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

A COMPARISON OF VERBAL AND NUMERICAL JUDGEMENTS IN THE ANALYTIC HIERARCHY PROCESS

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

Abstract: In the Analytic Hierarchy Process (AHP) decision makers make pairwise comparisons of alternatives and criteria. The AHP allows to make these pairwise comparisons verbally or numerically. Although verbal statements are intuitively attractive for preference elicitation, there is overwhelming evidence that people have very different numerical interpretations of the same verbal expressions. This study explores the consequences of these differences for the quality of the AHP analysis. The results of the laboratory study with 180 participants confirm that the 1-to-9 conversion table as is often used in the AHP tends to overestimate preferences. Concerning the outcome of the AHP analysis the numerical mode shows slightly better results (not significant). Given the preference of many people for the verbal mode we conclude that if accuracy is not of the highest importance, the ease and comfort of verbal expressions may be worth the small loss in decision quality.

Introduction

In the Analytic Hierarchy Process (Saaty, 1977, 1980) decision makers express on a 1-9 scale the extent to which they prefer one element (alternative or criterion) compared to the other. Based on these comparisons, the Analytic Hierarchy Process (AHP) computes the importances of criteria, the weights of criterion levels, and finally the preferences for alternatives. The AHP allows decision makers to make the pairwise comparisons verbally or numerically. In verbal comparisons decision makers select one phrase out of a list of nine phrases that best represents their opinion (for example 'moderately preferred' or 'extremely preferred'). Numerical comparisons measure directly the extent to which a decision maker prefers alternative A to B (for example three times).

From a theoretical point of view numerical judgements have a number of distinct advantages compared to verbal judgements. Numerical judgements are more precise, they permit communication to be less ambiguous and they can be used in calculations (Hamm, 1991). Yet, people frequently prefer to use verbal rather than numerical expressions. Verbal statements are intuitively attractive for preference elicitation. Many people resist expressing their opinions numerically, preferring instead to use nonnumerical terms (Budescu and Wallsten, 1985). People claim that they think in words, not numbers, and therefore better understand the meaning of words than of numbers. Zimmer (1983), as cited in Timmermans (1994), has argued that it is more natural to think and talk about uncertainty in verbal terms, and that people are more skilled in using the rules of language than in using the rules of probability. Nonnumerical phrases are supposed to be more useful to express the actual vagueness of an opinion than numbers (Budescu and Wallsten, 1985; see also Zwick, 1987). Timmermans (1994), for example, who studied decision making by physicians, states that physicians are reluctant to use numbers because numbers might suggest more precision than warranted by the available information.

Although the AHP allows decision makers to express preferences verbally or numerically, comparisons expressed in numbers are required to compute the importances or weights. Therefore all verbal judgements are converted into numbers. For example, if a decision maker indicates to prefer alternative A moderately to alternative B, the AHP translates this verbal statement into the numerical score 3. This means that the AHP assumes that the decision maker prefers alternative A three times as much as alternative B. It is doubtful whether the regular use of the phrase 'moderately preferred' refers to this value. Besides, several studies have shown that verbal expressions have different numerical meanings to different people (see section 1 for an extensive discussion). The ability to use verbal expressions is one of the attractive features of the AHP from the practitioner's point of view, but how does it affect the outcome of the AHP analysis? In this paper we will explore the consequences of the conversion process as proposed by Saaty (1980). The central research question is: does the mode of making pairwise comparisons, verbally or numerically, influence the quality of the AHP analysis?

In Section 1 we will review studies that have compared verbal and numerical judgements. Because most studies concentrated on the interpretation of probability phrases, we will also discuss the consequences of these findings for the AHP. The mode of expressing judgements can affect the quality of the AHP analysis in several ways, for example it can affect which alternative will be proposed or the ranking of the alternatives. For each of these aspects an hypothesis has been formulated in section 2. We have tested the hypotheses in a laboratory study with 180 participants. The design of this experiment will be briefly described in section 3. Section 4 will provide an overview of the results. The implications of our findings are discussed in the final section.

Numerical versus verbal judgements

In the AHP preference phrases as 'moderately preferred', 'strongly preferred', and 'extremely preferred' are converted into numbers. The quality of this conversion is the topic of our paper. In this section we will provide an overview of research in a related field, concerning probability phrases, and how these results relate to the AHP. The numerical interpretation of probability phrases has been a fruitful area of research in the past three decades, resulting in a rich literature. In many experiments researchers have tried to find out how people interpret terms as 'likely', 'probable', and 'almost certain'. Brun and Feigen (1988) refer to the titles of three studies which serve as indicators of the intruiging findings of these experiments: 'How often is often? (Hakel, 1968), 'How probable is probable?' (Beyth-Marom, 1982), and 'Sometimes frequently means seldom' (Pepper and Prytulak, 1974). There is overwhelming evidence that people considerably disagree on the numerical interpretation of most verbal expressions. The numerical values subjects assign to the same verbal expressions display large ranges (see also Simpson, 1944, 1963; Stone and Johnson, 1959; Lichtenstein and Newman, 1967; Budescu and Wallsten, 1985; Clarke et al., 1992). Verbal expressions may be inadequate for effective communication because the sender of information can have a different interpretation than the receiver. In a recent study, Timmermans (1994: 148) reported that the interpretations of the term 'very likely' ranged from 30% to 99%. In the AHP the same 1-to-9 conversion table is used for all decision makers regardless how they interpret the phrases. To avoid the interpretation problem decision makers should be aware of the contents of the 1-to-9 conversion table. In making judgements, they should realize that, for example, 'moderately preferred' implies the numerical score of three. However, this would diminish the attractiveness of the verbal scale considerably. Decision makers might feel forced to use a numerical scale labeled with verbal expressions instead of being able to use a verbal scale with their own interpretations.

Because many studies consistently found a high between-subject variability of interpretations, research was directed towards the factors influencing the variability. One potential factor is domain experience. Three studies in the area of medical decision making report conflicting results. Nakao and Axelrod (1983) found more consensus among physicians than among laypeople for the numerical meaning of verbal terms. However, Kong et al. (1986) and Timmermans (1994) found no effect of domain experience. The AHP is used by decision makers with varying experience. Due to the inclusive results of the referred studies, there is no reason to expect that the verbal or numerical mode of the AHP is more suitable for one group of decision makers, e.g. more experienced decision makers or less experienced decision makers.

A second factor supposed to influence the degree of variability in the interpretations is the order in which the verbal expressions are presented. Hamm (1991) found a higher variability in the numerical values assigned to verbal phrases when these phrases are presented in a random order. Clarke et al. (1992) report similar results. Hamm (1991) also found that subjects faced with randomly ordered lists of phrases are more likely to select phrases in the second half of the list. In the AHP the verbal phrases are shown in an ordered list, therefore less variability in the numerical interpretations is expected. The second effect, selecting phrases in the second half of the list, is also unlikely in the AHP. Besides the use of an ordered list, the number of phrases in the AHP is nine (five different levels and four intermediate levels), which is within the boundaries of the magical number seven plus or minus two (Miller, 1956). Hamm (1991) used in his study a list of 19 phrases.

A third factor that might influence the variability of the interpretations of verbal expressions is context. Does it make any difference whether a verbal expression is presented as part of a larger text? Literature provides conflicting clues. According to Mapes (1979) and Bryant and Norman (1983) the numerical interpretations of verbal terms seem to depend on context. Surprisingly, Beyth-Marom (1982) found that judgements were more variable when an expression appeared in context than when it was judged out of context. This finding was replicated in two of the three experiments of Brun and Teigen (1988). However, Bass et al. (1974) and Timmermans (1994) found no effect of context on the variability of interpretations. There seems to be no straightforward relationship between variability of interpretations and context, maybe because, as Budescu and Wallsten (1995) note, context is not a well-defined unidimensional concept. According to Brun and Teigen (1988: 402) the interpretations of phrases are biased by individual differences in opinion. They raise the question in *which* context variability is increased rather than decreased. Brun and Teigen conclude that variability of interpretations can be supposed to increase in contexts involving controversial topics. In their extensive review of literature, Budescu and Wallsten (1995) distinguish six situational variables that have been shown to affect the meanings of probability phrases: base rates, the number of possible alternatives, outcome severity, outcome valence (whether the outcome is positively or negatively valued), characteristics of available uncertainty vocabulary, and one's role in a dialogue. In the AHP the context in which verbal expressions are presented is constant and noncontroversial. The phrases are presented to the decision maker in a standard context always containing two elements (two alternatives or two criteria). This means that most of the situational variables mentioned above are expected not to influence the outcomes of the AHP. However, the design of an AHP analysis might not account for the effect of outcome severity, particularly important in medical decision-making, and outcome valence. Both Mullet and Rivet (1991) and Cohen and Wallsten (1992) found that the numerical interpretations of verbal expressions were higher when associated with a positive than with a negative outcome.

People assign highly different values to the same phrases. The between-subject variability of interpretations is high, but how about the within-subject variability? In the AHP the conversion table is not used once only, but for each set of pairwise comparisons of (sub)criteria or alternatives. Stability of the interpretations is therefore an important assumption. Both Johnson (1973) and Beyth-Maron (1982) found that subjects were relatively consistent in their assignment of numbers to phrases. Also, Budescu and Wallsten (1985) report that individuals have a relatively stable rank ordering of phrases over time.

Based on this discussion we conclude that between-subject variability is the single most important factor influencing the numerical interpretation of preference phrases. Two factors that correlate with the variability are order and context. Order can be left out because the AHP uses an ordered list of phrases. Important aspects of context include outcome severity, outcome valence and whether the decision (or parts of it) is controversial. A decision is controversial if it can have political consequences for the decision maker or can affect his or her image or status with the organization. To avoid these possible effects, we have selected a noncontroversial decision for our experiment.

Hypotheses

The goal of the AHP is to improve the quality of the decision making process. Thus it is in itself not important if one decision maker interprets for example 'moderately preferred' as the numerical score 3 while another decision maker interprets the same phrase as score 4. Crucial to the AHP is: do differences in interpretations influence the quality of the AHP analysis? Does the AHP only provide reliable results for decision makers who interpret the nine phrases in the same way as the 1-9 conversion table does? According to Pöyhönen et al. (1996) the AHP leads to erroneous results if a decision maker does not interpret the verbal statements as weights ratios from one to nine. However, it might be possible that, although each translation of a verbal judgement into a number contains an error, the errors average out. A possible explanation is that similar conversion errors are made in each set of comparisons. Then the final scores of the AHP analysis could still be accurate, although less than perfect. This idea is reflected in the warning of Dyer and Forman (1991: 90) that verbal judgements may be used for one set of judgements and numerical judgements for another, but that one should not mix the verbal and numerical modes for any set of judgements.

The AHP offers both modes, the verbal and the numerical mode. Therefore our basic hypothesis is that the mode itself does not influence the quality of the AHP analysis. We have investigated this hypothesis in several ways. The first category of analyses concentrate on the outcome of the AHP. We have investigated the choice, the ranking, and the preference scores as predicted by the AHP. Second, the intermediate results of the AHP analysis are investigated. Hypotheses have been formulated regarding the importances of criteria, the range of the weights of criterion levels, and the degree of linearity of the utility curves.

The first set of hypotheses refer to the outcome of the AHP analysis. The AHP analysis ends with an advice of which alternative the decision maker should select, this is the alternative with the highest preference score. The numerical and the verbal mode should be equally able to predict which alternative the decision maker will select from a given set of alternatives (Hauser and Koppelman, 1979; Elrod et al., 1992).

Hypothesis 1: The numerical and the verbal mode are equally able to predict which alternative a decision maker will choose from a set of alternatives.

The AHP analysis produces a ranking of the alternatives considered. If both modes are able to provide a good reflection of the preference structure of the decision maker, they both should be able to predict the ranking within a set of alternatives (Schoemaker and Waid, 1982; Tscheulin, 1991).

Hypothesis 2: The numerical and the verbal mode are equally able to predict the ranking of a set of alternatives.

Hypothesis 2 tests the ranking of the alternatives and refers to the ordinal characteristics of the preference data only. A more stringent test consists of the comparison of the actual and predicted preferences scores (Schoemaker and Waid, 1982; Akaah and Korgaonkar, 1983). Both the numerical and the verbal mode should be able to predict the scores assigned to alternatives by the decision maker.

Hypothesis 3: The numerical and the verbal mode are equally able to predict the preferences (scores) for a set of alternative.

The remaining hypotheses refer to the intermediate results of the AHP analysis. These results consist of the importances of the criteria and the weights of the criterion levels. Pairwise comparison of all criteria enables the AHP to compute the importance of each criterion. These importances are not only necessary to compute the preferences for alternatives, but are also used during the sensitivity analysis at the end of an AHP session. Sensitivity analysis, which can be easily performed by means of an AHP software package (Buede, 1992), shows the impact of changes in the importances on the preferences for alternatives. The smaller the impact of these changes, the more confident the decision maker is expected to be in the outcome of the AHP analysis. Given the significant role of the importances, the numerical and the verbal mode should produce similar importances.

Hypothesis 4: The numerical and the verbal mode are equally able to predict the importances of the criteria.

In the introduction of this paper we provided the example of a decision maker who prefers alternative A moderately to alternative B. The AHP interprets this verbal statement as the numerical score 3, implying that the decision maker prefers alternative A three times as much as alternative B. Given the meaning of the word 'moderately' in the regular use of language, the score 3 is probably an overestimation of the difference as perceived by the decision maker. The same applies to the other verbal judgements in the AHP. Pöyhönen et al. (1996) found that the 1-to-9 scale overestimates the ratios assigned to verbal expressions. Also alternatives to the 1-to-9 integer scale contain smaller numbers for the values in the range between 1 to 9 ((Ma and Zheng, 1991; Salo and Hämäläinen, 1993). Thus, in comparison to the numerical mode, the verbal mode is expected to predict larger differences between criterion levels. This implies a larger range between the weights of the most preferred criterion level and the least preferred criterion level.

Hypothesis 5: The verbal mode predicts a larger range of the weights of the criterion levels than the numerical mode.

The AHP enables decision makers to express non-linear utility of quantitive factors (like price or distance). Many models in decision theory assume linear utility, but there has also been substantive research towards non-linear decision models (see for example, Karmarker, 1978; Fishburn, 1988). Often it is assumed that non-linear utility curves are more realistic. When in the AHP the weights of the criterion levels are considered as single points on the utility curve, both the numerical and verbal mode can result in a non-linear utility curve. According to Dyer and Forman (1991: 123) humans have a tendency to overlook the non-linearities in their utility when using numerical comparisons. Therefore, they advocate to use verbal judgements for numerical data. Based on this line of reasoning, it is more likely to find linear weights curves for criteria with an underlying numerical continuum in case of numerical judgements.

Hypothesis 6: The numerical mode predicts weights curves for quantitive criteria with a higher degree of linearity than the verbal mode.

Research design

The hypotheses were tested in a laboratory study with 180 participants. The participants were students of the University of Groningen. The decision task they had to complete was the selection of a room to rent. We incorporated five criteria, rent, area, location, type of house, and facilities. The first two criteria are quantitative and have four levels. The other three criteria are qualitative and have three levels (Table 1). The levels were selected based on a survey to determine the acceptable levels. (See Huizingh and Vrolijk (1996) for a more extensive description of the research design.)

Table 1: The selected criteria and the levels of these criteria.

Criterion	Level 1	Level 2	Level 3	Level 4
Rent	Dfl 450	Dfl 375	Dfl 300	Dfl 225
Area	10 m2	15 m2	20 m2	25 m2
Location	centre	in between	near the university	
Type	small student house (3 students)	large student house (6 students)	student flats	
Facilities	garden	balcony	none	

Each participant completed three tasks relevant for this study (the study is part of a larger research project). These three tasks were (1) making the AHP pairwise comparisons, (2) comparing a small set of alternatives, and (3) rating a large set of alternatives. To prevent order effects in which the completion of one task influences the results of subsequent tasks, an activity that had nothing to do with the research project was inserted between each decision-making task (a quiz concerning sports and culture in the City of Groningen).

1. AHP task. We used a variant of the absolute measurement approach (Saaty, 1986). In this variant the criterion levels were described by means of absolute intensities (Huizingh and Vrolijk, 1996) instead of relative intensities (as low, medium, and high). To complete the AHP task the participants first compared the levels of each criterion pairwise and then similar judgements were made for the criteria. The participants were randomly assigned to the numerical or the verbal mode of the AHP. Half of the participants made verbal judgements, the others used numbers. A similar research design was used in Study 1 of Hamm (1991). We applied the same conversion table as is implemented in the software package Expert Choice (Table 2).

Table 2: The conversion table to translate verbal preferences into numbers.

Verbal Judgement	Numerical Judgement
Equally preferred	1
Equally to moderately	2
Moderately preferred	3
Moderately to strongly	4
Strongly preferred	5
Strongly to very strongly	6
Very strongly preferred	7
Very strongly to extremely	8
Extremely preferred	9

. By means of the other two tasks preference scores of alternatives were collected which could be used to evaluate the AHP predictions.

2. *Comparing task.* The participants divided 100 points among four alternatives. This task is similar to decision tasks in practice and AHP applications as reported in literature. In these tasks decision makers simultaneously compare a limited number of alternatives.
3. *Rating task.* The participants rated 25 rooms on a 0 - 100 scale. This task is relevant because the AHP and other Multiple Criteria Decision-Making methods assume that decision makers have an implicit decision model. This model enables decision makers to judge or score a number of alternatives independent of each other.

We will refer to these two tasks as the evaluation tasks. The room descriptions for both evaluation tasks were generated by the orthoplan procedure in SPSS (SPSS, 1990). One of the authors developed a computer program to support the data collection. 91 Participants completed the numerical mode of the AHP, and 89 the verbal mode. The data of the AHP task were used to generate the predictions for the alternatives in both evaluation tasks (solution method was the right eigenvector). The hypotheses were tested with these predictions.

Results

In the next sections we will discuss the results of the laboratory study with respect to the hypotheses.

1. Choice

Table 3 shows the number (and percentage) of cases in which the numerical and verbal mode of the AHP predict the same choice as the decision maker made in the evaluation task. The number of correct predictions is not an integer value because of ties. If two alternatives were preferred most in an evaluation task and AHP computed the highest preference for one of them, the value 0.5 was assigned, instead of 1 (Leigh et al., 1984). Consequently table 3 shows a conservative estimate of the predictive ability of both methods. For the rating task both methods show almost equal results (49% and 51% correct predictions). For the comparing task the difference between both modes is larger (53% versus 47%), but still not significant. This means that both evaluation tasks confirm the hypothesis that the numerical and the verbal mode of the AHP are equally able to predict which alternative a decision maker will choose from a set of alternatives.

Table 3: Number and percentage of correct choices predicted by the numerical mode (n = 91) and the verbal mode (n = 89) (hypothesis 1; binominal test).

Choice	Correct predictions		Differences	
	N	%	%	Z-score (p-value)
Rating task				
- Numerical mode	44.91	49.35	-1.69	0.227
- Verbal mode	45.43	51.04		(.82)
Comparing task				
- Numerical mode	48.50	53.30	6.30	0.845
- Verbal mode	41.83	47.00		(.40)

2. Ranking of alternatives

The second hypothesis refers to the ability of the numerical and the verbal mode to predict the ranking of a set of alternatives. Kendall's tau is used to measure the extent to which the ranking of the AHP resembles the ranking of the validation task (Green et al., 1972; Timmermans, 1985; Tscheulin, 1991). The results are displayed in Table 4. For the rating task the mean of the Kendall's taus of both groups is almost equal (numerical 0.47 and verbal 0.46), resulting in a very large p-value of the Mann-Whitney test (this test was used because of the non-normality of the

Kendall's taus). The comparing task shows similar results, although the difference between the means of Kendall's taus for both methods is larger (numerical 0.48 and verbal 0.43). The numerical mode is slightly better than the verbal mode in both cases, but the Mann-Whitney test indicates that the differences between the numerical and the verbal mode are not significant. This means that we can accept the hypotheses that the numerical and the verbal mode are equally able to predict the ranking of a set of alternatives.

Table 4: Rank order correlation coefficient of the numerical mode (n = 91) and the verbal mode (n = 89) (hypothesis 2; Mann-Whitney test).

Rank order correlation Correlation between	Mean of Kendall's Tau	Mann-Whitney test		
		Mean of ranks	U	Z-score (p-value)
Rating task and				
- Numerical mode	.47	91.4	3970	-.228
- Verbal mode	.46	89.6		(.82)
Comparing task and				
- Numerical mode	.48	94.4	3696	-1.02
- Verbal mode	.43	86.5		(.31)

3. Preferences for alternatives

The third hypothesis deals with the differences of the preference scores as predicted by both methods. The Pearson correlation coefficient is used to measure the extent to which the calculated preferences resemble the assigned preferences of the comparing and rating task (Akaah and Korgaonkar, 1983; Van der Lans and Heiser, 1992). The results are displayed in Table 5. The means of the correlation coefficients for both methods are almost equal for the rating task (numerical 0.63 and verbal 0.60). A similar result is found for the comparing task, the mean of the correlation coefficients of the numerical mode is 0.60, and of the verbal mode 0.56.). Again, the numerical mode is slightly better than the verbal mode, but for both tasks the Mann-Whitney test is not significant. Therefore, we accept our hypothesis that the numerical and the verbal mode are equally able to predict the preference scores for a set of alternatives.

Table 5 Pearson correlation coefficient of the numerical mode (n = 91) and the verbal mode (n = 89) (hypothesis 3; Mann-Whitney test).

Pearson correlation Correlation between	Mean of correlation coefficients	Mann-Whitney test		
		Mean of ranks	U	Z-score (p-value)
Rating task and				
- Numerical mode	.63	94.2	3712	-.966
- Verbal mode	.60	86.7		(.33)
Comparing task and				
- Numerical mode	.60	94.4	3863	-.532
- Verbal mode	.56	86.5		(.59)

4. Importance of criteria

According to the fourth hypothesis the numerical and the verbal mode should produce similar importances of the criteria. This hypothesis has been tested for all five criteria (see Table 6). The first two columns contain the means of the importances as computed by the numerical and the verbal mode. A two-tailed t-test was performed to test the hypothesis of equal means. As the p-values in table 6 show, none of the differences is significant. In case of criterion Rent the difference was almost significant (p = 0.06), the other criteria have much larger p-values. This implies that, as hypothesized, the numerical and the verbal mode compute almost equal importances for all five criteria.

Table 6: The importance of criteria as computed by the numerical mode (n = 91) and the verbal mode (n = 89) (hypothesis 4; t-test).

Criterion	Mean of importances:		Difference	T-test	
	Numerical mode	Verbal mode	in means	t-value	p-value (two-tailed)
Rent	.28	.24	.040	1.93	.06
Area	.22	.24	-.014	-.74	.46
Location	.21	.21	-.001	-.06	.96
Type	.17	.19	-.026	-1.25	.21
Facilities	.12	.12	.001	.05	.96

5. Range of weights

The fifth hypothesis assumes a difference between the numerical and the verbal mode. We expect the verbal mode to predict larger differences between criterion levels. For each criterion the difference has been computed between the weight of the most preferred criterion level and the weight of the least preferred criterion level. Table 7 contains the results of this analysis. For all five criteria the range of the verbal mode is larger than the range of the numerical mode. A one-tailed t-test was performed to test the hypothesis that the means of the verbal ranges was significantly different. As displayed in the last column of table 7 all t-values are significant ($p < 0.05$). Most of them are highly significant ($p < 0.000$). This finding confirms our hypothesis that the verbal pairwise comparisons were significantly more extreme than the numerical pairwise comparisons. In other words: the AHP conversion table tends to overestimate the differences as perceived by decision makers.

Table 7: The range of the weights of the most preferred criterion level and the least preferred criterion level as computed by the numerical mode (n = 91) and the verbal mode (n = 89) (hypothesis 5; t-test).

Criterion	Mean of range:		Difference	T-test	
	Numerical mode	Verbal mode	in means	t-value	p-value (one-tailed)
Rent	.51	.54	-.037	-2.12	.018
Area	.46	.54	-.072	-4.70	.000
Location	.50	.59	-.091	-4.80	.000
Type	.53	.60	-.067	-3.78	.000
Facilities	.52	.57	-.047	-2.46	.008

6. Linearity of weights curve

The final hypothesis refers to the degree of linearity of the weights curve. Because it is much easier to attain a linear weights curve when expressing judgements numerically, we expect that the weights curves for the numerical mode are more linear. This hypothesis can be tested only for the quantitative criteria Rent and Area. Both criteria have four levels. For each respondent a regression line was computed. In this analysis the criterion levels were the independent variable and the weights the dependent variable. Next, the sum of squares of the residuals (SSR) was computed for each respondent. Because of the non-normality of the squared and summed residuals, the Mann-Whitney test was used. A one-tailed test was performed to test the hypothesis that the means of the verbal SSR's was significantly larger than the means of the numerical SSR's. The results are displayed in table 8. For both criteria the means of the verbal SSR's was significantly larger ($p < 0.05$). This implies that the weights curve was more linear when the numerical mode had been used. This finding confirms our hypothesis.

Table 8: Degree of linearity measured as the sum of squares of the residuals (SSR) of the numerical mode (n = 91) and the verbal mode (n = 89) (hypothesis 6; Mann-Whitney test).

	Mean of SSR	Mann-Whitney test		
		Mean of ranks	U	Z-score (one-tailed p-value)
Criterion Rent				
- Numerical mode	.0322	84.0	3455	-1.70
- Verbal mode	.0421	97.2		(.04)
Criterion Area				
- Numerical mode	.0243	82.3	3306	-2.13
- Verbal mode	.0304	98.9		(.02)

Summary and Conclusions

In the Analytic Hierarchy Process (AHP) pairwise comparisons are made on a 1-9 scale. The nine levels can be labeled with numbers (the numerical mode) or with preference phrases (the verbal mode). When the verbal mode is used, a conversion table is applied to translate the verbal preferences into numbers. In the area of probability phrases, many studies have shown that people assign very different numbers to the same phrases. In this study we have explored the consequences of different interpretations on the quality of the AHP analysis.

The first part of our analyses concentrated on the outcome of the AHP. We studied (1) the alternative proposed by the AHP, (2) the ranking of the alternatives, and (3) the preference scores of the alternatives. In almost all cases the numerical mode showed better results than the verbal mode, but none of the tests were significant. The same conclusion was drawn for the importances of the five criteria in our experiment. We found small but insignificant differences. However, two other intermediate results were significantly different. The range of weights between the most preferred and the least preferred criterion level was significantly larger for the verbal mode. This applied for all five criteria. We can conclude that, as expected, the conversion table of the AHP overestimates the preferences of decision makers. For example, 'moderately preferred' is assigned the score 3, for most decision makers it should be a lower score. The second significant finding is related to the degree of linearity of the weights curve. For both quantitative criteria we found that this curve tends to be more linear when the decision makers used the numerical mode.

Based on the results of our experiment, we conclude that using the verbal mode without knowing how people interpret the preference phrases leads to a small loss of decision quality. This finding is consistent with other studies. Budescu, Weinberg and Wallsten (1988) had subjects bid for two-outcome gambles. The gambles varied according to domain (gain or loss) and mode of probability presentation (numerical, verbal, or graphical). Subjects earned significantly more money in the numerical (and graphical) mode than in the verbal one, but the difference was small in each domain. Over all gambles with positive expected value (the gain games), subjects earned only 1.23% less money in the verbal condition than in the other two conditions combined. The losses were larger by only 4.65%. Erev and Cohen (1990) found similar results. They asked four basketball experts to assess the probabilities of upcoming basketball game events. The probabilities were expressed both in numbers and in phrases. Undergraduates used the assessments for making gambling decisions. The average difference between the decision makers' profits using the numerical and the verbal mode was in favor of the numerical mode but far from significant. Given the well-known preference of people for verbal instead of numerical judgements (Zimmer, 1983; Budescu and Wallsten, 1985; Zwick, 1987; Hamm, 1991), the optimal solution would be to include a personal conversion table for each decision maker. Although optimal, it may not be a practical solution. Therefore, if accuracy is not very important, we can support the conclusion of Hamm (1991): in these situations the ease and comfort of verbal expressions may be worth the small cost.

The results of the verbal mode of the AHP might be improved by using an alternative AHP conversion table. Examples are the 9/9-to-9/1 scale (Ma and Zheng, 1991) and the balanced scale (Salo and Hämäläinen, 1993). Systematic research is needed on the quality of these and other scales that can be applied within AHP. These studies should also pay attention to the context of the decision situation. For example, outcome severity and outcome valence have found to affect the meanings of probability phrases. These findings directly address the validity of the AHP assumption that a single conversion scale can be applied for all decision makers in all circumstances.

References

Akaah, I.P., P.K. Korgaonkar (1983), An Empirical Comparison of the Predictive Validity of Self-Explicated, Huber-Hybrid, Traditional Conjoint, and Hybrid Conjoint Models, *Journal of Marketing Research,* Vol. XX, 187-197.

Bass, B.M., W.F. Cascio, E.J. O'Conner (1974), Magnitude estimation of expressions of frequency and amount, *Journal of Applied Psychology,* 59, 313-320.

Beyth-Marom, R. (1982), How probable is probable?, Numerical translation of verbal probability expressions, *Journal of Forecasting,* 1, 257-269.

Brun, W., K.H. Teigen (1988), Verbal Probabilities: Ambiguous, Context-Dependent, or Both?, *Organizational Behavior and Human Decision Processes,* 41, 390-404.

Bryant, G.D., G.R. Norman (1980), Expressions of probability: Words and Numbers, *The New England Journal of Medicine,* 302, 411.

Budescu, D.V., T.S. Wallsten (1985), Consistency in Interpretation of Probabilistic Phrases, *Organizational Behavior and Human Decision Processes,* 36, 391-405.

Budescu, D.V., T.S. Wallsten (1990), Dyadic Decisions with Numerical and Verbal Probabilities, *Organizational Behavior and Human Decision Processes,* 46, 240-263.

Budescu, D.V., T.S. Wallsten (1995), Processing linguistic probabilities: General Principles and empirical evidence, in: J. Busemeyer, R. Hasties and D.L Medin (eds.), *Decision Making from a Cognitive Perspective,* Academic Press, 275-318.

Budescu, D.V., S. Weinberg, T.S. Wallsten (1988), Decisions Based on Numerically and Verbally Expressed Uncertainties, *Journal of Experiment Psychology: Human Perception and Performance,* 14, no. 2, 281-294.

Buede, D.M. (1992), Software Review: Three Packages for AHP: Criterium, Expert Choice and HIPRE III+, *Journal of Multi-Criteria Decision Analysis,* 1,2, 119-112.

Clarke, V.A., C.L. Ruffin, D.J. Hill, A.L. Beaman (1992), Ratings of orally presented verbal expressions of probability by a heterogeneous sample, *Journal of Applied Social Psychology,* 22, 638-656.

Dyer, R.F. and E.H. Forman (1991), *An Analytic Approach to Marketing Decisions,* Prentice Hall, Englewood Cliffs.

Elrod, T., J.J. Louviere and S.D. Krishnakumar (1992), An Empirical Comparison of Ratings-Based and Choice-Based Conjoint Models, *Journal of Marketing Research,* Vol. XXIX, 368-377.

Erev, I.; B.L. Cohen (1990), Verbal versus Numerical Probabilities: Efficiency, Biases and the Preference Paradox, *Organizational Behavior and Human Decision Processes,* 45, 1-18.

Fishburn, P.C. (1988), *Nonlinear Preference and Utility Theory,* John Hopkins University Press, Baltimore, Maryland.

Green, P.E., F.J. Carmone and Y. Wind (1972), Subjective evaluation and conjoint measurement, *Behavioral Science,* Vol. 17, 288-299.

Hakel, M. (1968), How often is often?, *American Psychologist,* 23, 533-534.

Hamm, R.M. (1991), Selection of Verbal Probabilities: A Solution for Some Problems of Verbal Probability Expression, *Organizational Behavior and Human Decision Processes,* 48, 193-223.

Hauser, J.R. and F.S. Koppelman (1979), Alternative Perceptual Mapping Techniques: Relative Accuracy and Usefulness, *Journal of Marketing Research,* Vol. XVI, 495-506.

Huizingh, K.R.E., H.C.J. Vrolijk (1996), Extending the Applicability of the Analytic Hierarchy Process, *Socio-Economic Planning Sciences* (forthcoming).

Johnson, E.M. (1973), *Numerical encoding of qualitative expressions of uncertainty,* Techn. Paper 250, U.S. Army Research Institute for the Behavioral and Social Sciences, Alexandria, VA.

Karmarker, U. (1978), Subjectively Weighted Utility, *Organisational Behavior and Human Performance,* 21, 61-72.

Kong, A., G.O. Barnett, F. Mosteller, C. Youtz (1986), How medical professionals evaluate expressions of probability, *The New England Journal of Medicine,* 315, 740-744.

Leigh, T.W., D.B. MacKay and J.O. Summers (1984), Reliability and Validity of conjoint analysis and self-explicated weights: a comparison, *Journal of Marketing Research,* Vol. XXI, 456-462.

Lichtenstein, S., J.R. Newman (1967), Empirical scaling of common verbal phrases associated with numerical probabilities, *Psychonomic Science,* 9, 563-564.

Ma, D., X. Zheng (1991), 9/9 - 9/1 Scale Method of AHP, *Proceedings of the 2nd International Symposium on the AHP,* vol. I, Pittsburgh, PA, 197-202.

Mapes, R.E.A. (1979), Verbal and Numerical Estimates of Probability in Therapeutic Contexts, *Soc Science Medicine,* 13, 277-282.

Miller, G.A. (1956), The magical number seven plus or minus two: Some limits on our capacity for processing information, *Psychological Review,* 63, 81-97.

Nakao, M.A., S. Axelrod (1983), Numbers are better than words: Verbal specifications of frequency have no place in medicine, *The American Journal of Medicine,* 74, 1061-1063.

Pepper, S., L.S. Prytullak (1974), Sometimes frequently means seldom: Context effects in the interpretation of quantitative expressions, *Journal of Research in Personality,* 8, 95-101.

Pöyhönen, M.A., R.P. Hämäläinen, A.A. Salo (1996), An experiment on the numerical modelling of verbal ratio statements, *Journal of Multi-Criteria Decision Analysis,* Vol. 5 (forthcoming).

Saaty, T. (1977), A Scaling Method for Priorities in Hierarchical Structures, *Journal of Mathematical Psychology,* 15, 234-281.

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

Saaty, T.L. (1986), Absolute and relative measurement with the AHP, the most livable cities in the United States, *Socio-Economic Planning Science,* Vol. 20, 327-331.

Salo, A. and R.P. Hämäläinen (1993), *On the measurement of Preferences in the Analytic Hierarchy Process,* Helsinki University of Technology, System Analysis Laboratory Research Reports, Espoo.

Schoemaker, P.J.H. and C.C. Waid (1982), An Experimental Comparison Of Different Approaches to Determining Weights in Additive Utility Models, *Management Science,* Vol. 28, 182-196.

Simpson, R.H. (1944), The specific meanings of certain terms indicating differing degrees of frequency, *Quarterly Journal of Speech,* 30, 328-330.

Simpson, R.H. (1963), Stability in meanings for quantitative terms: A comparison over 20 years, *Quarterly Journal of Speech,* 49, 146-151.

SPSS (1990), *SPSS Categories,* SPSS Inc., Chicago.

Stone, D.R., R.J. Johnson (1959), A study of words indicating frequency, *Journal of Educational Psychology,* 50, 224-227.

Timmermans, D. (1994), The Roles of Experience and Domain of Experience in Using Numerical and Verbal Probability Terms in Medical Decisions, *Medical Decision Making,* vol. 14, no. 2, 146-156.

Timmermans, H. (1985), Hybrid and Non-Hybrid Evaluation Models for Predicting Outdoor Recreation Behavior: A Test of Predictive Ability, *Leisure Science,* Vol. 9, 67-76.

Tscheulin, D. (1991), Ein Empirischer Vergleich der Eignung von Conjoint-Analyse und Analytic Hierarchy Process (AHP) zur Neuproduktplannung, *Zeitschrift fur Betriebswirtshaft,* Vol. 61, 1267-1279.

Van der Lans, I.A., W.J. Heiser (1992), Constrained part-worth estimation in conjoint analysis using the self-explicated utility model, *International Journal of Research in Marketing,* Vol. 9, 325-344.

Zimmer, A.C. (1983), Verbal vs. numerical processing of subjective probabilities, in: R.W. Scholz (ed.), *Decision Making Under Uncertainty,* North-Holland, Amsterdam, The Netherlands.

Zwick, R. (1987), *Combining stochastic uncertainty and linguistic inexactness: Theory and experimental evaluation,* Ph.D. dissertation, Unversity of North Carolina, Chapel Hill.