Does AHP help us make a choice? - An experimental evaluation

Alessio Ishizaka and Dieter Balkenborg and Todd R Kaplan

University of Portsmouth, University of Exeter, University of Exeter and University of Haifa

2. August 2010

Does AHP help us make a choice? - An experimental evaluation

Alessio Ishizaka[1*], Dieter Balkenborg[2] and Todd Kaplan[2]
[1] University of Portsmouth, Portsmouth Business School, PO1 3DE Portsmouth, United Kingdom
Email: Alessio.Ishizaka@port.ac.uk
[2] University of Exeter, Department of Economics, EX4 4PU Exeter, United Kingdom
Emails: D.G.Balkenborg@exeter.ac.uk, T.R.Kaplan@exeter.ac.uk

ABSTRACT

In this paper, we use experimental economics methods to test how well Analytic Hierarchy Process (AHP) fares as a choice support system in a real decision problem. AHP provides a ranking that we statistically compare with three additional rankings given by the subjects in the experiment: one at the beginning, one after providing AHP with the necessary pair-wise comparisons and one after learning the ranking provided by AHP. While the rankings vary widely across subjects, we observe that for each individual all four rankings are similar. Hence, subjects are consistent and AHP is, for the most part, able to replicate their rankings. Furthermore, while the rankings are similar, we do find that the AHP ranking helps the decision-makers reformulate their choices by taking into account suggestions made by AHP.

Subject areas: Decision analysis, Multiple Criteria Decision Aid, Analytic Hierarchy Process (AHP), Validation, Experimental Economics

1. Introduction

Companies grow, prosper or fail as a consequence of the decisions taken by their management and stakeholders. Good decision making is therefore vital for the success of enterprises and administrations. Several multiple-criteria decision methods have been developed to help managers in this respect. The Analytic Hierarchy Process (AHP) (Ishizaka and Labib 2009; Saaty 1977; Saaty 1980) is probably the most widely used of these. It has been applied in a diverse range of areas including: Information Systems (Ahn and Choi 2007), Supply Chain Management (Akarte et al. 2001; Sha and Che 2005), Public services (Fukuyama and Weber 2002; Mingers et al. 2007), Health (Lee and Kwak 1999; Li et al. 2008), Strategy (Leung et al. 2005), E-learning (Tavana 2005), Defence (Wheeler 2005), Maritime Ports (Yeo et al. 2009), and Manufacturing (Banuelas and Antony 2006). There is no clear evidence, however, that AHP provides its users with their "best" choice and not an

* Corresponding author

arbitrary one. Perhaps managers want only to claim to use a scientific process for their decisions but would have taken the same decisions without AHP.

The aim of this research is to verify the practicality of AHP using the methods of experimental economics. Experimental economics studies the behaviour of human subjects in real decision problems under controlled laboratory conditions. To give appropriate incentives, subjects are rewarded, based upon their decisions, with an amount of money or goods comparable to what they could gain elsewhere. The use of laboratory experiments as a tool in empirical economic analysis has grown in economics over the last twenty years, culminating in the Economics Nobel Prizes for Daniel Kahnemann and Vernon Smith in 2002 (see the advanced information of the Nobel committee available at http://nobelprize.org/nobel_prizes/economics). The approach has also been successful in other areas, as, for instance, in Accounting (Callaghan et al. 2006), Environmental Sciences (Sturm and Weimann 2006), Social Preferences (Karlan 2005), Supply Chain Management (Croson and Donohue 2002), and Marketing (Beil 1996).

Our experiment attempts to reproduce a real decision problem in the laboratory. A failure for AHP to pass the controlled laboratory test on a basic everyday decision would, in our view, cast serious doubt on the use of AHP on more important problems. The decision problem that we tested is the problem of selecting a box of chocolates among five possibilities. The decision problem is not trivial (at least for some of us) because one has to select among a variety of high quality chocolates from leading brands at different prices. Hence the question of whether AHP can help to improve this basic consumer choice is of interest and a thoroughly negative answer would cast serious doubts on AHP.

We presented the decision problem to twenty-one University of Exeter undergraduates. The task in the experiment involved subjects, endowed with a budget, having to buy exactly one box of chocolate from a list. The subjects kept the money that was in excess of the price. Our procedures also involved asking subjects to give rankings on three different occasions. In addition, AHP was used to generate an additional ranking making a total of four rankings:

A) The first ranking is by participants after tasting the chocolates but before using AHP,
B) The second ranking is also by participants after they had provided the necessary input required by the AHP software,
C) A ranking is calculated by AHP,
D) The last ranking is by participants (the third generated by them) and was completed immediately after the ranking calculated by AHP was revealed to them.

In our experiment, four hypothetic scenarios are plausible:

(a) All three subject rankings (A, B and D) and the AHP ranking (C) are identical. In this case, AHP works correctly but it will be superfluous to use it because no new information is added.
(b) All three subject rankings are identical but are different to the AHP ranking. In this case, AHP is not a useful method as AHP offers advice that the subjects do not agree with.

(c) The subjects' second ranking is different from the AHP ranking and the subjects subsequently adopt the AHP ranking. In this case AHP is a useful method because it helps to reformulate the choices.

(d) The subjects' first and second ranking are different, the subjects' second ranking and final ranking are identical, but different from the AHP ranking. In this case, the process of using AHP is useful but not the result.

While the rankings vary widely from individual to individual, we find, by using a variety of nonparametric statistical tests, that for each individual the ranking generated by AHP is typically in reasonable agreement with the rankings provided by each participant. While we find that AHP detects clear top and least priorities well, we also find that the three rankings given by the subjects tend to be closer to each other than they are to the AHP ranking. We also find that there is evidence that the subjects tend to follow the ranking provided by AHP. By finding at least some support for scenario (c), our experiment provides evidence that AHP is a useful decision tool.

2. Literature review

AHP is a popular Multi-Criteria Decision Method (MCDM), where the key ingredient is that all evaluations are made by pair-wise comparisons on a scale 1 to 9 in a matrix **A** (Saaty 1977; Saaty 1980). In a first step, the decision-maker compares each pair of *n* alternatives in regard to each of *m* criteria. For each criterion *c* local priorities are calculated from the comparison matrix $\mathbf{A_c}$ by the eigenvalue method:

$$\boldsymbol{A}_c \cdot \vec{p}_c = \lambda_c \cdot \vec{p}_c \quad \text{where} \quad \boldsymbol{A}_c \text{ is the comparison matrix,} \qquad (1)$$

$\vec{p}_c$ is the vector of the local priorities,

λ_c is the maximal eigenvalue.

The local priorities yield a cardinal comparison of the various alternatives based upon a single criterion. In a second step, the importance of the criteria is compared pair-wise and weights are calculated again with the eigenvalue method as in (1). The global priorities are then calculated by weighting the local priorities with the corresponding weights for each criterion:

$$\vec{p} = P\vec{w} \quad \text{where} \quad \vec{p} \text{ is the vector of global priorities,} \qquad (2)$$

$\vec{w}$ is the vector of the weights,

P is the matrix of all vectors of local priorities.

The global priorities yield a cardinal comparison of the various alternatives based upon all criteria, in particular, it yields a ranking of the alternatives.

AHP has been extensively used in practice. Several papers have compiled the numerous AHP success stories (Forman and Gass 2001; Golden et al. 1989; Ho 2008; Kumar and Vaidya 2006; Liberatore and Nydick 2008; Omkarprasad and Sushil 2006; Saaty and Forman 1992; Shim 1989; Vargas 1990; Zahedi 1986), but its popularity does not verify that the AHP recommendation is always the best alternative. In fact, AHP has been sharply criticised on several points (Bana e Costa and Vansnick 2008; Barzilai 2001; Belton and Gear 1983; Dodd and Donegan 1995; Donegan et al. 1992; Dyer 1990; Holder 1991; Johnson et al. 1979; Pöyhönen et al. 1997; Salo and Hamalainen 1997; Webber et al. 1996). Many papers have theoretically compared or at least grouped multi-criteria decision methods by similarities (Al-Shemmeri et al., 1997; Guitouni and Martel, 1998; Guitouni et al., 2007; Kornyshova and Salinesi, 2008; Simpson, 1996). These articles stress that choosing a multi-criteria method is a multi-criteria problem. No method has been found to be better on all aspects. Therefore, experiments have been conducted to validate MCDM methods. They can be divided into two groups:

- *Techniques validating outputs calculated by MCDM methods against verifiable objective results*. These experiments assume that the decision is about measurable criteria like the correct estimation of the area of geometric figures or the volume of a type of drink (coffee, tea, whisky, etc) consumed in a country (e.g., Millet 1997; Saaty 2005; Saaty 2006a; Saaty 2006b; Whitaker 2007). These validations give convincing support for AHP; however, they do not address real life decision problems where alternatives are often more difficult to compare because more subjective criteria are involved as, for example, matters of taste or judgements of non-verifiable probabilities.
- *Techniques applied to problems incorporating subjective criteria* (e.g., Brugha 2000; Brugha 2004; Keeney et al. 1990; Korhonen and Topdagi 2003; Hobbs and Meier, 1994; Huizingh and Vrolijk, 1997; Linares 2009). At the end of the decision process, participants were asked by questionnaires or interviews about the process and their satisfaction with the results. For example, Linares (2009) asked 18 students to rank cars with AHP. Thereafter, inconsistencies in the AHP matrices were removed by an automatic algorithm and a new ranking was generated. In the final questionnaire, the majority of the students said that when intransitivities were removed, their preferences were not better represented. In another experiment (Keeney et al. 1990), subjects were asked to provide a direct (informal) ranking of alternatives and then went through a multiattribute utility (MAU) (formal) assessment. After the formal assessment they were encouraged to compare the direct and the MAU evaluations and resolve any inconsistencies. Of all the subjects 80% changed their initial rank order and 67% changed their preferred alternative; most of the changes were in the "direction" of the MAU evaluation. In other words, MAU produced a different ranking from the initial ranking but was helpful to readjust the final ranking. Huizingh

and Vrolijk (1997) designed an experiment where participants were asked to select a room to rent. They observed that participants were more satisfied with the AHP result than with a random selection.

Table 1 summarizes the theoretical and experimental validation techniques. It has been observed that MCDA method selection depends on the problem and the user. To better understand MCDA methods, experiments were used. The experimental validation with subjective results is more convincing than the techniques with verifiable objective results because they deal with problems where AHP is more likely to be applied. In all of these studies, the decision problem was hypothetical and subjects were not rewarded according to their success. Therefore the motivation for their behaviour in the experiment is not clear. Our approach is not only in line with the techniques of the second group (experimental validation with subjective results), but also follows the experimental economics methodology, aiming to give appropriate incentives and make the decisions real and not hypothetical.

Validating method	References	Outcome
Theoretical validation	Simpson, 1996	Comparison of Multi-Attribute Value Theory (MAVT) and ELimination et Choix Traduisant la REalité (ELECTRE). Author concludes that competing tools are not exclusive and should be applied to the same problem for comparison.
	Al-Shemmeri et al., 1997	Listing of a large number of criteria to evaluate methods. Authors conclude that the selection of method may depend on the problem.
	Guitouni and Martel, 1998	Comparison of 29 MCDA methods. Authors conclude that no method is clearly better than the others.
	Kornyshova and Salinesi, 2008	Review of 9 MCDA selection approaches. Authors conclude that there is no perfect method. The selection of a method depends on the characteristics of the problem and the information available.
Experimental validation with verifiable objective results	Millet, 1997; Saaty 2005; Saaty 2006a; Saaty 2006b; Whitaker 2007	Area of geometric figures, volume of drink consumption in a country or distance between cities are evaluated by asking directly an estimate and derived indirectly from pair-wise comparisons (as in AHP). AHP appears to provide more accurate results.

Experimental validation with subjective results	Keeney et al. 1990	21 participants had to select hypothetical long term energy policy. MAU helped them to readjust their initial direct evaluation.
	Hobbs and Meier, 1994	In an hypothetical planning problem, 6 methods are experimentally compared by 12 persons and they concluded that no single MCDA method is unambigously more valid than the others.
	Huizingh and Vrolijk, 1997	180 participants were asked to solve the hypothetical problem of choosing a room to rent. It was observed that AHP give better result than choosing at random.
	Brugha, 2000	Two groups of 10 students were proposed to solve the hypothetical problem of career and car selection. It was observed that participants preferred to use Scoring With Intervals (scoring with respect to a reference) than relative measurement (as in AHP) but relative measurement is preferred when intervals are difficult to identify. The final results calculated by the methods were not compared, probably because it was a fictitious problem.
	Korhonen and Topdagi 2003	4 vegans and 4 non-vegans used AHP to rank meals described on paper. AHP was able to estimate utility and disutility of meals (e.g. vegans dislike meat).
	Brugha 2004	53 students were asked to choose what they would do next year. It was observed that they prefer to use simple methods for screening and more elaborate methods for ranking. The final results calculated by the methods were not analysed, probably because it was a fictitious problem.
	Linares, 2009	18 students rank cars with AHP in a hypothetical problem. It has been observed that when intransitivity is removed, the participants' preferences were not better represented.

Table 1: Summary of validation techniques.

3. Description of the experiment

3.1 Experimental Design

In our laboratory experiment, twenty-one University of Exeter undergraduates are asked to make a straightforward, but not necessarily easy choice in a real decision problem, namely, choosing among five different high quality boxes of chocolates. The five chocolates boxes are:

- Marks & Spencer plc (Chocolate selection), £9.99, 765 g, UK
- Sainsbury's (Belgian chocolate assortment), £7.99, 380 g, Belgian
- Thorntons (Continental white selection), £8.25, 300 g, UK
- Ferrero Rocher (Ferrero Rocher), £4.25, 300 g, Italy
- Lindt (Lindor Cornet), £3.29, 200 g, Switzerland

The full description of the chocolates including ingredients was distributed to the participants.

3.2. Demography of the participants

Twenty-one subjects, eight female and thirteen male, recruited with advertisements among the economics and business students of the University of Exeter, took part in our experiment. Participants were mainly from year three of their undergraduate studies and British. They were in the range of 18-23 years old except for one mature student who was 27 years old (see Table 2). As with most university students, they have a limited work experience; internships are not required in their study. - None of the subjects were aware of AHP before the experiment. Our results did not vary according to the small differences in demographic characteristics in our sample. Only the participants who did not taste the chocolates are outliers (see section 4.4). This missing information is crucial in making the decision because the final purpose of the chocolates is naturally to eat them.

#	Domicile	Age	Gender	Study	Year
1	UK	20	F	BA Business Economics	3
2	UK	20	M	BA Business Economics	3
3	UK	27	M	BA Economics and Politics	3
4	UK	21	F	BA Business Economics	3
5	UK	22	M	BA Business Economics	3
6	Hong Kong	21	M	BA Economics	3
7	UK	21	M	BA Business Economics	3
8	UK	21	M	BA Economics and Politics	3
9	UK	21	M	BA Economics	3
10	Singapore	23	M	BA Business Economics	2
11	UK	21	F	BA Business and Management with Euro Study	4
12	UK	21	F	BA Philosophy and Political Economy	3
13	UK	22	M	BA Economics and Politics	3
14	UK	20	F	BA Business Economics	3
15	UK	22	F	BA Business Studies	2
16	UK	19	M	BA Business Economics	1
17	UK	18	M	BA Economics w Euro Study	1
18	UK	18	M	BA Accounting and Finance	1
19	UK	19	F	BA Economics	1
20	Hong Kong*	22	M	BA Accounting and Finance	2
21	UK	22	F	BA Business and Management with Euro Study	3

* This participant had British nationality

Table 2: Demography of the participants.

3.3 Experimental Procedures

The subjects were given £15 with which they had to buy one box of chocolates at the retail store price, keeping the remainder. This was a highly subjective decision, depending on taste, previous experience of the chocolates, external knowledge of chocolate in general, the value given to some of the ingredients, the money and the quantity.

The experiment lasted slightly less than one hour and was divided into five steps:

i) The subjects received the full description of the chocolates and were then asked to taste them. (Two subjects refused to do so due to dietary restrictions. We hence excluded them from the statistical evaluation. The criterion "taste" has a high importance for the decision: it is central to this experiment and neglecting it could distort the results. The arguments in section 4.2 will give further support for our decision.) After tasting the chocolates, the subjects had to form a first ranking of the chocolates (Ranking A).

ii) In the core part of the experiment we used the implementation of AHP by Expert Choice (http://www.expertchoice.com). The subjects were asked to enter their comparisons for the following problem model:

- **Goal**:
  Buy a box of chocolates.

- **Criteria:**
  - *Value for money*. In order to give a more subjective aspect to this criterion, we chose to use the term value for money instead of price. In fact, it is difficult to isolate the criterion price and convert it on the comparison scale.
  - *Brand reputation*. This reflects previous or exterior knowledge of the chocolates.
  - *Quantity*: Quantity is not necessarily a linear criterion where more is always better. Some subjects might prefer to have only a small or medium sized box because they live alone or think of their waistline.
  - *Ingredients*: They can be an index of quality different from taste. Moreover, the criterion can be very important for subjects with allergies or those with strong ethical or religious beliefs.
  - *Taste*: Surely, the most subjective criterion.
- **Alternatives:**
  The five choices described previously: Marks & Spencer, Sainsbury's, Thorntons, Ferrero Rocher, and Lindt.

This experiment is based on this particular modelling. We are aware that a missing criterion, considered by the subjects but not by AHP, would lead to different rankings. To minimise this risk, we spent considerable time and ran pilot experiments to carefully select the principal criteria. To assess our choices, a questionnaire was handed out at the end of the experiment. The subjects agreed with our selection of criteria. Only one subject mentioned that the packaging could have an influence on the selection of a box of chocolates. While this criterion should be kept in mind for future experiments, the responses support that the chosen set of criteria were sufficient to capture the decision problem. It should, however, have a marginal influence because we cannot observe any data difference between this subject and the others.

Before the subjects get to know the results of AHP, they have to rank the chocolates again (Ranking B). With this step, we aim to see if the use of AHP has an impact on their judgment.

iii) The ranking of AHP is revealed (Ranking C).

iv) The subjects make a final ranking (Ranking D). This is used to test whether the AHP advice has an impact on the subject's priorities.

v) For the payoff only three randomly selected chocolates of the five would be made available. This technique should be a good motivator for subjects to give us honest rankings. In fact, we induce subjects not to overweight their first choice and to evaluate carefully the bottom of the ranking since those chocolates then have a reasonable likelihood of being selected.

Then, one of the first three rankings A, B, or C (the first two by the subjects and the one by AHP) would be selected at random. The subjects would be allocated the available chocolate that was best according to this ranking. If it differed from the best available alternative from the final ranking D, they were given an opportunity to switch as follows. In addition to the price difference that they would have to pay or receive as compensation, the subjects had to propose a transaction fee between £0 and £1 that they were willing to pay. Then, the computer would draw a random transaction fee. If the drawn number was equal to or lower than the proposed transaction fee, the subject was allowed to exchange the chocolate on payment of the transaction fee.

This procedure is called the Becker–DeGroot–Marschak method (BDM), (Becker et al. 1964). In the original experiment, the subject formulates a bid. The bid is compared to a price determined by a random number generator. If the subject's bid is greater than the price, he or she pays the price and receives the item being auctioned. If the subject's bid is lower than the price, he or she pays nothing and receives nothing. It is a widely used experimental method to measure individuals' valuations for consumption goods. We selected this mechanism in order to ensure that subjects had an incentive to provide sincere rankings and in order to test whether subjects may simply be indifferent among several rankings. When the AHP ranking was randomly drawn, we would be able to see how much a subject was willing to pay in order to be able to switch from the alternative selected by AHP to his own final choice.

4. Results

4.1 Introduction

As we will see, the four rankings A, B, C and D tended to be similar for each subject. The first evidence for this is that the BDM procedure for altering the box of chocolates that the subject received was not invoked for any subject. There was only one case where the highest ranked available chocolate was different for the two rankings, but the subject refused to switch at a positive price, indicating indifference. In this case the two rankings differed only by a single swap. Recall that the ranking of A, B, or C for which the box of chocolates that the subjects received was randomly selected. In our experiment, ranking A was selected eight times, ranking B six times and ranking C five times. If the selection of the box of chocolates had been based solely on the AHP ranking (ranking C) still only one clash with D would have occurred, demonstrating that AHP reflects the subject's preferences quite well. The next sections will further examine the similarities of the rankings.

4.2 Criteria

The criteria rankings made by the subjects are concordant with a Kendall concordance coefficient of 0.55, significant at a 5% level. If we leave out the criterion taste, the concordance coefficient is 0.3 and still significant. Concordance is no longer significant if we also leave out the criterion ingredients (coefficient of 0.16). Taste is indisputably the most important criterion (cf. Table

3). It obtains more than twice the weight of the second-most important criterion “value for money”; only five of the nineteen subjects do not select it in the first place but they do select it in the second place.

Criteria	Average weight ± standard deviation
Taste	0.432 ± 0.016
Value for money	0.198 ± 0.015
Quantity	0.141 ± 0.009
Brand reputation	0.140 ± 0.018
Ingredients	0.089 ± 0.007

Table 3: Average weight and standard deviation of the criteria.

If we compare one-by-one the criterion “taste” with the other criteria, we can test the hypothesis that most subjects consider taste more important than another criterion against the zero hypothesis that both criteria are equally often considered more important. In all pair-wise comparisons the zero hypothesis is rejected by a sign test (see line three of Table 4). The same test shows that the criterion “ingredients” is significantly the least important criterion.

Value	Taste	Brand	Taste	Quantity	Taste	Ingredients	Taste
2	17	2	17	1	18	0	19
0.036 %		0.036 %		0.004 %		0.0002 %	

Table 4: Number of times “taste” is more important than another criterion.

Observation 1: *The criterion “taste” is significantly the most important and the criterion “ingredients” the least important.*

4.3 Chocolates

No chocolate was clearly preferred or disliked, as can be seen from the final ranking D (see Table 5). A concordance between the subjects’ rankings does not exist, as the low Kendall coefficient of 0.029 demonstrates. We do not have a niche brand like fat-free or organic chocolates. Our chocolates selection was as homogenous as possible in order to have a very subjective decision varying greatly from person to person. We view this as support for the adequacy of our experimental design. The choice problem does not have an obvious solution and depends on subjective criteria.

Chocolates	Ranked best	Ranked worst
Marks & Spencer plc	2	3
Sainsbury's	6	3
Thorntons	3	5
Ferrero Rocher	3	4
Lindt	5	4

Table 5: Number of times a chocolate box is classified first and last in the final ranking.

Observation 2: *No chocolate is significantly preferred or disliked. They are all considered valid alternatives.*

4.4 Inconsistencies

In order to determine the AHP ranking, the subjects were asked to enter pair-wise comparisons. It is possible to be inconsistent with these comparisons. For instance, one can violate transitivity, i.e., enter data stating that the Lindt chocolate tastes better than Thorntons, the Thorntons tastes better than the Sainsbury's, and the Sainsbury's tastes better than the Lindt. One can also satisfy transitivity, yet be cardinally inconsistent. For instance, one can enter data stating that the Lindt chocolate tastes better than the Thorntons by a factor of 2, the Thorntons tastes better than Sainsbury's by a factor of 2, and Lindt tastes better than Sainsbury's by a factor of 6.

AHP has a means for measuring any inconsistencies by a formula called the consistency ratio (Saaty 1977; Saaty 1980). A ratio of 0 means perfect consistency while any ratio over 0.1 is considered inconsistent. Only 31% of the subjects had a consistency ratio equal or lower than this limit; however, we did not ask the subjects to reconsider the values in the matrices because it would have been a difficult and time-consuming process.

One potential reason for inconsistencies could simply be indifferences among the alternatives and they use AHP more as a lottery system than as a support decision tool. This indifference would also be reflected in a variation of the rankings during the experiment. To examine this, we have studied each subject's relationship between the inconsistencies of the subject's comparisons and the variation of the subject's rankings. One might have expected that subjects who change their rankings often are also more inconsistent in the pair-wise comparisons required by AHP. However, we discovered that this not the case.

To do this we compared the consistency measure with a measure of the distance between two rankings, namely the squared Euclidian distance:

$$\sum_{chocolates} (ranks\ shifted)^2. \qquad (3)$$

Example for the Euclidian distance:

Take the two following rankings, where each number represents a particular alternative:

Ranking A: 2, 1, 5, 3, 4,

Ranking B: 3, 5, 4, 2, 1.

Notice that 2 moves three places between the rankings A and B; item 1 moves three places; item 5 moves one place; item 3 moves three places; and item 4 moves two places. Hence, the Euclidian distance between the two rankings is $32 = 3^2 + 3^2 + 1^2 + 3^2 + 2^2$.

The distance increases quadratically. Thus, when an item moves two places the Euclidean distance is more important than two independent swaps. A linear distance would have given them the same weight.

Example:

Take the two following rankings, where each number represents a particular alternative:

Ranking A: 1, 2, 3, 4, 5,

Ranking B: 3, 2, 1, 4, 5,

Ranking C: 2, 1 3, 5, 4.

The Euclidean distance between ranking A and B is 8 and between ranking A and C is 4. The linear distance between the rankings is 4 in both cases. The linear distance does not distinguish between a double shift and two independent swaps.

We measured the variation across several rankings by adding the Euclidian distances of any two of them. Figure 1 shows that the inconsistency in comparisons has no correlation with the variation across the different rankings A, B and D.

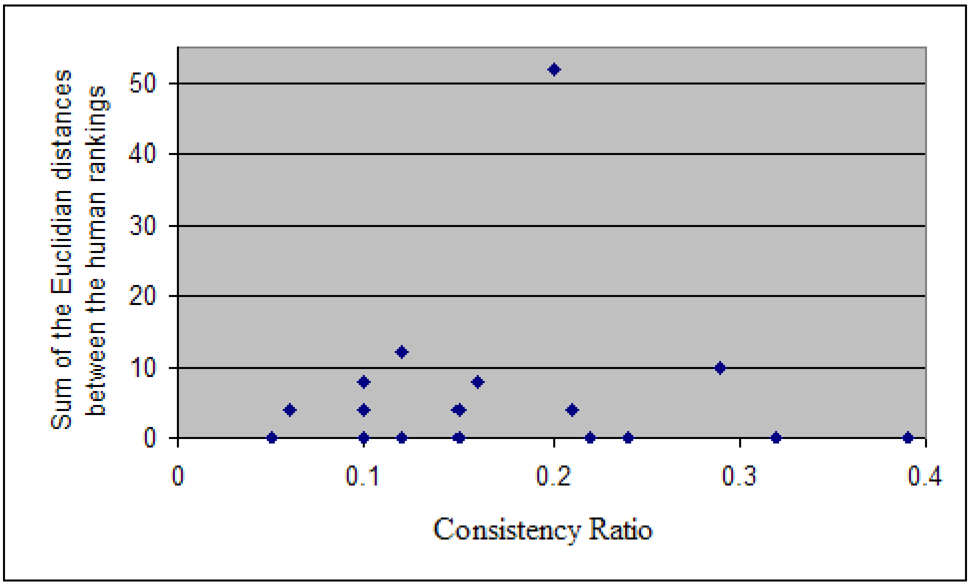

Figure 1: No correlation between the consistency ratio and the variation across the rankings A, B and D.

Figure 2 shows that subjects 16 and 19, having not tasted the chocolates, had far more variation across their rankings. Subject 20, who confessed in the post-experiment questionnaire that he had difficulty differentiating the taste of the chocolates, had the same uncertainty. This indicates that the criterion taste has a high importance for the ranking of the chocolates.

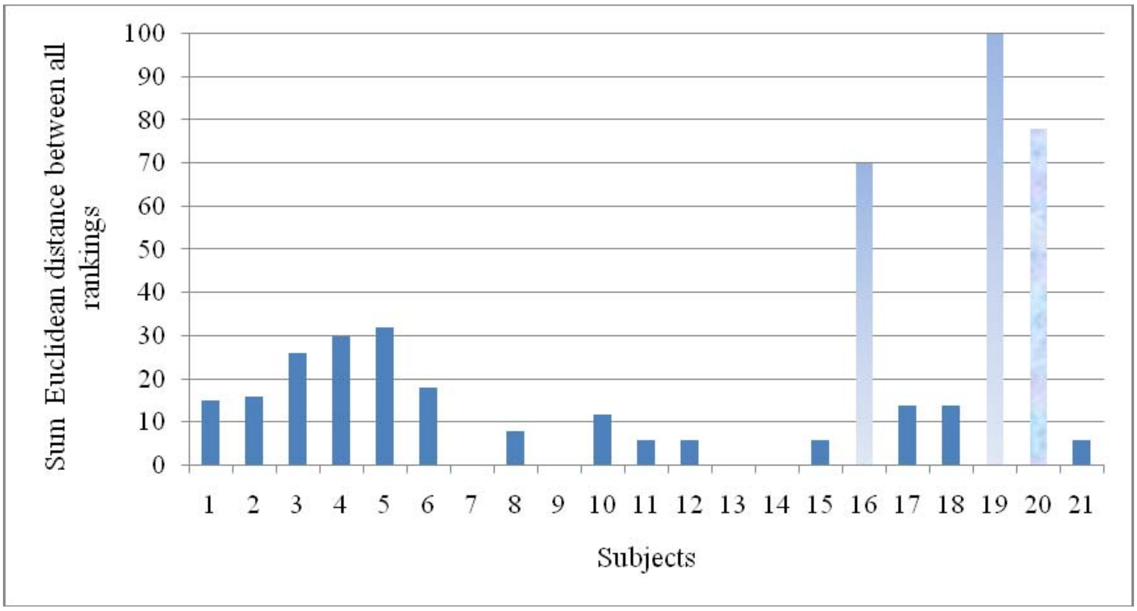

Figure 2: Euclidian distance across rankings of each subject.

Observation 3: *The degree of inconsistency in the pair-wise comparisons has no relation with the variation in the rankings.*

4.5 Closeness of the AHP ranking with the decision-maker's ranking

As a computational experiment, we generated all 120 possible rankings with five alternatives and calculated the Euclidian distances with formula (3) to a fixed ranking (Table 6). The median of the Euclidian distance is between 18 and 20. This means that if all rankings were randomly selected with equal probability, 50% of the rankings would have a Euclidian distance of 18 or less and 50% a Euclidian distance of 20 or more. If we compare the distances between the AHP ranking C and the rankings A, B or D of the subjects (Figure 3), no single one is 20 or higher. Therefore, we can reject by a sign test the hypothesis that ranking C and the other rankings are unrelated. The AHP ranking C is very close to those given by the subjects: 61% of the distances are 0 (same ranking) or 2 (single exchange of two adjacent alternatives).

Euclidian distance	2	4	6	8	10	12	14	16	18
Number of rankings	4	3	6	7	6	4	10	6	10

20	22	24	26	28	30	32	34	36	38	40
6	10	6	10	4	6	7	6	3	4	1

Table 6: Number of rankings with a given Euclidian distance to one particular ranking.

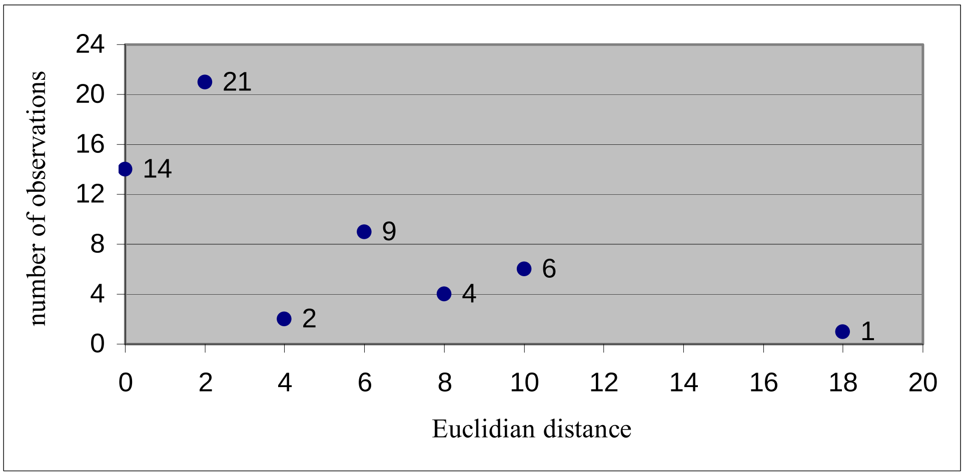

Figure 3: Euclidian distance between the AHP rankings and the subjects' rankings.

Observation 4: The AHP ranking, C, is close to the direct rankings, A, B and D.

4.6 The impact of learning the AHP ranking

The subjects' rankings varied slightly during the experiment. Any variation between the rankings A and B would only be due to a subject learning introspectively by entering comparisons while using the AHP program. On the other hand, any variation between B and D would be due to learning about the information provided by AHP in the form of a ranking. In this section we study whether the advice of AHP was used by the subjects. In order to examine this, we used two methods: the number of changes in the direction of and against the AHP advice and, as before, the Euclidean distance between two rankings.

4.6.1 Changes for or against the AHP advice

For any two rankings of a subject, say, B and D, we look at all cases where an alternative changes position both in the ranking from B to D and from B to C. If the change is in the same direction, the change from B to D is consistent; otherwise it is inconsistent with the AHP advice. For each subject we count whether the majority of such changes is consistent or inconsistent with the AHP advice (Table 7). The zero-hypothesis is that both possibilities are equally likely. If the probability of observation (p-value) is lower than 5%, then we can reject the zero-hypothesis with one sided sign test and if lower than 2.5%, we can rejected with a two sided sign test.

	Majority of changes with AHP advice	Majority of changes against AHP advice	Significance (p-value)
From ranking A to B	2	2	-
From ranking B to D	6	0	**1.6 %**
From ranking A to D	7	0	**0.8 %**

Table 7: How the AHP advice is considered when a subject decides to change his ranking.

The subjects do not yet know of the AHP advice when they form rankings A and B. Thus, it may not be surprising that the number of changes from A to B in the direction of and against the AHP advice is the same. This suggests that the process of filling in the AHP matrices has no visible influence on the direction of the changes. In contrast, subjects clearly follow the AHP ranking, or at least do not act against it, once they see it. Some subjects have written in the feedback questionnaire that AHP reminded them to weight some criteria more strongly and they therefore followed the AHP advice. AHP clearly helps the subjects in their choices.

4.6.2 Euclidean distance between rankings

The prior section underlines the influence of learning the AHP ranking and the non influence of the act of filling in the AHP matrices on the subject's own ranking. Here we show that these observations can also be made by comparing the Euclidian distance between the rankings. We assume that the last ranking D most accurately reflects their true preferences and thus would be more satisfied by receiving the chocolate using that ranking.

We consider five research hypotheses, the zero hypothesis is always that all Euclidian distances are equal:

- *Euclidian distance AC* > *Euclidian distance CD*: the ranking D is nearer to the ranking C than A is to C. It implies that in order to build ranking D, the subjects take into account the advice of AHP and modify their previous direct ranking.
- *Euclidian distance BC* > *Euclidian distance CD*: the ranking D is nearer to the ranking C than B is to C. Again, it implies that in order to build ranking D, the subjects take in account the advice of AHP and modify their previous direct ranking.
- *Euclidian distance AD* > *Euclidian distance BD*: the ranking B is nearer to the ranking D than A is to D. It implies that the process of filling in the AHP matrices moves the subjects closer to the final ranking. This would indicate that the process itself may help the subjects improve their final ranking.

- *Euclidian distance AD < Euclidian distance CD*: the ranking A is nearer to the ranking D than C is to D. It implies that the first ranking is a better representation of the subjects' preferences than the AHP ranking.
- *Euclidian distance BD < Euclidian distance CD*: the ranking B is nearer to the ranking D than C is to D. It implies that the second ranking is a better representation of the subjects' preferences than the AHP ranking.

Three of these research hypotheses are significant with sign test (Table 8). The subjects utilise the advice of AHP for their final decision but the process of filling in the matrices does not move them closer to their final ranking. The ranking after filling in the matrices is significantly more representative of a subject's true preferences than the AHP ranking. It could be therefore be unwise to base the final decision only on the AHP ranking.

Hypothesis	True	Indeterminate	False	Significance (p-value)
AC > CD	7	11	1	3.5 %
BC > CD	6	13	0	1.6 %
AD > BD	3	16	0	-
AD < CD	9	7	3	-
BD < CD	10	6	3	4.6 %

Table 8: Five research hypotheses based on the Euclidean distance.

Observation 5: *Seeing the AHP ranking helps the subjects and they follow its advice. Still, the direct ranking after the process of filling in the matrices is a significantly better representation of the subjects' preferences than the AHP ranking.*

4.7 Divergence of the AHP ranking from the direct rankings

In this section we study the differences amongst the three direct rankings versus differences between the AHP ranking and the three direct rankings. The Euclidian distances amongst the direct rankings AB, BD and AD are summed and compared with the sum of the Euclidian distances between the AHP ranking and the three direct rankings AC, BC and DC. The former number is higher than the latter number for two subjects, equal for five subjects and smaller for twelve subjects. The differences between the AHP ranking and the three direct rankings are hence significantly higher (0.03 %) than the differences amongst the three direct rankings.

Observation 6: *The AHP ranking is the most different among all four rankings.*

4.8 Clear top priority

In this section, we would like to see if AHP detects a clear top priority. A clear priority is defined when an alternative is identically ranked in all the three direct rankings (A, B and D). We then check whether AHP ranks this alternative as highest in agreement with the three other rankings (Table 9).

If we consider all five alternatives, 12 subjects out of 19 have a clear preference and AHP confirms it for 11 subjects. If AHP would randomly rank alternatives, each alternative would have a 20% probability of being ranked first. By a binomial test, we can reject the hypothesis that AHP is randomly ranking the top alternative. In order to see if a clear priority is confirmed in lower ranked alternatives, we remove subsequently 1, 2 and 3 alternatives. In the case of 4 alternatives, we have 95 rankings to verify: 95=19*5 where the 5 is the number of possible single alternatives that could be removed and 19 is the number of subjects. If AHP would randomly rank alternatives, each alternative would now have a 25% probability of being ranked first. By a binomial test, we can reject the hypothesis that AHP is randomly ranking its top alternative among the subsets of alternatives. For the case of three and two alternatives, we can also reject that AHP randomly ranks its top alternative.

	5 alternatives	4 alternatives	3 alternatives	2 alternatives
Total possibilities	19	95	190	190
Clear top priority	12 (63%)	70 (74 %)	157 (83 %)	171 (90 %)
AHP confirmation	11 (58 %)	59 (62 %)	133 (70 %)	153 (81 %)

Table 9: Number of times a clear top priority (in rankings A, B and D) is confirmed by the AHP ranking C.

Observation 7: AHP duplicates very well a clear top priority.

4.9 Clear least priority

In contrast to section 4.8, we examine if a clearly least priority is detected by AHP (Table 10). The number of clear least priorities is higher than the number of clear top priorities (see section 4.8). This observation may be due to the design of the experiment, which is a selection and not an exclusion problem (e.g. to reduce the number of available chocolates boxes from 5 to 3 in a retail shop). This leads subjects to be more concerned to modify the top range of alternatives, which affects their rewards. If AHP would randomly allocate their alternative, an alternative would have a 20% probability of being ranked last. By a binomial test, we can reject the hypothesis that AHP is randomly ranking the last alternative. This rejection also occurs when we remove successively one, two and three alternatives.

	5 alternatives	4 alternatives	3 alternatives	2 alternatives
Total possibilities	19	95	190	190
Clear least priorities	16 (84%)	82 (86 %)	169 (89 %)	171 (90 %)
AHP confirmation	12 (63 %)	66 (69 %)	142 (74 %)	153 (81 %)

Table 10: Number of times a clear least priority (in rankings A, B and D) is confirmed by the AHP ranking C.

Observation 8: AHP duplicates very well a clear least priority.

5. Conclusions

AHP has been both highly praised and strongly criticised. This dichotomy is largely due to the difficulty of testing the AHP method (Yüksel and Dagdeviren 2007) because AHP incorporates quantitative and qualitative criteria. The novelty of our approach is to use experimental economic methods to test AHP on an elementary decision problem with non-measurable decision criteria. More specifically, we used AHP to help subjects in a controlled laboratory environment to make a real, although reduced-scale, decision, namely to buy a box of chocolates. This decision problem shares essential features with several decision problems where AHP has been used, in particular with problems where one criterion is dominant. We observe that AHP provides rankings which are very close to the three subject rankings: 61% of them have the same ranking or agree with it up to a single interchange of two adjacent alternatives.

Differences in the rankings may also arise when important criteria are left out in the AHP evaluation. Apparently this was not the case in our experiment as subjects agreed with the proposed AHP model, as written in the post-experiment questionnaire. An inappropriate weighting of criteria or a biased evaluation of pair-wise comparisons may also be a reason for inconsistencies.

AHP is a useful decision aid method in the sense that it would help the decision-maker to make his decision using its advice without totally overriding the initial, tentative, choice. The reliability of AHP is very high as it detects top and least priorities. These observations suggest that AHP has been probably an adequate support decision tool in many decision problems. .

Using the tools from experimental economics we have shown that AHP is useful in assisting the decision-making process, especially when the problem incorporates a dominant criterion. In future work we plan to apply our experimental approach to other multi-criteria methods and other decision objectives which may not always have a dominant criterion.

ACKNOWLEDGEMENTS

We gratefully acknowledge the financial support of the Freiwillige Akademische Gesellschaft Basel and the University of Exeter. We wish to thank Tim Miller for valuable research assistance and comments. We also thank also the two anonymous reviewers for the valuable feedback and constructive criticism.

References

Ahn B and Choi S (2007). ERP system selection using a simulation-based AHP approach: a case of Korean homeshopping company. *J Oper Res Soc* 59(3):322-330.

Akarte M, Surendra N, Ravi B, and Rangaraj N (2001). Web based casting supplier evaluation using analytical hierarchy process. *J Oper Res Soc* 52(5):511-522.

Al-Shemmeri T, Al-Kloub B and Pearman A (1997) Model choice in multicriteria decision aid, *Eur J Opl Res* 97(3):550-560.

Bana e Costa C, and Vansnick J (2008). A Critical Analysis of the Eigenvalue Method Used to Derive Priorities in AHP. *Eur J Opl Res* 187(3):1422-1428.

Banuelas R, and Antony J (2006). Application of stochastic analytic hierarchy process within a domestic appliance manufacturer. *J Oper Res Soc* 58(1):29-38.

Barzilai J (2001). Notes on the analytic hierarchy process. Proceedings of the NSF Design and Manufacturing Research Conference. Tampa. p 1-6.

Becker GM, De Groot MH, and Marschak J (1964). Measuring utility by a single-response sequential method. *Behav Sci* 9(3):226-232.

Beil R (1996). Laboratory Experimentation in Economic Research: An Introduction to Psychologists and Marketers. *Psychol & Mark* 13(4):331-340.

Belton V, and Gear A (1983). On a Shortcoming of Saaty's Method of Analytical Hierarchies. *OMEGA* 11(3):228-230.

Brugha C (2000). Relative measurement and the power function. *Eur J Opl Res* 121(3):627-640.

Brugha C (2004). Phased multicriteria preference finding. *Eur J Opl Res* 158(2):308-316.

Callaghan C, Gabriel A, and Sainty B (2006). Review and Classification of Experimental Economics Research in Accounting. *J of Acc Lit* 25:59-126.

Croson R, and Donohue K (2002). Experimental Economics and Supply Chain Management. *Interfaces* 32(5):74-82.

Dieckmann A, Dippold K, and Holger D (2009). Compensatory versus noncompensatory models for predicting consumer preferences. *Judgm Decis Mak* 4(3):200-213.

Dodd F, and Donegan H (1995). Comparison of priotization techniques using interhierarchy mappings. *J Oper Res Soc* 46(4):492-498.

Donegan H, Dodd F, and McMaster T (1992). A new approach to AHP decision-making. *The Statician* 41(3):295-302.

Dyer J (1990). Remarks on the Analytic Hierarchy Process. *Mngt Sci* 36(3):249-258.

Forman E, and Gass S (2001). The Analytic Hierarchy Process – An Exposition. *Opns Res* 49(4):469-486.

Fukuyama H, and Weber W (2002). Evaluating public school district performance via DEA gain functions. *J Oper Res Soc* 53(9):992-1003.

Gigerenzer G, and Goldstein D (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychol Rev* 103(4):650-669.

Golden B, Wasil E, and Harker P (1989). The Analytic Hierarchy Process: Applications and Studies. Springer-Verlag: Heidelberg.

Guitoni A, Martel J-M (1998). Tentative guidelines to help choosing an appropriate MCDA method. *Eur J Opl Res* 109(2), 501-521.

Ho W (2008). Integrated analytic hierarchy process and its applications - A literature review. *Eur J Opl Res* 186(1), 211-228.

Hobbs B, and Meier P (1994). Multicriteria methods for resource planning: an experimental comparison. *IEEE Trans. Pow. Sys*, 9(4): 1811-1817.

Holder R (1991). Response to Holder's Comments on the Analytic Hierarchy Process: Response to the Response. *J Oper Res Soc* 42(10):914-918.

Huizingh E, and Vrolijk H (1997). Extending the Applicability of the Analytic Hierarchy Process. *Socio-Econ Plann Sci* 31(1):29-39.

Ishizaka A, and Labib A (2009). Analytic Hierarchy Process and Expert Choice: benefits and limitations. *OR Insight* 22(4):201–220, 2009.

Ishizaka A, Balkenborg D, and Kaplan T (2010). Influence of aggregation and measurement scale on ranking a compromise alternative in AHP, *J Oper Res Soc*, doi:10.1057/jors.2010.23(advance online publication).

Johnson C, Beine W, and Wang T (1979). Right-Left Asymmetry in an Eigenvector Ranking Procedure. *J Math Psychol* 19(1):61-64.

Karapetrovic S, Rosenbloom ES (1999). A quality control approach to consistency paradoxes in AHP. *Eur J Opl Res* **119(3)**: 704-18.

Karlan D (2005). Using Experimental Economics to Measure Social Capital and Predict Financial Decisions. *Am Econ Rev* 95(5):1688-1699.

Keeney R, Von Winterfeldt D, and Eppel T (1990). Eliciting public values for complex policy decisions. *Mngt Sci* 36(9):1011-1030.

Korhonen P, and Topdagi H (2003). Performance of the AHP in comparison of gains and losses. *Mathl Comput Modelling* 37(7-8).

Kornyshova E, and Salinesi C (2007). Selecting MCDM Techniques: State of the Art, Proceedings of the International IEEE Symposium on Computational Intelligence in Multicriteria Deicsion Making, Honolulu, 22-29.

Kumar S, and Vaidya O (2006). Analytic hierarchy process: An overview of applications. *Eur J Opl Res* 169(1):1-29.

Kwiesielewicz M, and van Uden E (2004). Inconsistent and Contradictory Judgements in Pairwise Comparison Method in AHP. *Comput Opns Res 31*, 713-719.

Lee C, and Kwak N (1999). Information resource planning for a health-care system using an AHP-based goal programming method. *J Oper Res Soc* 50(12):1191-1198.

Leung L, Lam K, and Cao D (2005). Implementing the balanced scorecard using the analytic hierarchy process & the analytic network process. *J Oper Res Soc* 57(6):682-691.

Li X, Beullens P, Jones D, and Tamiz M (2008). An integrated queuing and multi-objective bed allocation model with application to a hospital in China. *J Oper Res Soc* 60(3):330-338.

Liberatore M, and Nydick R (2008). The analytic hierarchy process in medical and health care decision making: A literature review. *Eur J Opl Res* 189(1):194-207.

Linares P (2009). Are inconsistent decisions better? An experiment with pairwise comparisons. *Eur J Opl Res* 193(2):492-498.

Millet I (1997). The effectiveness of alternative preference elicitation methods in the Analytic Hierarchy Process. *J Multi-criteria Decis Anal* 6(1), 41-51

Mingers J, Liu W, and Meng W (2007). Using SSM to structure the identification of inputs and outputs in DEA. *J Oper Res Soc* 60(2):168-179.

Omkarprasad V, and Sushil K (2006). Analytic hierarchy process: an overview of applications. *Eur J Opl Res* 169(1):1-29.

Payne J, Bettman J, and Johnson E (1988). Adaptive strategy selection in decision making. *J Exp Psychol Learn Mem Cogn* 14(3):534-552.

Pöyhönen M, Hamalainen R, and Salo A (1997). An Experiment on the Numerical Modelling of Verbal Ratio Statements. *J Multi-criteria Decis Anal* 6(1):1-10.

Saaty T (1977). A scaling method for priorities in hierarchical structures. *J Math Psychol* 15(3):234-281.

Saaty T (1980). The Analytic Hierarchy Process. McGraw-Hill: New York.

Saaty T (2005). Making and validating complex decisions with the AHP/ANP. *J Syst Sci and Syst Eng* 14(1):1-36.

Saaty T (2006a). Rank from Comparisons and from Ratings in the Analytic Hierarchy/Network Processes. *Eur J Opl Res* 168(2):557-570.

Saaty T (2006b). There is no mathematical validity for using fuzzy number crunching in the analytic hierarchy process. *J Syst Sci and Syst Eng* 15(4): 457-464.

Saaty T, and Forman E (1992). The Hierarchon: A Dictionary of Hierarchies. RWS Publications: Pittsburgh.

Salo A, and Hamalainen R (1997). On the Measurement of Preference in the Analytic Hierarchy Process. *J Multi-criteria Decis Anal* 6(6):309-319.

Sha D, and Che Z (2005). Supply chain network design: partner selection and production/distribution planning using a systematic model. *J Oper Res Soc* 57(1):52-62.

Shim J (1989). Bibliography research on the analytic hierarchy process (AHP). *Socio Econ Plan Sci* 23(3):161-167.

Simpson L (1996). Do Decision Makers Know What They Prefer?: MAVT and ELECTRE II. *J Oper Res Soc* 47(7): 919-929.

Sturm B, and Weimann J (2006). Experiments in Environmental Economics and Some Close Relatives. *J Econ Surv* 20(3):419-457.

Tavana M (2005). A priority assessment multi-criteria decision model for human spaceflight mission planning at NASA. *J Oper Res Soc* 57(10):1197-1215.

Vargas L (1990). An overview of the analytic hierarchy process and its applications. *Eur J Opl Res* 48(1):2-8.

Webber S, Apostolou B, and Hassell J (1996). The sensitivity of the analytic hierarchy process to alternative scale and cue presentations. *Eur J Opl Res* 96(2):351-362.

Wheeler S (2005). An analysis of combined arms teaming for the Australian defence force. *J Oper Res Soc* 57(11):1279-1288.

Whitaker R (2007). Validation examples of the Analytic Hierarchy Process and Analytic Network Process. *Math Comput Model* 46(7-8):840-859.

Yeo G, Song D, Dinwoodie J, and Roe M (2009). Weighting the competitiveness factors for container ports under conflicting interests. *J Oper Res Soc* advance online publication:10.1057/jors.2009.1088.

Yüksel I, and Dagdeviren M (2007). Using the analytic network process (ANP) in a SWOT analysis - A case study for a textile firm. *Inf Sci* 177(16):3364-3382.

Zahedi F. (1986). The analytic hierarchy process: a survey of the method and its applications. *Interface* 16(4):96-108.