierarchy Process (AHP) requires decision makers to compare alternatives and criteria pairwise. In the verbal mode the decision maker can for example specify that alternative I is moderately more preferred than alternative II. These verbal comparisons are converted into numbers. Since the conception of the AHP in the late seventies, there has been a debate about the correctness of these conversions. This debate is based on the fact that AHP converts verbal judgments into numbers on a 1 to 9 integer scale. For instance, if the decision maker considers alternative A to be slightly more preferable than alternative B, AHP awards this judgment a score of 2, which means that A is twice as preferable as B. Many researchers find it questionable whether this is what the decision maker meant when indicating that A is slightly more preferable than B. Several studies in probability assessment have shown that different people have different interpretations of verbal phrases (Beyth-Marom, 1982; Timmermans, 1994; Budescu and Wallsten, 1995). Dyer and Forman (1991) suggest that this is not necessarily a problem in the AHP, as long as verbal and numerical judgments are not mixed. The preferential hierarchy of the alternatives will be in accordance with the decision maker's perception, but this is not necessarily the case for the calculated differences in priority between the alternatives.

In recent years a number of alternative scales have been proposed in literature. Ma and Zheng (1991) have proposed the 9/9-to-9/1 scale; Salo and Hämäläinen (1993) have introduced the balanced scale and Lootsma (1992) has proposed a natural scale based on Weber's Law of Just Noticeable Differences. An important characteristic of all these alternative scales is the increasing difference between successive levels, whereas the original 1-9 integer scale is based on constant differences.

In this paper we address the question whether the different scales perform equally well. Some previous research has been done to study the effect of the scale (Schoner and Wedley, 1989; Pöyhönen et al., 1996). However, these studies were aimed at measurement tasks (proportion of colors, size of objects) with only marginal managerial implications. Our experiment involves a task in which the preferences of the participants have been measured.. We compared the four scales to test whether the different scales result in similar weights. Furthermore we investigated the predictive validity of all scales. The predictive validity was established based on two sets of evaluation alternatives. We tested the quality of the scales in predicting the preferences of the decision maker for both evaluation sets. Additional analyses focused on the effect of the scale on the consistency ratio. The consistency ratio is based on the consistency index of the judgment matrix and the consistency index of randomly generated reciprocal matrixes (Saaty, 1980; Vargas, 1982; Forman, 1990). To be able to calculate these consistency ratios for the other scales we repeated the simulations with a sample size of 100,000 to calculate the random consistency index. Three scales and matrices of dimensions of 2 to 9 resulted in 2.4 million random matrices evaluated.

Our analyses are based on the data of a laboratory study with 89 participants. The decision task in the laboratory study involved the selection of an apartment. The respondents made pairwise comparisons of the attribute levels and pairwise comparisons of the attributes by selecting preference phrases (verbal mode). Furthermore they evaluated two sets of evaluation alternatives. Their preferences with respect to the evaluation alternatives were used to test the predictive validity of the scales. Our results show significant differences between the four scales with respect to the predictive validity. Further analyses have shown that the selection of the scale also affects the consistency of the judgments. Our results clearly show that AHP as a decision method can be further improved by seriously paying attention to the selection of the scale. Further research should be aimed at identifying rules prescribing which scale to apply in a certain decision context.