Environment and Planning B: Planning and Design 1995, volume 22, pages 21 - 34

Can decisionmakers express multiattribute preferences using AHP and MUT? An experiment

S-K Lai
Department of Land Economics and Administration, National Chung Hsing University, Taipei, Taiwan ROC
L D Hopkins
Department of Urban and Regional Planning, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
Received 17 August 1993; in revised form 17 March 1994

Abstract. In the research design, iteration between two multiattribute techniques is used to compare their ability to elicit preferences. Each subject used one technique, then a second, and iterated between the two. Previous judgments from each technique were presented as two anchors for each succeeding judgment. The final, converged judgment is an estimate of intended preferences, or at least of constructed preferences. The technique that in its first use yielded preferences closer to the converged preference is therefore the better technique. Three techniques (MUT, AHP, and AHP′) were compared in two experiments. MUT was found to be more effective than AHP′ and not significantly different from AHP. The results also imply that judgments of ratios between atrribute intervals are more difficult to make than judgments of equivalence of intervals.

1 Introduction

The majority of multiattribute decisionmaking techniques elicit from the decisionmaker judgments either on the equivalence of attribute intervals or on the ratio between attribute intervals. The comparison of multiattribute techniques, therefore, provides insights both into the effectiveness of particular techniques and into people's ability to make equivalence judgments compared with ratio judgments. Effectiveness is defined here as the subjects' ability to express the preferences they intend to express. There is, however, no externally observable standard of intended preferences. We introduce a research design that uses a standard internally derived from iterative application of the techniques being compared. This standard of effectiveness can be justified by viewing preferences either as given and stable or as constructed in the process of eliciting them.

This research design requires mathematical transformations of judgments between techniques for two reasons. First, these transformations permit the presentation of the results of judgments made via one technique as anchors for judgments by the other technique. Second, they provide the computation of a converged preference structure as the internally generated standard. These transformations themselves must be justified for the techniques being compared, as explained earlier (Lai and Hopkins, 1989). If the validity of the internally generated standard and the validity of the transformations are acknowledged, this research design can answer the question: which technique more effectively elicits the preferences the decisionmaker intends to express?

We compare three techniques in the two experiments. In the first we compare the analytic hierarchy process (AHP) (Saaty, 1980) with multiattribute utility theory (MUT) (Keeney and Raiffa, 1976). The second compares a modified version of AHP, which we label AHP′, with MUT. AHP′ asks subjects to make judgments as ratios of intervals between performance scores of two alternatives or attributes. The original AHP asks subjects to judge ratios between absolute levels of performance.

We chose these techniques for comparison because they are used in practice and are advocated by different groups of authors.

Multiattribute judgments are only one part of any problem-exploration process. Any results based on this judgment task must also be effective in a larger problem-solving context (Brill et al, 1990).

2 Evaluating multiattribute techniques

The difficulty in evaluating multiattribute decisionmaking techniques is establishing a criterion for comparison. Von Winterfeldt and Edwards (1986) reviewed three validation paradigms that are used to test the validity of such techniques: convergent, predictive, and axiomatic. The convergent approach compares results from one technique to results from a second technique. The second technique is the decisionmaker's holistic judgments of preference among alternatives. The predictive approach compares the results from a technique with the future performance of evaluated objects, such as applicants to graduate schools. The axiomatic approach tests the logical assumptions of techniques, such as preferential independence among attributes.

None of these approaches provides a satisfactory answer to the question of effectiveness. Axiomatic standards test only whether the technique is logically consistent, not whether it is properly understood and applied by a decisionmaker. It thus does not determine whether a decisionmaker is successful in expressing intended preferences. Predictive standards imply that the preference structure is an attempt to predict actual performance, thus testing, for example, how good a decisionmaker is at a priori judging of students or projects. The ability to predict performance based on leading indicators is not equivalent to the ability to properly express one's value preferences. The convergent approach relies on the expectation that different techniques for estimating the same value should be correlated. A technique that correlates well with some central tendency of other techniques is likely to be effective. Most such evaluations, however, compare a technique to holistic judgment. Such comparisons imply that a decisionmaker can make judgments more effectively by relying on holistic judgment without any explicit technique. The fundamental claim of multiattribute techniques is that they are more effective than holistic judgment, which requires them to yield preferences that are *not* convergent with holistic judgment. Holistic judgment has less face validity than other techniques, not more. Evidence (for example, Slovic et al, 1977) suggests that subjects perceive alternatives differently when making decomposed judgments than when making holistic judgments. Different results should be expected and thus the claim that holistic judgments are valid as a standard is unjustified.

Comparisons using the convergent approach (for example, Schoemaker and Ward, 1982) are reviewed by von Winterfeldt and Edwards (1986). The most frequent approach is to have each subject use multiple techniques and compare the results between techniques for each subject. It is difficult to elicit judgments using several techniques from each subject or to justify any one technique as a valid standard. We thus introduce an iterative approach that allows two techniques to be compared with an internally generated standard based on those two techniques. Hopkins (1984) proposed a similar approach for evaluating techniques applied to ill-defined problems. That standard was generated across subjects, as implemented by Brill et al (1990), whereas the standard in these experiments is generated within subjects. An internally generated standard for comparison must confront the question of whether preferences are constructed in the process of expressing them or are known a priori and are stable during the process of elicitation.

If we take the constructivist view, then iterative use of a technique should result in a sequence of adjustments based on increased understanding as the results are displayed. When a person stops iterating, we would expect the final judgments to be that person's best estimate from that technique. In this view, a technique that yields preferences on the *first* iteration that are closer to those on the *final* iteration is either more effective at estimating constructed preferences or is less effective in helping to construct preferences. In the first interpretation, the first estimate is good because it is nearly the same as the last. In the second interpretation, the technique does not succeed because it does not lead us to change our first estimate. If we accept each of these interpretations as equally likely, then we may be able to distinguish between two techniques. Ask subjects to use each technique, then iterate between the two with the results of the immediately preceding judgment from each technique as two anchors for succeeding judgments. The resulting constructed preferences are no longer uniquely influenced by the constructive effectiveness of one technique. Thus, the technique for which the first use yields preferences closer to the final preferences is a better technique.

This conclusion is still open to the argument that the 'better' technique merely has more influence on the constructed preferences. Even if one technique is less useful in constructing preferences, because iterative use does not lead to changes in preference, it is still effective. That is, even if one technique is more effective at helping us to construct our preferences (that is, it leads to greater change), the technique that gave, after a single use, the best estimate of our final constructed preferences is the most effective technique, *if* we use a technique only once. Most multiattribute evaluation applications use a technique only once. We do not use these techniques iteratively in practice. Thus the technique that is most effective when used *once* is the technique of interest.

This research design is also, and more directly, valid if we assume that preferences are known a priori and can be estimated as 'true preferences'. The iteration of the estimation process is then an attempt to remove the error in the judgments through successive approximations. The magnitude of the anchoring effects (for example, Tversky and Kahneman, 1982) should be the same for the two techniques because the results of previous judgments are both presented in the same form. Thus the final estimate of the 'true preference', though it may still contain error, should not be biased toward either technique. A counter argument, that one technique has greater influence on the estimate, implies a constructivist view of preferences and thus the above argument relative to constructivist approaches applies. In the a priori true-preference view a subject who chooses to change a preference structure, after seeing it displayed, is changing it because it does not match the subject's 'true preference', which the subject can recognize when it is presented. Thus the question is, can the subject use the technique to generate that true preference in a single try or not?

Comparison of AHP and MUT through this research design requires that the preferences resulting from judgments in one be transformable into preferences as presented by the other. Each technique generates a scale of relationships among alternatives on each attribute and weights or scaling constants among attributes. Although there is disagreement about whether any transformation is valid (see Belton, 1986; Belton and Gear, 1983; Dyer, 1990; Harker and Vargas, 1990), Kamenetzky (1982) concluded that AHP and MUT are comparable in the structure of value functions, but that the meanings of weights are ambiguous. Belton (1986) concluded that the mathematical meaning of an attribute weight in AHP is the relative contribution of a unit of the MUT value for that attribute. Zahedi (1987), in a

more formal way, found that the two techniques mathematically lead to the same results if weight ratios in pairwise comparison matrices were utility ratios and utilities on each attribute were normalized by a constant scalar. These two techniques should thus yield equivalent results if the weights and values are represented on the same scale. The transformation used between the two techniques is based on an earlier paper (Lai and Hopkins, 1989).

3 Experiments

This research design was the basis for two experiments. One experiment compared AHP with MUT. The other compared MUT with AHP′, a modification of the traditional AHP. AHP′ elicits judgments as pairwise ratios of attribute intervals (Lai, 1990). The aim behind considering AHP′ is to distinguish between the elicitation effectiveness of the pairwise comparison format in AHP compared with the mid-value splitting format in MUT, and the elicitation effectiveness of the importance questions in the traditional AHP compared with interval comparisons elicited in the same format by AHP′. Each subject alternately applies each of the two techniques, iterating until a stopping rule is met. One group of subjects starts with each technique, forming a 2 by 2 (two application orders and two techniques) factorial experiment with repeated measurements.

Our first hypothesis was that MUT would be more effective than AHP. The importance questions of the classical AHP are less precisely related to the resulting mathematical representation of preferences than are the elicitation questions of MUT, which are concretely related to attribute intervals. We also used the original approach to AHP (for example, Saaty, 1982), which first compares attributes to attributes, then alternatives to alternatives within attributes. If this sequence is reversed, improved judgments might result because subjects would have a more concrete sense of the attribute measures being compared. This distinction between the two methods might be checked in other experiments. We would expect this reversal of order to improve the performance of AHP, but we tested the standard AHP process as it is currently used. Our second hypothesis was that AHP′ would be more effective than MUT because the pairwise comparison format is easier for subjects to use, and the elicitation questions have been modified so as to be more concrete representations of the attribute scales.

3.1 Procedure

A hypothetical problem of choosing apartments in a college town was chosen as appropriate for the student subjects. Four apartments were presented in terms of four attributes: walking time to work in minutes, floor area in square feet, rent in US dollars, and walking time to parking place in minutes. The subjects were undergraduate engineering students; they were paid five dollars per hour in ten minute increments. Twenty subjects, ten in each group, completed experiment 1. Sixteen subjects, seven using MUT first and nine using AHP′ first, completed experiment 2. The system was implemented on a Macintosh computer system.

Describing the steps for a single subject will help to explain the procedure. The subject received a sheet showing the performance on each attribute for each apartment and brief instructions on how to use the computer system. A subject in group 1, which used MUT first, was then presented with a screen similar to that shown in figure 1. The screen is divided into three parts. The four diagrams in the upper part of the screen are for the mid-value splitting technique. The three diagrams in the lower part of the screen are for tradeoff judgments. The gray area in the middle part is for displaying information that is relevant to the current evaluation and for

ranking the attribute ranges. The interface guides the subject through the tasks of assessing the value functions for each of the four attributes. Using the mid-value splitting technique as explained by Keeney and Raiffa (1976), the subject splits the range of attribute performance shown on the horizontal axis so that the value of the change from the lowest performance to the split is the same as the value from the split to the highest performance. This process is then repeated for the quarter points of value, shown on the vertical axis, thus the subject is estimating a value function that transforms performance levels on the attribute to value levels. In figure 1, the value functions for walking time to work and flow area have already been estimated. The screen shows the value function for rent being estimated. Note that the elicitation question states the meaning of the judgment being made and that the instruction box tells the subject how to input the judgment—by dragging the curve along the particular mid-value line. As the curve is dragged the statement of meaning is modified so that the subject knows what the system is receiving.

After completing the mid-value splitting technique for four attributes, the subject then assesses the scaling constants for the attributes as shown in figure 2 (see over). Figure 2 presents the screen display for this part of the MUT evaluation procedure. Four numerical cards and four white statement boxes are displayed in the gray area. The subject drags the four numerical cards into the four white, round-cornered boxes. The text in each of the four boxes is the statement of the difference between the worst and the best level of an attribute. The subject ranks the four attribute ranges by considering the worths of changes from the worst to the best levels. The attribute range worth the most is ranked first and the attribute range worth the least is ranked last. After the ranking is determined, the screen is cleared and the order

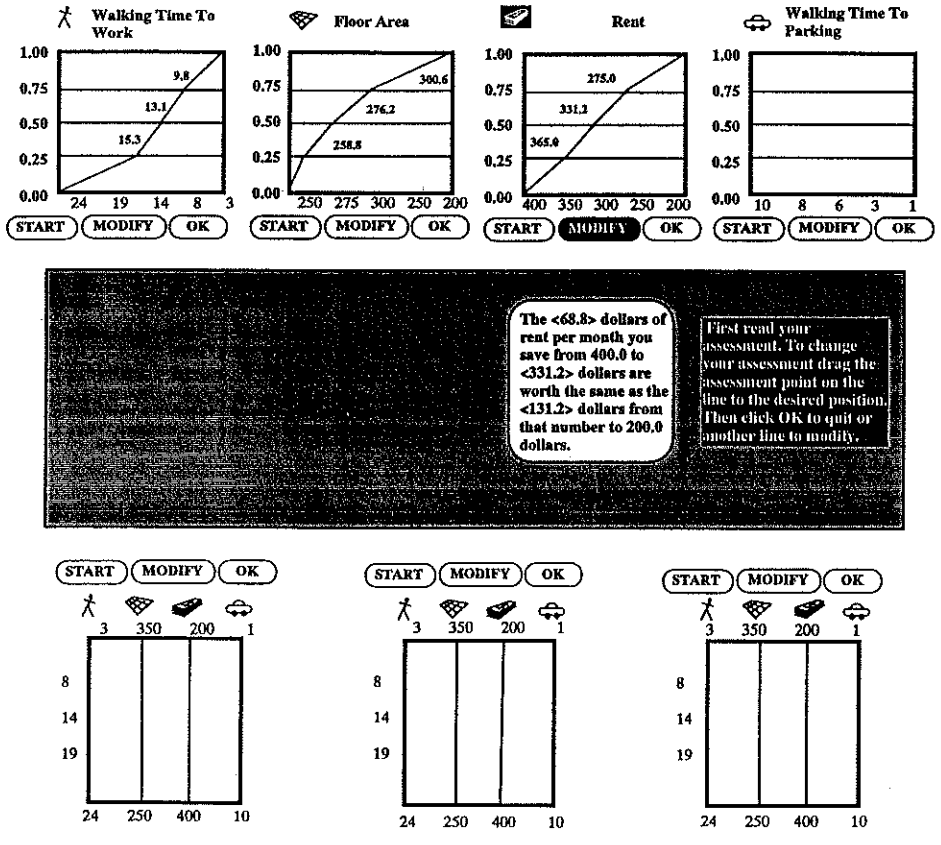

Figure 1. Screen display for the mid-value splitting technique in MUT.

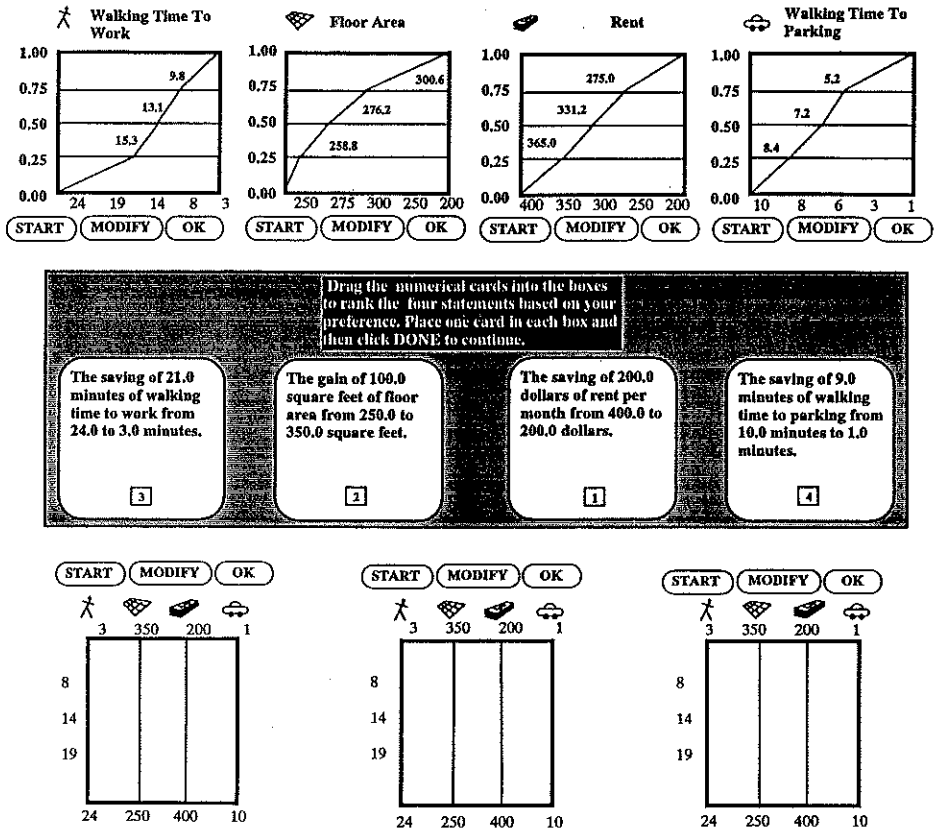

Figure 2. Screen display for attribute assessing scaling constant: step 1.

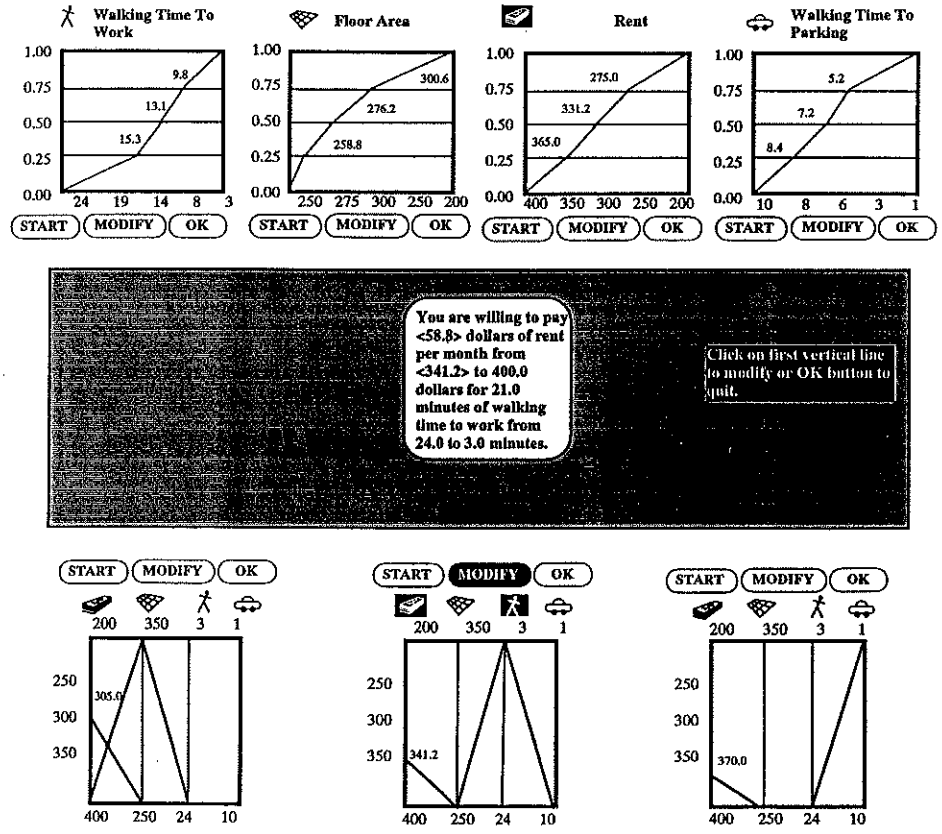

Figure 3. Screen display for attribute assessing scaling constant: step 2.

of the icons above the three diagrams in the lower part of the screen is changed according to the ranking just made.

Figure 3 is the screen display for the third part of the MUT procedure: tradeoff judgments. For each of the three diagrams, the subject compares the first-ranked attribute, which has the largest value range, with one of the other three lower-ranked attributes in an ascending order according to the ranking. For each diagram, the subject adjusts the profile on the vertical line for the first-ranked attribute down to a point such that the change from the worst level to the level of the first-ranked attribute is worth the same as the change from the worst to the best levels of the lower-ranked attribute being compared. Determining the scaling constants completes the use of MUT.

Then the screen for AHP appears as in figure 4. In this experiment, as in the classical AHP, the subject first determines pairwise comparisons among attributes by dragging the bar on a ruler set to the traditional nine-point scale. Pairs can be compared in any order chosen by the subject by clicking on the appropriate cell in the matrix. Then the subject compares alternatives on each attribute as in figure 5 (see over). This sequence completes the firstiteration in the experiment: the application of each technique in the order prescribed for the particular group of subjects. The other group of subjects completes AHP first, then MUT.

Then the subject is presented with a screen like that in figure 6 (see over). This is the same as the MUT screen in the first iteration except that it shows the results of both the first elicitation based on MUT and the first elicitation based on AHP. The subject is asked to make the judgments again. In this experiment, the subjects

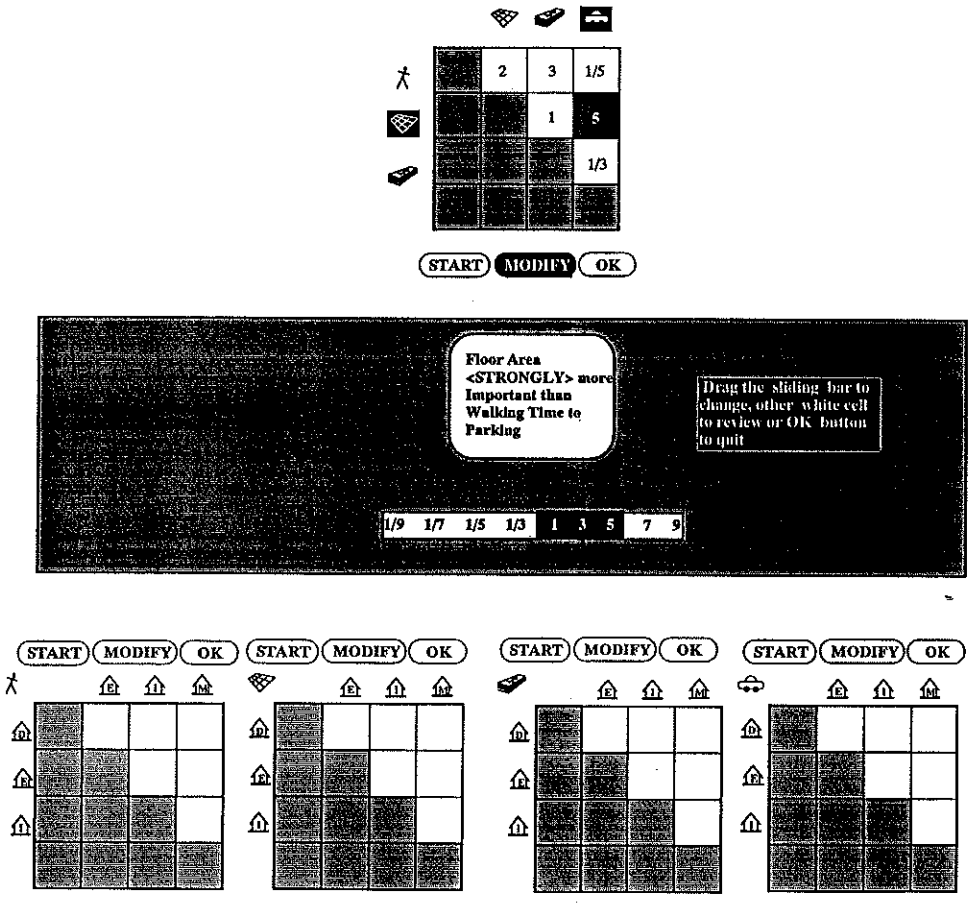

Figure 4. Screen display for a pairwise comparison among attributes in AHP.

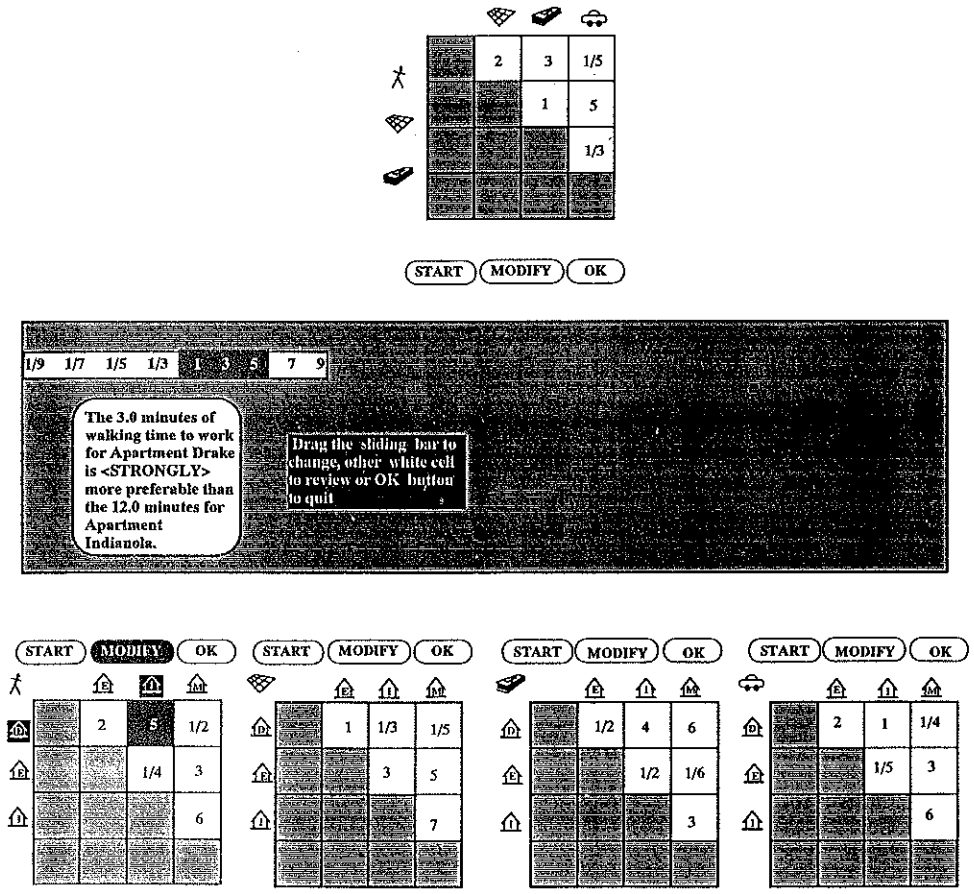

Figure 5. Screen display for pairwise comparisons within attributes in AHP.

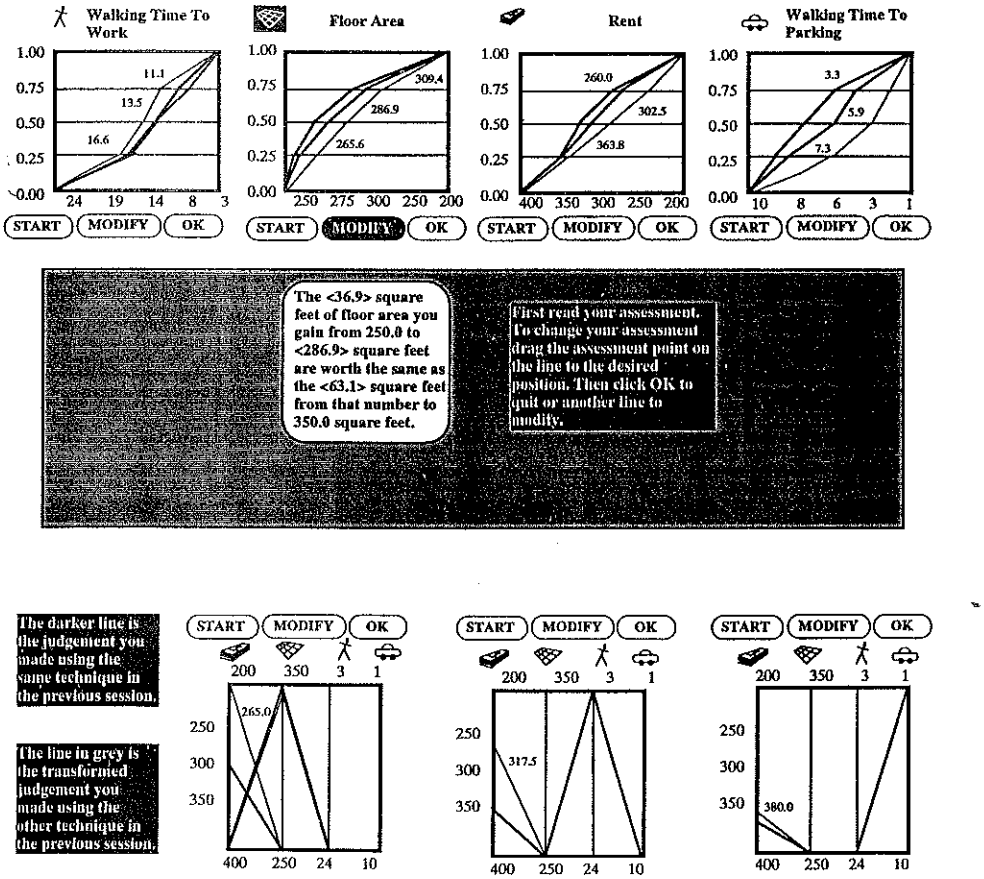

Figure 6. Screen display for the mid-value splitting technique in MUT after iteration 1.

know which reported judgments resulted from which technique. The research design might be improved if they did not know which was which. Then, the subject is presented with a screen like that in figure 7, which shows the AHP elicitations with the results of prior judgments displayed. The subject then makes new judgments using AHP. Again, the other group of subjects follow the same procedure, but started with AHP rather than MUT to control for order effects.

This process is repeated, with the displays showing the results from the most recent previous judgments of each technique, until a stopping rule is reached. Two stopping rules were used: (1) the differences of weights and value functions between the current and previous iterations are less than 10% of the weight and value ranges (or 0.1 for both weights and value functions), or (2) the number of iterations is greater than six and the elapsed time is greater than $1\frac{1}{2}$ hours.

The second experiment, comparing MUT and AHP′, was carried out in the same way. Figure 8 (see over) shows the elicitation question for AHP′. The ranges identified are the ranges for those attributes across all alternatives. Compare the elicitation statement in figure 8, which considers a ratio of intervals, to that in figure 3, which shows the traditional AHP elicitation question.

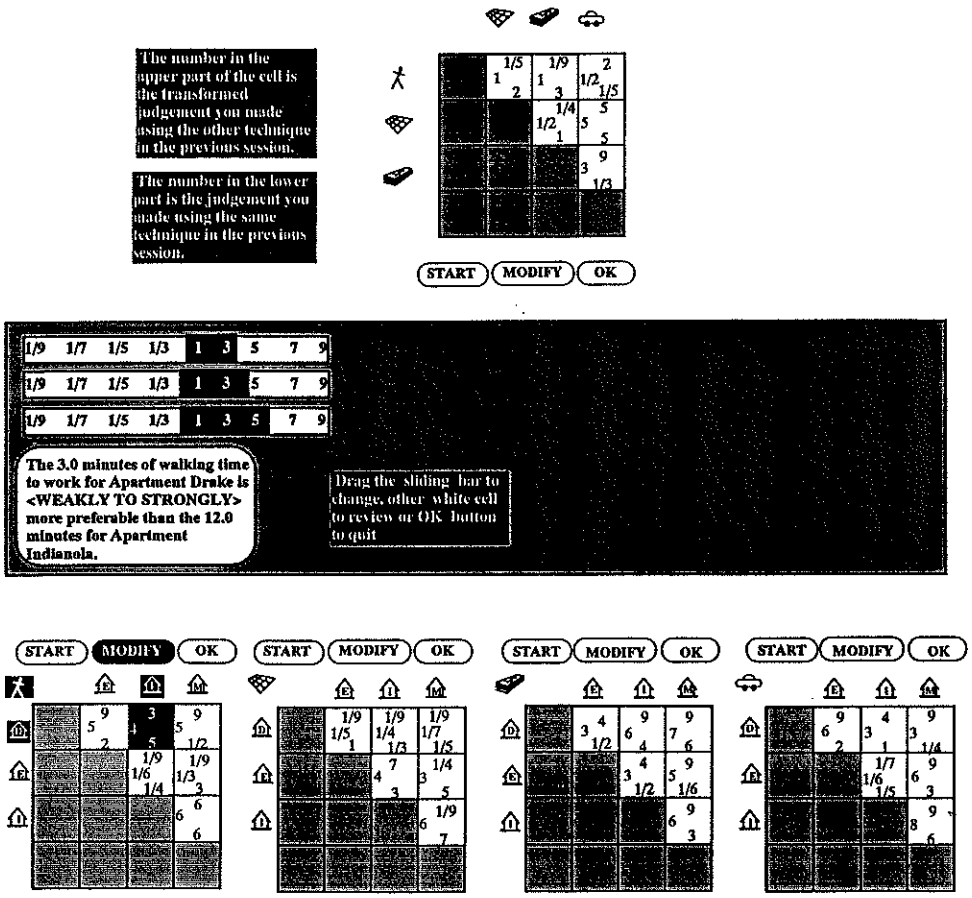

Figure 7. Screen display for pairwise comparisons within attributes in AHP after iteration 1.

3.2 Convergence index

A convergence index was then computed for each subject, comparing the initial judgments by each technique with the final judgments. As shown in equation (1) (see over), the convergence index, C, estimates how well one preference structure converges to the other by comparing preference structures for the initial use of each

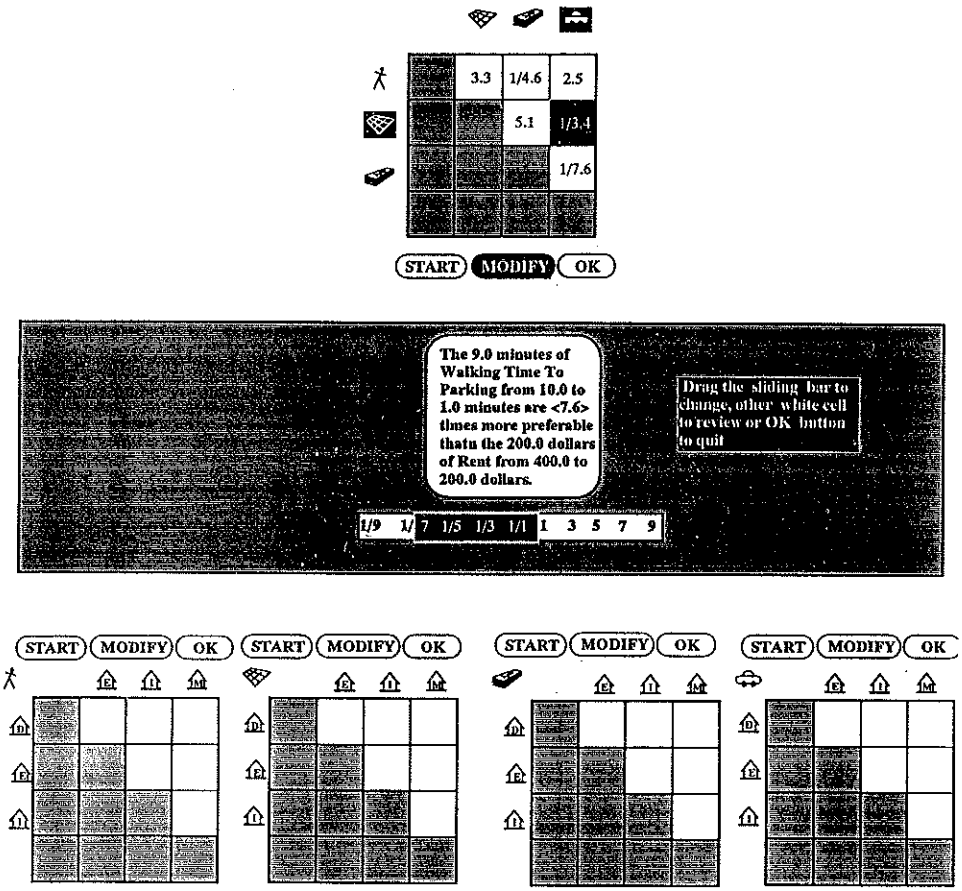

Figure 8. Screen display for a pairwise comparison among attributes in AHP/.

technique with the final preference structure. Equations (2) and (3) are the sums of the absolute differences between two value functions, S_V, at the 100 interval points as defined in the convergence index for comparing value functions, and the sum of the absolute differences between two sets of weights, S_W, for comparing weights. S_V and S_W were designed for stopping the iterative procedure.

$$C = \sum_{i=1}^{m} \sum_{j=1}^{n} |\bar{w}_i^f \bar{v}_{ij}^f - w_i^t v_{ij}^t| \,, \tag{1}$$

$$S_V = \sum_{i=1}^{m} \sum_{j=1}^{n} |\bar{v}_{ij}^f - v_{ij}^t| \,, \tag{2}$$

$$S_W = \sum_{i=1}^{m} |\bar{w}_i^f - w_i^t| \,, \tag{3}$$

where

m is the number of attributes,

n is the number of horizontal intervals for a given value function (arbitrarily set to 100),

w_i^t is the weight of attribute i in iteration t, where the technique is first applied,

$\bar{w}^f$ is the average weight of attribute i between the final iteration f and the preceding one $f-1$,

v_{ij}^t is the value calculated for attribute i at interval j in iteration t, and

$\bar{v}_{ij}^f$ is the average value calculated for attribute i at interval j between the final iteration f and the preceding one $f-1$.

The converged weights and values are the means of the weights and values between the last two iterations, or $\bar{w}_i{}^f = \frac{1}{2}(w_i{}^f + w_i{}^{f-1})$ and $\bar{v}_{ij}^f = \frac{1}{2}(\bar{v}_{ij}^f + v_{ij}^{f-1})$. These averages neutralize the criterion relative to the two techniques being compared by averaging across the last application of each technique, rather than just using the weights and values of whichever technique happened to be used last.

4 Results

The analyses are divided into two parts: (1) comparison of techniques within experiments, and (2) comparison of techniques across experiments. Four quantitative dependent variables are included in the analyses: preference structures, value functions, weights, and times. The measurements for the first three dependent variables are defined in the previous section. Times are measured in minutes. Twenty subjects completed experiment 1 (ten in each group). Sixteen subjects completed experiment 2 (seven in the MUT group and nine in the AHP′ group). The statistical tests on the results are summarized in table 1.

Table 1. Results of tests of significance.

	Preference structure[a]		Value functions[a]		Weights[a]		Times[a]	
	F/t	p/t	F/t	p/t	F/t	p/t	F/t	p/t
MUT vs AHP $F(1, 18)$								
Interaction effect	1.65	>0.10	0.64	>0.10	0.73	>0.10	6.71	<0.02*
Technique effect	0.11	>0.10	0.24	>0.10	0.0	>0.10	10.95	<0.01*
Order effect	0.32	>0.10	0.18	>0.10	0.11	>0.10	2.76	<0.06
MUT vs AHP′ $F(1, 14)$								
Interaction effect	0.99	>0.10	1.04	>0.10	0.39	>0.10	6.31	<0.03*
Technique effect	7.47	<0.02*	10.48	<0.01*	0.89	>0.10	2.88	>0.10
Order effect	0.0	>0.10	0.03	>0.10	0.06	>0.10	0.75	>0.10
Across experiments								
MUT means	−0.4532	>0.10	0.6348	>0.10	−0.429	>0.10	0.87	>0.10
AHP vs AHP′	1.98	<0.06	1.611	>0.10	0.7113	>0.10	2.3252	<0.04*

[a] F for MUT versus AHP and AHP′ and t across experiments.
* Statistical tests were significant at $p/t = 0.05$.

4.1 Experiment 1: comparing MUT and AHP

The only significant effects in experiment 1 are the interaction between the technique and order with respect to times and the technique with respect to times. MUT is not significantly different from AHP in terms of its effectiveness in expressing preference structures, values, or weights. Subjects spent less time in AHP than in MUT.

Based on Dunn's multiple comparison technique, the difference between mean times of MUT (20.33)and AHP (13.12) when MUT is applied first is significantly different ($p < 0.01$). The mean times for MUT when initially applied (20.33) and when applied after AHP (13.82) are also significantly different ($p < 0.05$). The times spent on MUT are thus affected by being preceded by AHP.

4.2 Experiment 2: comparing MUT and AHP′

The only significant effects in experiment 2 are the interaction between the technique and order with respect to times, techniques with respect to preference structures, and techniques with respect to value functions. The hypothesis that AHP′ is more effective in expressing preference structures than MUT is contradicted; MUT is

more effective than AHP′. MUT is also more effective in expressing value functions than AHP′, but there is no significant difference in expressing weights. Subjects spent less time in AHP′ than in MUT and the difference is significant. The reduction in the time for MUT from the initial session to iteration 1 is more dramatic than that for AHP′. Based on Dunn's multiple comparison technique, the mean times for MUT (23.93) and AHP′ (16.35) when MUT is applied first are significantly different ($p < 0.05$). The mean times for MUT when initially applied (23.93) and when applied after AHP′ (17.72) across the two subject groups are also significantly different ($p < 0.05$). The times spent on MUT are thus affected by being preceded by AHP′, which was also true for AHP in experiment 1.

4.3 Comparisons across experiments

To compare the technique performances across experiments, we first compare the MUT convergence indices, or other dependent variables, in experiment 1 with those in experiment 2 to determine whether the convergence indices are consistent in the two experiments. If the deviations of the initial MUT judgments from the final average judgments, that is, the convergence indices for MUT, are not significantly different in the two experiments, then the two subject pools for the two experiments are homogeneous in their understanding and ability to apply techniques in the iterative procedure. Only if this homogeneity among subjects holds can we then compare AHP and AHP′ across experiments. The comparison is based on t tests of differences of means on these dependent variables for techniques as initially applied in the iterative procedure. The comparisons across experiments were all valid, but only the difference in the times was significant.

The mean convergence indices for MUT in experiments 1 and 2 (21.79 and 18.84, respectively) are not significantly different. The subjects in the two experiments can thus be treated as homogeneous with respect to their ability to work through the iterative procedure. We can thus compare the mean convergence index for AHP (18.67) in experiment 1 with that for AHP′ (26.26) in experiment 2. This difference is not significant so AHP and AHP′ are not distinguishable with respect to their effectiveness in expressing preference structures.

The mean values for MUT (51.05) in experiment 2 and experiment 1 (44.01) are not significantly different. Although comparison between AHP and AHP′ is thus valid, the difference across the experiments is not significant. The results for the weights are the same.

Subjects spent slightly more time in MUT (23.93) in experiment 1 than in experiment 1 (20.33), but the difference is not significant to AHP and AHP′ can be compared. Subjects spent more time in AHP′ (17.72) in experiment 2 than in AHP (13.82) in experiment 1 and the difference is significant.

5 Discussion

We hypothesized that MUT would be more effective than AHP because the elicitation questions in MUT are more directly related to the mathematical meanings of the preference structures than are the elicitation questions in AHP. We also hypothesized that AHP′ would be more effective than MUT because the pairwise comparison format is easier to use than the mid-value splitting technique though the meanings of the elicitation questions are equivalent. The results, however, suggest that MUT and AHP are not different in their effectiveness and that MUT is more effective than AHP′. Although the results of one experiment based on a new research design should not be treated as conclusive, we should at least consider possible explanations for these results.

The lack of differences between MUT and AHP and between AHP and AHP′ imply that any effects of differences in the meanings of elicitation questions are insufficient to affect judgments based on a translation to a common meaning. The many difficulties reported in the literature that humans have in expressing consistent judgments may subsume the effects of differences in elicitation. To distinguish these effects we would need a research design that accounts for biases and errors in human judgments, independent of the elicitation questions.

The results suggest that ratio judgments on intervals are more difficult to make than equivalence judgments on intervals. AHP′, which requires such ratio judgments, was less effective than MUT, which uses equivalence judgments. AHP, which is identical to AHP′ in every respect except that it does not require ratio judgments on intervals, was equally as effective as MUT. Thus, our second hypothesis, that AHP′ would be more effective than MUT, may be contradicted because the effect of increased difficulty from ratio judgment of intervals is greater than the effect from the easier to use pairwise judgment format of AHP′. AHP does yield a time advantage over MUT, suggesting that the general AHP format is easier to use, but apparently not sufficiently so to override other effects in this case. Use of MUT as the first technique before AHP or AHP′ does not reduce the time then spent using AHP or AHP′. Use of AHP or AHP′ as the first technique does, however, reduce the time then spent in using MUT. An asymmetric learning effect of some kind apparently occurs. What a subject learns about making these judgments from making them with AHP or AHP′ appears to be transferable to MUT, but not vice versa. It is difficult from these experiments to narrow the possible interpretations. The AHP′ judgments were apparently harder to make than the MUT judgments, but not the AHP judgments. Thus the common learning effect of AHP and AHP′ does not appear to result from the difficulty of making judgments. It might result from characteristics of the format of judgment elicitation, which AHP and AHP′ have in common but in which both differ from MUT.

Our claims for the validity of the research design were supported by the results, though certainly not conclusively. The effectiveness of MUT on all four dependent variables was not significantly different between the two experiments. That is, the behavior of the subjects in MUT was not affected by the differences between AHP and AHP′. If we argue that we are testing for the ability to elicit true preferences, then the implication is that true preferences were not affected by differences in techniques. The possibility of true preferences is thus not contradicted. If we argue that we are testing for the construction of preferences, then the implication is that constructed preferences are independent of combinations of techniques. Either implication is consistent with the validity of the convergence approach to assessing the elicitation of preferences. It is not simply a heavier reliance on MUT that explains the consistency because MUT was equivalently effective to AHP, but more effective than AHP′. These results are tentative at best. The claims about the research design are not contradicted, but they are also not confirmed.

6 Conclusions

MUT was not significantly different from AHP in effectiveness, but MUT was more effective than AHP′. The result that MUT was more effective than AHP′ implies that equivalence judgments on intervals are more effective than ratio judgments on intervals. AHP was better than MUT with respect to time. These first experimental results should prompt further experimental work rather than be taken as conclusive for practice. Additional experiments must also contend with further validation of the research design.

Acknowledgements. We would like to acknowledge the advice and comments of John Flach and Penelope Sanderson in the development of this research. This research was funded in part by a grant from the Research Board of the University of Illinois at Urbana-Champaign. Suggestions from referees helped to improve the presentation of these results.

References

Belton V, 1986, "A comparison of the analytic hierarchy process and a simple multi-attribute value function" *European Journal of Operational Research* **26** 7-21

Belton V, Gear T, 1983, "On a short-coming of Saaty's melthod of analytic hierarchies" *Omega* **11** 228-230

Brill Jr E D, Flach J M, Hopkins L D, Ranjithan S, 1990, "MGA: A decision support system for complex, incompletely defined problems" *IEEE Transaction on Systems, Man, and Cybernetics* **20** 745-757

Dyer J S, 1990, "Remarks on the analytic hierarchy process" *Management Science* **36** 249-258

Harker P T, Vargas L G, 1990, "Reply to 'remarks on the analytic hierarchy process' by J S Dyer" *Management Science* **36** 269-273

Hopkins L D, 1984, "Evaluation of methods for exploring ill-defined problems" *Environment and Planning B: Planning and Design* **11** 339-348

Kamenetzky R D, 1982, "The relationship between the analytic hierarchy process and the additive value function" *Decision Sciences* **13** 702-713

Keeney R L, Raiffa H, 1976 *Decisions with Multiple Objectives* (John Wiley, New York)

Lai S-K, 1990 *A Comparison of Multiattribute Decision Making Techniques Using an Iterative Procedure to Derive a Convergent Criterion* PhD dissertation, Department of Urban and Regional Planning, University of Illinois at Urbana-Champaign, Urbana, IL

Lai S-K, Hopkins L D, 1989, "The meanings of trade-offs in multiattribute evaluation methods: a comparison" *Environment and Planning B: Planning and Design* **16** 155-170

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

Saaty T L, 1982 *Decision Making for Leaders* (RWS Publications, Pittsburgh, PA)

Schoemaker P J H, Waid C C, 1982, "An experimental comparison of different approaches to determining weights in additive utility models" *Management Science* **28** 182-196

Slovic P, Fischhoff B, Lichtenstein S, 1977, "Behavioral decision theory" *Annual Review of Psychology* **28** 1-39

Tversky A, Kahneman D, 1982, "Judgment under uncertainty: heuristics and biases", in *Judgment Under Uncertainty: Heuristics and Biases* Eds D Kahneman, P Slovic, A Tversky (Cambridge University Press, Cambridge)

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

Zahedi F, 1987, "A utility approach to the analytic hierarchy process" *Mathematical Modelling* **9** 387-395