Memory & Cognition
1989, 17 (3), 249-264

Timed magnitude comparisons of numerical and nonnumerical expressions of uncertainty

AMANDA JAFFE-KATZ and DAVID V. BUDESCU
University of Haifa, Haifa, Israel

and

THOMAS S. WALLSTEN
University of North Carolina, Chapel Hill, North Carolina

Two experiments involving paired comparisons of numerical and nonnumerical expressions of uncertainty are reported. Subjects were timed under two opposing sets of instructions ("choose higher probability" vs. "choose lower probability"). Numerical comparisons were consistently faster and easier than their nonnumerical counterparts. Consistent distance and congruity effects were obtained, illustrating that both numerical and nonnumerical expressions of uncertainty contain subjective magnitude information, and suggesting that similar processes are employed in manipulating and comparing numerical and verbal terms. To account for the general pattern of results obtained, Holyoak's reference point model (1978) was generalized by explicitly including the vagueness of the nonnumerical expressions. This generalized model is based on the notion that probability expressions can be represented by membership functions (Wallsten, Budescu, Rapoport, Zwick, & Forsyth, 1986) from which measures of location for each word, and measures of overlap for each pair can be derived. A good level of fit was obtained for this model at the individual level.

The concept of subjective probability refers to an individual's opinion regarding the uncertainty of an event, or the truth of a statement. Such opinions are manifested in behavior, and need not be explicitly verbalized or quantified. However, when it is necessary to communicate the opinions to others, people can use one of two modes of expression, numerical or verbal. Most decision theories tacitly assume that the internal assessment of opinion is independent of the mode of response used to transmit its outcome. However, a given mode of response (verbal vs. numerical) may aid this translation of internal opinion to external judgment, or indeed even affect the formation of the internal opinion. For example, Zimmer (1984) regards the verbal mode to be the more natural one for processing probabilistic information. It follows, he claims, that if one is forced to provide numerical estimates, one is obliged to operate in a mode that requires "more mental effort" and is therefore more prone to interference from biasing tendencies.

This notion is consistent with results of a survey of over 400 people concerning numerical and verbal assessment of probabilities (Wallsten, Zwick, Kemp, & Budescu, 1988). A majority (77%) thought that most people prefer to express uncertainty verbally in everyday life. When asked about their own preferences, 65% stated that they personally prefer to express uncertainty verbally to other people, while 70% preferred to receive it numerically from others. Furthermore, 36% of the respondents preferred expressing information in one mode and receiving it in the other. Thus, it appears that people consider both modes feasible and acceptable, at least under certain circumstances.

Sometimes this reality is used to justify the claim that the various expressions of certainty can be directly mapped onto the 0-1 interval of probabilities. Most recently this claim was made by Kong, Barnett, Mosteller, and Youtz (1986), who suggested that such a mapping would enhance communication among medical professionals. Numerous attempts to decide just "how often is often?" or "how likely is likely?" have appeared in the psychological literature in the past 40 years (for an almost exhaustive list of references see Budescu & Wallsten, 1987, and Wallsten, Budescu, Rapoport, Zwick, & Forsyth, 1986). Most of these studies have involved quantification of verbal expressions of uncertainty, and they have shown that there is a large amount of variability in the numerical values assigned by subjects to nonnumerical probabilistic terms, and that a great deal of overlap among such terms exists. No data are available regarding the quality of conversions of numerical judgments to verbal probability expressions, but Budescu, Weinberg, and Wallsten (1988) show that when numerical and verbal judgments are provided for the same events, the former are more reliable, precise, and consistent.

This research was supported by a grant from the Binational Science Foundation U.S.-Israel (Jerusalem), and a Hugo Bergmann award to David V. Budescu. We wish to thank Hannah Shtrominger for writing the computer programs for the experiments, and Ilana Schindler and Taly Shilon for assistance in data collection. For reprints, please write Thomas S. Wallsten, Department of Psychology, University of North Carolina, Chapel Hill, NC 27599-3270.

When evaluating this body of research, it is important that one identify the source of the large variability. In the few studies that have allowed decomposition of the variance (Beyth-Marom, 1982; Budescu & Wallsten, 1985; Johnson, 1973), it was established that while individuals are relatively consistent in their assignment of numbers to phrases, there is high *intersubject* variability. More recent work with membership functions (Rapoport, Wallsten, & Cox, 1987; Wallsten, Budescu, Rapoport, et al., 1986) suggests that probability phrases have vague meanings for individuals also.

The purpose of the present study is to probe further the similarities and differences between numerical and verbal expressions of probability. Unlike previous studies, ours goes beyond the simple task of conversion from one mode to another, and uses instead a paired-comparison paradigm to investigate the cognitive processes employed in the comparison of probability words and numbers. The subjects are presented with two probability numbers (NN), two probability words (WW), or a word and a number (WN), and they must choose, while being timed, which of the two items describes a higher/lower level of certainty. A voluminous literature documents several consistent and systematic choice-time phenomena for magnitude comparisons in other domains. The first goal of the present study is to demonstrate that the same phenomena are obtained when comparing pairs of probability values, probability phrases, and mixed pairs. Such a demonstration would lend support to the notion that both modes of expressing uncertainty are based on the same internal representation of beliefs. Moreover, it is of interest to investigate and compare these effects across the three types of pairs (NN, WW, and WN).

Empirical Phenomena in Paired Comparisons

The paired-comparison paradigm applied to mental representations of semantic information has established two major robust empirical phenomena. The first, known as *the (symbolic) distance effect* refers to the fact that performance, whether measured by accuracy or response time (RT), improves as the distance between the quantities denoted by the two stimuli increases. In other words, the closer the pair members are, the harder it is to compare them. This effect was first described by Moyer and Landauer (1967) for pairs of digits, and it has since been replicated (e.g., by Buckley & Gillman, 1974). The distance effect has since been found ubiquitously for every symbolic attribute used in paired comparisons: digit names (e.g., by Foltz, Poltrock, & Potts, 1984); pairs of letters in the alphabet (e.g., by Parkman, 1971); size of animals or other objects (e.g., by Jamieson & Petrusic, 1975; Paivio, 1975); natural semantic orderings such as time, temperature and quality (e.g., by Holyoak & Walker, 1976); artificial orderings (e.g., by Moyer & Bayer, 1976; Potts, 1974). It should be noted that the time, temperature, and quality scales employed by Holyoak and Walker (1976) differ from the other continua for which the symbolic distance effect has been reported. These concepts appear to be ordered on the basis of linguistic knowledge; for example, *decade* versus *day* or *torrid* versus *cool*. In the present study, the probability expressions also form a linguistically-based scale.

The second phenomenon, *the (semantic) congruity effect*, describes the interaction between the direction of judgment required by the instructions and the position of the stimuli along the judged continuum. Comparisons are faster when the instructions are congruent (or "match") the stimuli than when they "disagree" with them. That is to say, it is easier to identify the larger of two large stimuli or the smaller of two small stimuli than it is to select the larger of two small stimuli or the smaller of two large stimuli. The interaction may be of any form (in the anticipated direction), and need not necessarily display a crossover pattern. This effect was demonstrated initially for pairs of digits by Banks, Fujii, and Kayra-Stuart (1976).

Unidimensionality of Stimuli

The distance and congruity effects are based on the notion that the stimuli are located along a unidimensional continuum. The numbers fall naturally on such a scale, but probability words do not necessarily satisfy this assumption (Budescu & Wallsten, 1985). Therefore, the second goal of this work is to test the unidimensionality of the probability phrases and of the combined scale of phrases and numbers. For this purpose, additional untimed paired comparisons were run, in which subjects judged ratios of likelihoods. The ratio judgments were then subjected to scaling procedures, to determine whether the stimuli could be scaled in one dimension, and if so, to derive the scale values. While the scale values are of interest in their own right as a further means of comparing numerical and verbal expressions of uncertainty, they also formed part of the basis for investigating the distance effect.

A Model of Magnitude Comparisons

We hypothesized that verbal and numerical descriptors of probability can be conceived of as located along a single continuum of uncertainty, and that similar cognitive processes are invoked in manipulating and comparing their magnitude. It is important to emphasize that the two modes of description clearly differ in other ways, and that we do not expect the various types of comparisons to be indistinguishable. Previous work (Budescu & Wallsten, 1985; Budescu et al., 1988; Wallsten, Budescu, & Zwick, 1986) has established some of these differences, which would necessarily have an impact on the ease and rate of the comparisons. Thus the third, and final, goal of this work is to develop a general model of the comparison of verbal and numerical descriptors of probabilities. We elected to build upon the model proposed and successfully tested by Holyoak (1978) for the NN comparisons. The model is relatively simple, has considerable intuitive and psychological appeal, and has been strongly supported empirically (Holyoak, 1978; Holyoak & Patterson, 1981).

Finally, as we hope to show, it can be easily and naturally extended to incorporate the special features of WW and WN comparisons.

Holyoak's (1978) Reference Point Model

This analog comparison model assumes that scales in memory are conceptually bounded at each end. A decision about relative magnitude is based on the ratio of the distance between each item and a reference point determined by the question. The question may imply one of the endpoints or some other value as the reference. The model further postulates a mental process in which the distances of the two stimuli from the reference point are compared repeatedly until a decision criterion is reached. Because large differences in magnitude will be detected with few comparisons, whereas concepts relatively similar in magnitude will require a larger number of comparisons, it is predicted that RT will decrease monotonically as the ratio of the two distances departs from unity. This, of course, is the distance effect. The congruity effect follows from Weber Law considerations, according to which differences at low stimulus magnitudes are more discriminable than equal differences at high stimulus magnitudes. Specifically, two terms will be discriminated more quickly when they are close to the reference point indicated by the question than when they are far from it, because the distances in the ratio will be smaller in the former than in the latter case.

To generalize this model to our situation, we assume that under the "choose lower/higher" instructions the subject operates with the implicit reference points located at 0 and 1, respectively, with the relevant reference point determined by the nature of the instructions. Calculation of the distance between the relevant reference point and the two stimuli, as required by the model, is a simple and straightforward procedure for precise numerical values. However, phrases are vague and do not have a unique numerical representation. Therefore, we must assume the existence of a stage in which the vagueness of each nonnumerical stimulus is tentatively resolved and is represented by a single value. In calculating distances from the reference point, the subjects treat these values like the numerical stimuli. The existence of this stage implies that all comparisons not involving phrases (i.e., NN) would be performed faster, as no resolution is necessary. To keep the model as simple as possible, we further assume initially that the time required for this preliminary resolution of vagueness is approximately fixed, and that it does not depend on the actual terms compared.

The final stage in Holyoak's model assumes that the subject performs repeated observations of the relevant magnitudes in order to estimate their true values with sufficient precision. The number of repeated observations and, consequently, the time taken to compare two stimuli depend on the ratio of their distances from the relevant reference point. This assumption is based on the notion that all stimuli are equally precise, and that problems of comparison are related only to their proximity. However, in the case of nonnumerical representations, an additional attribute of the stimuli—their vagueness/precision—must be considered. Specifically, we assume that the number of repeated observations will also be related to the level of confusability, or overlap, between the stimuli. Thus, the RT for WW and WN pairs will be longer not only because of the need to resolve the phrases, but also because a greater number of repeated observations will be required to reach the decision criterion.

We describe two experiments designed to test this particular extension of Holyoak's model (1978). In both experiments, the subjects were asked to perform paired comparisons of verbal and numerical descriptors of probability under different instructions. In the first experiment, we sought to establish (1) the unidimensionality of the continuum of uncertainty underlying both types of representations, (2) the qualitative resemblance of the cognitive processes involved in all types of comparisons, and (3) the quantitative differences among the three types of comparisons, as predicted by the model. In the second experiment, we introduced a more specific characterization of the vagueness of the phrases that allowed us to derive measures of overlap among the various stimuli and to test quantitatively the fit of the proposed model.

EXPERIMENT 1

Method

Subjects

Twenty native English speakers were paid for their participation in this experiment. There were 6 male and 14 female subjects, most of whom were aged between 21 and 30. Four of the original subjects had to be replaced due to their high degree of inconsistency during the first experimental session. (A maximum of 15 inconsistent judgments was arbitrarily held to be permissible, based on a pilot run.)

Materials

Besner and Coltheart (1979) suggest that perception of digits may involve different (i.e., ideographic) cognitive processing from that of word perception (alphabetic processing). Therefore, the numerical expressions were presented as digit *names*, in an attempt to match the time necessary for reading the stimuli, thus allowing a meaningful comparison of RTs. Seven nonnumerical expressions of uncertainty (IMPOSSIBLE, IMPROBABLE, UNLIKELY, POSSIBLE, LIKELY, PROBABLE, CERTAIN) and seven numerical expressions of uncertainty (FIVE%, TWENTY%, THIRTY-FIVE%, FIFTY%, SIXTY-FIVE%, EIGHTY%, NINETY-FIVE%) were paired, generating 91 stimuli as follows:

$(7\times6)/2 = 21$ word-word (WW) comparisons;
$(7\times6)/2 = 21$ number-number (NN) comparisons; and
$7\times7 = 49$ word-number (WN) comparisons.

The experiment was performed on a Visual 200 Terminal connected to a DEC LSI-11/23 computer. A three-buttoned keyboard was employed for recording the subjects' responses.

Procedure

The subjects were presented with pairs of probability descriptors (words or numbers) and were required to make (1) a speedy magnitude judgment as to which member of the pair conveyed a greater or lesser degree of uncertainty, and subsequently (2) an untimed ratio judgment regarding the two terms in the pair. Each subject participated in two experimental sessions; the mean intersession interval was 9 days. In Session 1, half of the subjects were required to choose the term denoting the *higher* probability in the pair, and the other half the term denoting the *lower* probability. This was reversed in Session 2. The subjects were first familiarized with the list of probability expressions, including several additional terms that were subsequently presented only in the warm-up trials. Ten practice trials were presented, the results of which were discarded from the RT analysis. The subjects were then able to opt for further practice (the same 10 trials repeated), or to proceed immediately with the experiment. Three subjects requested additional practice.

A series of 182 trials was presented, each of the 91 paired comparisons appearing randomly twice, with the constraint that no 2 consecutive trials could present the same pair. The response-stimulus interval was 1 sec. The two terms were displayed side by side within a rectangular frame on the terminal screen, employing 5×7 dot matrix characters. On the second presentation, the locations of the respective terms alternated, such that each pair was balanced within the session with regard to left/right presentation. If the subject made inconsistent judgments regarding a pair, it was presented a third time. In such cases, the respective left/right location of the two members of the pair was selected randomly. The total number of paired comparisons per session, therefore, could surpass the 182 minimum. While artificially limited, the occurrence of these inconsistent judgments indicated the "error rate" of an otherwise free-choice task. The RTs of these "tie-breakers" were not included in the analyses, and these judgments served merely to obtain a "dominant" choice to be used in the second task (the ratio judgment).

Once all the timed paired comparisons had been made, the subject was asked to provide a ratio judgment, with no time constraints, for each pair. This was achieved by moving a cursor on a 30-point scale, depicting the ratio by which one term represented greater certainty than the other. For each pair, the *greater* term was determined by the subject's previous magnitude judgments. The extremities of the scale were accompanied by verbal labels, that is "1:1 ratio" and "maximal ratio," and the cursor was presented at "1:1 ratio."

In the second session, the 91 pairs were again presented (at least) twice, but now the judgment was made according to the alternative set of instructions ("choose higher probability" or "choose lower probability"). Again, the subjects made a ratio judgment for each paired comparison based on their own dominance judgments.

In total, in the course of the two experimental sessions, each subject made (a minimum of) 364 (91×2×2) magnitude judgments under RT constraints, and 182 (91×2) ratio judgments.

Results

Scaling the Data

We used three alternative methods, with appropriate measures of goodness of fit, for deriving ordinal scale values at the individual level for each of the probability expressions employed: (1) Total vote count (TVC) (Coombs, 1964): each word was given a rank based on the total number of times it was judged to be higher than every other word in the paired comparisons dominance judgments. The poorness of fit measure is the number of intransitive triples. (2) Geometric means (Crawford & Williams, 1985; Torgerson, 1958) and (3) eigenvector/eigenvalues (Saaty, 1977, 1980) were derived from the ratio judgments. By Saaty's method, a subjective scale is determined for the elements based on an eigenvector analysis of the matrix of pairwise comparisons. The eigenvector provides the priority orderings and the eigenvalue is a measure of the consistency of judgment. Crawford and Williams (1985), however, suggest that as an estimator of ratio scales, the geometric mean vector is preferable to the dominant eigenvector in several respects. Both these methods were applied to the data, and they in fact produced similar results. Indices of fit for ratio scales have been discussed by Budescu, Zwick, and Rapoport (1986), who have also calculated critical values based on Monte Carlo results. All the indices of fit indicated that the three scales were unidimensional for most (87%) of the subjects.

The numerical expressions were simply ranked 1-7 in ascending order, except in the case of 1 subject, who in Session 2 judged 50% to convey greater certainty than 65%. This subject's idiosyncratic ordering was employed in his particular case. Ordinal scales were obtained for each subject for each session, combining the three methods for obtaining scale values outlined above. The overall distribution of ranks of each of the probability terms in these combined WW and WN scales are summarized across subjects in Table 1.

An additional scaling was obtained from the ratio judgments at the *group* level, by applying the ALSCAL MDS (multidimensional scaling) procedure (Takane, Young, & de Leeuw, 1977). A good one-dimensional solution resulted, with a Kruskall Stress 1 value of 0.083. The scale values thus derived for the 14 probability terms were subjected to a linear transformation rescaling them on a 0 to 1 range, and they are presented graphically in Figure 1.

Do the words take on a different position when combined with numbers on a single scale? For 9 subjects, the words maintained their positions relative to one another, on both scales. Of the remaining 11 subjects, for only 2 did inversions occur in both sessions. Of these 11 subjects, 85% reflected but a single discrepancy in the ranking of words in the WN as compared to the WW scale. The lowest Kendall rank correlation obtained between the two rankings was 0.714, indicating three inversions.

We also examined the Kendall rank correlations across and between subjects. The NN scale was, as expected, consistent except for the one anomaly mentioned above. For WN and WW, the highest correlations were found within subjects across sessions (0.910 and 0.905, respectively). The between-subjects correlations ranged from 0.799 to 0.826 for WW, and 0.819 to 0.831 for WN, demonstrating high intersubject agreement with regard to ranking the probability expressions. Despite this overall consistency and stability in ranking, the individuals' rankings were used for the purpose of all other analyses involving RTs. That is, we determined semantic distance individually for each subject, according to their own rankings. Thus, in the group analyses of RT, the rank distances (or lags), rather than specific paired expressions, remain constant across subjects.

Analysis of Response Times

The data collected include RTs for 7,280 trials. The overall mean and median rates of response (reciprocal RTs) are 0.695/sec and 0.681/sec, respectively. In the following analyses, the dependent variable reported is *median rate of response* (see Wainer, 1977, for a discussion

Table 1
Percentage of Rankings Received by Any Term Across Subjects

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
						WW Ranks								
IMPOSSIBLE	92.5	5.0	2.5											
IMPROBABLE	5.0	60.0	32.5	2.5										
UNLIKELY	2.5	30.0	65.0	2.5										
POSSIBLE		5.0		67.5	17.5	10.0								
LIKELY				17.5	50.0	32.5								
PROBABLE				10.0	32.5	57.5								
CERTAIN							100.0							
						WN Ranks								
IMPOSSIBLE	85.0	12.5		2.5										
5%	12.5	42.5	40.0	5.0										
IMPROBABLE	2.5	27.5	20.0	40.0	7.5	2.5								
20%		15.0	10.0	65.0	10.0									
UNLIKELY		15.0	22.5	37.5	15.0	7.5	2.5							
35%			2.5	5.0	70.0	17.5		5.0						
POSSIBLE		2.5	2.5	2.5	7.5	7.5	25.0	25.0	10.0	5.0	10.0	2.5		
50%						2.5	45.0	35.0	10.0	5.0	2.5			
PROBABLE								12.5	15.0	22.5	27.5	15.0		
LIKELY							2.5	20.0	27.5	25.0	25.0			
65%								7.5	32.5	40.0	17.5	2.5		
80%											17.5	77.5	5.0	
95%													85.0	15.0
CERTAIN										2.5		2.5	10.0	85.0

of the relative merits of mean and median RTs and their reciprocals).

The effects of the independent variables session, instruction set, and type of judgment (NN, WW, or WN) on rate of responding are shown in Table 2. All three main effects are significant, while none of the interactions are. The rate of responding was greater in Session 2 than in Session 1 [0.760 vs. 0.669; $F(1,119) = 8.63, p < .01$] and under "choose higher" than under "choose lower" instructions [0.749 vs. 0.680; $F(1,119) = 4.82$, $p < .05$]; and it was fastest for NN [0.885; $F(2,119) = 29.99, p < .01$], with no significant difference between WW (0.644) and WN (0.615).

The distance effect. The expected monotonic pattern of RTs was obtained; it is displayed in Figure 2. The effect of distance is significant in all three cases [NN, $F(5,95) = 17.95$, $p < .01$; WW, $F(5,95) = 57.10$, $p < .01$; and WN, $F(12,213) = 35.10, p < .01$]. Similarly, instruction mode has a significant effect, reflecting the consistent tendency for decisions to be made quicker under the "choose higher probability" condition [NN, $F(1,19) = 6.40$, $p < .05$; for WW, $F(1,19) = 5.96$, $p < .05$; and for WN, $F(1,19) = 5.22, p < .05$]. There are no significant interactions of distance with instruction mode, and thus the curves are plotted to represent all the responses in both experimental sessions.

The congruity effect. Following Banks and Flora's (1977) grouping of stimuli, paired comparisons were designated *low*, if both members were ranked below the midpoint(s) of the scale, *high* if both members were ranked above the midpoint(s) of the scale, or *mixed*, if comprising one small and one large member, or the midpoint itself. The mixed pairs were excluded from the analyses in order to test the principle of congruity with the two sets of instructions. The question of interest, then, is: are high pairs judged more quickly under "choose higher probability" instructions than under "choose lower probability" instructions, and does the opposite hold for low pairs? The median rates of response are presented graphically in Figure 3.

A preliminary three-way ANOVA (size × instruction × type of judgment) was performed. Since the three-way interaction of size × instruction × type was significant [$F(2,38) = 4.27, p < .05$], we conducted separate two-way (size × instruction) ANOVAs for NN, WN, and WW judgments. Significant two-way interactions of size × instruction obtain for each of the three cases, all in the predicted direction [NN, $F(1,19) = 7.40, p < .05$; WW, $F(1,19) = 23.26, p < .01$; and WN, $F(1,19) = 20.92$, $p < .01$]. The variable *size* has differential effects on each of the three continua: for NN, high pairs are always judged at a slower rate [$F(1,19) = 5.91, p < .05$]; for WW, the converse is true—that is, low pairs are always judged more slowly [$F(1,19) = 44.67, p < .01$]; and for WN, there is no main effect of size [$F(1,19) = 2.18$], and a crossover effect obtains. Instruction mode bears a significant main effect only for WN [$F(1,19) = 10.94$, $p < .01$]; it is expressed by "choose higher probability" facilitating a speedier rate of response.

Note in Figure 3 that the NN decisions were made at a much faster rate, regardless of pair size and instruction, than either WW or WN choices were. What other effects do pair size, instruction mode, and type have on rates of response? A number of post hoc *t*-tests were performed comparing the WW and WN data; because of the multiplicity of post hoc tests applied to the data, only those

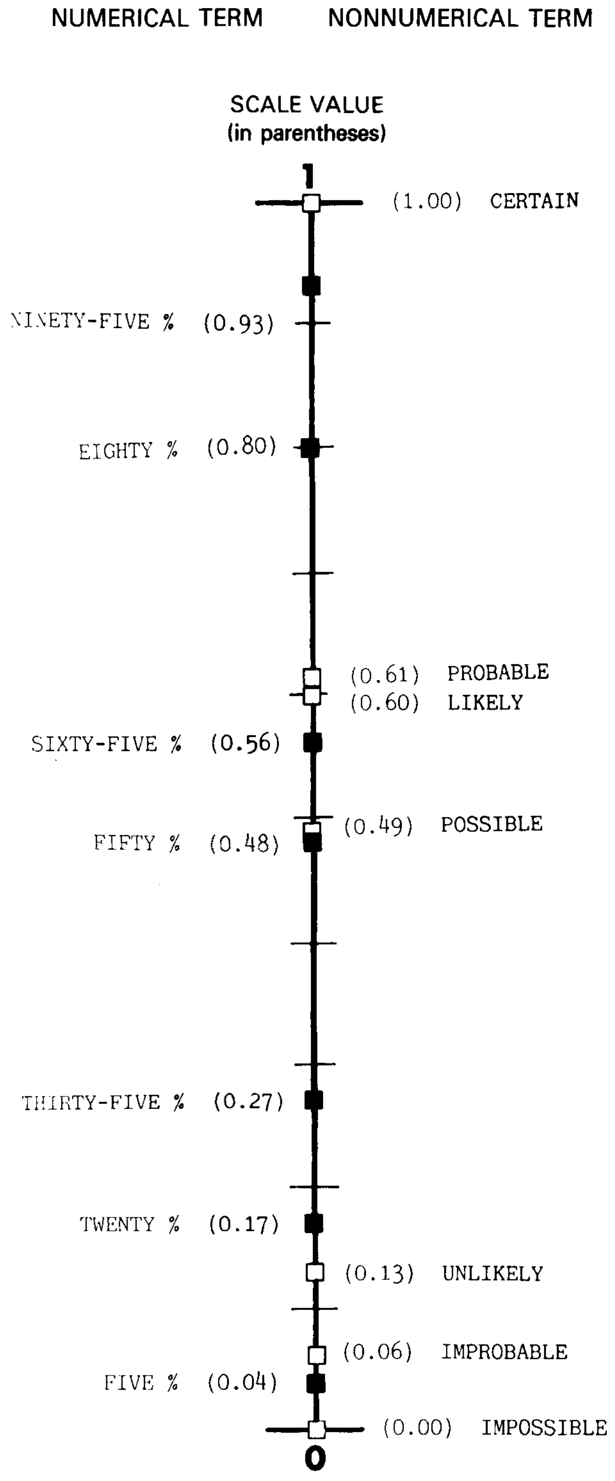

Figure 1. Scale values of the 14 probability terms derived from ratio judgments (converted to 0-1 scale).

which attained the more conservative (.01) level of significance are considered further. The results of these tests reveal that under "choose higher probability" instructions,

$$\text{WW high} = \text{WN high} > \text{WN low} > \text{WW low},$$

while *no* significant differences among the four compared groups obtain under "choose lower probability" instructions. Across instructions, WW high and WN high are both judged more quickly under "choose higher probability" than under "choose lower probability" instructions, while these comparisons are not significantly different when made for low pairs.

Error Rates (Inconsistent Choices)

There were fewer such errors made under "choose higher probability" instructions, [$F(2,119)$ = 7.73, $p < .01$]; and the mean number of such inconsistencies was lowest for NN (1.7%), followed by WW (9.3%) and WN (12.4%) [$F(2,119) = 93.43, p < .01$]. As predicted by the reference point model, there is no speed-accuracy tradeoff: the harder comparisons both require longer processing and are more prone to error.

Test of the Reference Point Model

All model analyses were performed on RTs, not rates, to correspond to the metric in which the model is formulated. The present data only allow for a very limited quantitative test of this model, based on 12 of the 20 subjects who ranked the expressions IMPOSSIBLE as lowest and CERTAIN as highest in both sessions, according to our combined ranking. For these subjects, we assumed, for simplicity, that the 14 stimuli are evenly spaced on the 0-1 interval with these two anchor terms at its endpoints. For each subject, the RTs within session were standardized in order to eliminate practice effects, and two regression equations were fitted, each representing a different version of the general model. In the first version the mean RT per pair was regressed on the ratio of distances from the relevant reference point. This corresponds to Holyoak's version of the model, because it does not distinguish among NN, WN, and WW pairs. The second version incorporated the notion of the constant time required for the resolution of the vagueness for words, by means of a 0-1 dummy variable distinguishing between NN and the other pairs. The fit of the two models is presented in Table 3. A moderate degree of fit was achieved, with a clear advantage to the improved model for all subjects (median R^2 = 0.31 vs. 0.16 for the standard version). We did not perform significance tests of the two R^2 for the observations are not independent and such tests would be misleading.

Discussion

Unidimensionality of Probability Expressions

The various measures of fit indicate almost exclusively that all comparisons (NN, WW, and WN) yield unidimensional scales. This is supported not only by the excellent unidimensional solution attained by the ALSCAL procedure on the ratio judgments for the 14 terms considered together, but also for the 7 numerical, 7 nonnumerical, and 14 word-with-number comparisons analyzed separately, for which, without exception, good one-dimensional solutions were obtained. Whether or not the words are indeed unidimensional, subjects are able to

Table 2
Median Rate of Response and Standard Error of Mean as a Function of Session, Instruction, and Type of Judgment: Experiment 1

	Rate (per second)								
	Session 1			Session 2			Across Sessions		
Instruction	NN	WN	WW	NN	WN	WW	NN	WN	WW
Lower Rate	0.829	0.548	0.576	0.876	0.616	0.638	0.852	0.582	0.607
SEM	.027	.047	.043	.057	.073	.060	.031	.043	.037
Higher Rate	0.833	0.599	0.629	1.001	0.698	0.734	0.917	0.648	0.681
SEM	.069	.073	.058	.023	.040	.047	.040	.042	.039
Across Rate	0.831	0.573	0.603	0.938	0.657	0.686	0.885	0.615	0.644
SEM	.036	.043	.036	.033	.042	0.039	.026	.030	.027

respond to a single dimension, without interference from other dimensions (e.g., vagueness), when answering the unidimensional questions that we asked.

Psychometric Properties of the NN, WW, and WN Types of Judgment

Within-subject correlations of rankings indicate stability of the ordering of the expressions of uncertainty for a given individual over time, and between-subject correlations describe group consensus. Predictably, and consistent with previous results obtained by Budescu and Wallsten (1985), the correlations are highest within subjects, across the two sessions. Specifically, 95% of all rankings are invariant across sessions, and on the average, only one pair is reversed. There is also a high degree of consistency among subjects in the ranking of the seven probability words, as well as in their interleaving with the numerical expressions in the WN scale.

Cognitive Processes Involved in Paired Comparisons

It was hypothesized that the total processing and decision time for any given pair of terms would vary as a function of their mode of expression. The results of the group ANOVA and individual regressions suggest that the critical feature is whether any vagueness resolution has to take place—that is, whether any probability words have to be translated into the number scale. As reported above (see Table 2), NN judgments were indeed faster than WW or WN, and WN were judged more slowly than WW (although not significantly so).

Two strong effects of experimental session and instruction mode were obtained, regardless of type of judgment (see Table 2). First, the rate of responding always increased in Session 2. This was clearly a product of practice, and an indication that in future work, longer practice sessions must be allowed. Second, the rate of responding was always greater under "choose higher probability" instructions. This latter result may be due to the principle of congruence, according to which "a question should be easier to answer whenever the question and the premise containing the answer are congruent in their underlying representations" (Clark, 1969). The principle would apply here if it were the case that probability representations are generally learned and stored in the form "*x* implies greater probability than *y*" rather than "*y* implies less probability than *x*." This explanation, which of course is ad hoc, must be independently substantiated before being accepted.

The distance effect. We only interpret these data at the ordinal level since step sizes are neither equal nor comparable across the three types of pairs. For NN the step sizes are ostensibly objective units of 15, from FIVE% to NINETY-FIVE%, (but see Figure 1 for a comparison of objective vs. subjective values of NN rescaled on a 0-to-1 scale), whereas for WW and WN they are subjective, unequal units, and furthermore the units differ between WW and WN. Relative spacings might be unequal due either to the particular choice of phrases or to the

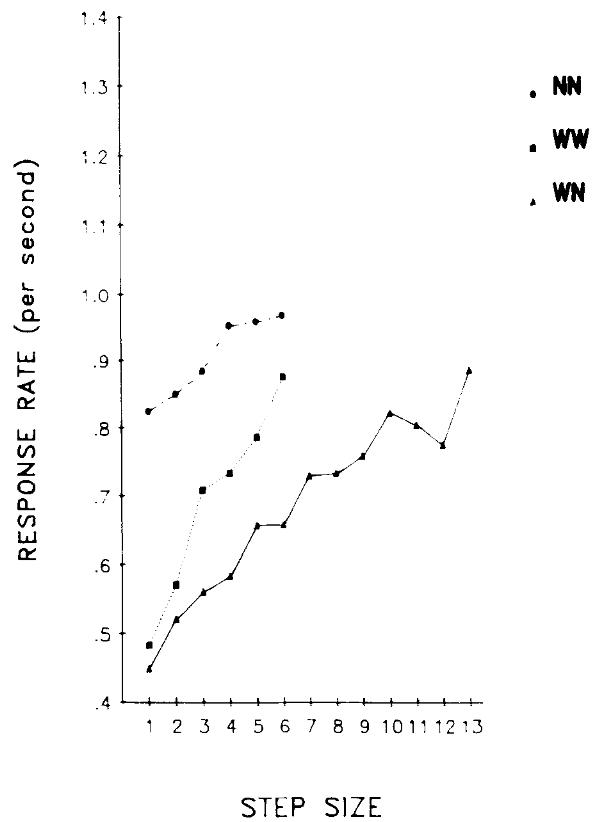

Figure 2. Median response rate as a function of step size (distance) between paired terms: Experiment 1.

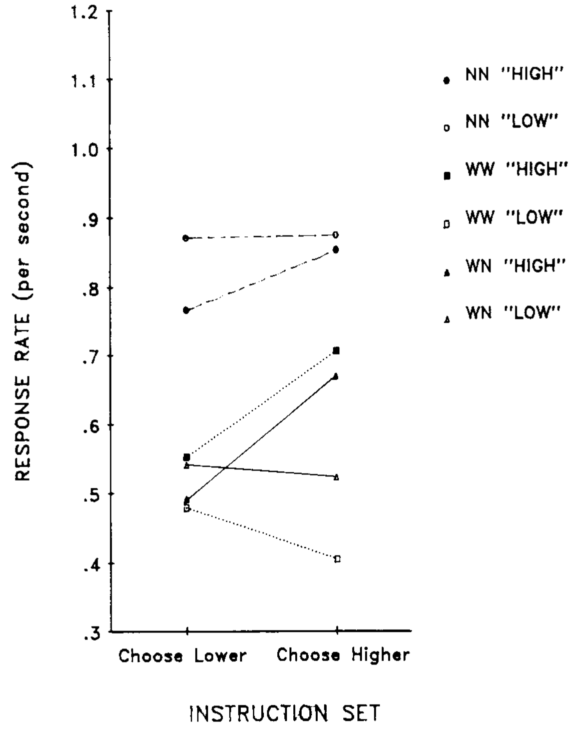

Figure 3. Median response rate as a function of instruction set and pair size: Experiment 1.

greater fuzziness of the words. Clearly, the RTs are ordered correctly and monotonically as a function of ordinal step size in all three types of judgments. The WW distance effect increases monotonically in the anticipated pattern, and 89% of the variance can be accounted for by a linear component. The WN distance effect is very similar in shape, and a partial trend analysis indicates that over 90% of the effect can be accounted for by a linear component. The rate of response here is slower, as predicted by the model, with additional processing time required to resolve the phrase and to make comparisons. The NN curve looks much flatter, and cannot be accounted for easily by a trend analysis. This is due, primarily, to the minimal variance in response rates in this case.

The congruity effect. This effect was obtained in all three types of judgments, but it took a slightly different pattern in each case. It is apparent in Figure 3 that the effect is always stronger for high pairs. This can be attributed to the fact that low pairs are closer together (the average mean distance between low pairs, as derived from the ALSCAL solution, is 0.48, compared to a mean interstimuli distance of 0.76 for high pairs), and therefore, according to Holyoak's model, they required more repeated observations for subjects to reach a predetermined decision criterion.

Additional support for this interpretation is obtained from the fact that, for WW and WN pairs, the responses are always slower for low pairs, which is consistent with an observation made in another context by Wallsten, Fillenbaum, and Cox (1986). They found the meanings of low probability phrases to be less sensitive to base rates than were the meanings of high and neutral phrases, and they interpreted the result as suggesting that low phrases are less vague than the other two types. One explanation they offer for the suggestion that low expressions are less vague and context dependent is that because most events have more than two possible outcomes, neutral probabilities are not fixed and are generally below 0.5. Therefore, the allowable range for low probability phrases is much smaller than for the high or neutral phrases. The present observation that low words are very close together supports the notion that they have less room within which to move.

General Discussion of Cognitive Phenomena

The existence of distance and congruity effects for all three types of judgments demonstrates that the semantic representations of subjective probability terms, both numerical and nonnumerical, contain subjective magnitude information based on the same internal representation of uncertainty. Combined with the derivation of unidimensional magnitude scales, and the recorded differences between median rates of response under the various presentation modes, this result provides strong qualitative support to our own extended version of Holyoak's model. The quantitative test of the model yielded only moderately successful results, however.

Several design shortcomings may account for this latter fact. First, we did not have reliable data regarding the location of the resolved values of the phrases. In the absence of such data, we only used the joint rankings of the words and numbers and assumed equal spacing to derive numerical values. However, the equal spacing assumption was incorrect, since the low phrases were clustered together. Furthermore, the ranking of the stimuli was determined from several scaling procedures, based on either paired comparisons of magnitudes or judgments of ratios of magnitude. These are all *relative* and *comparative* data, while the resolved value is assumed to be the result of a judgment process performed separately and indepen-

Table 3
Goodness of Fit (R^2) of Two Versions of the Model

Subject	Ratio of Distances	Resolution
02	20	42
05	11	23
06	09	21
08	24	41
10	08	20
11	20	31
13	08	26
14	17	39
15	09	30
16	21	31
17	15	22
19	28	37
Median	16	31
All	15	30

Note—Decimal point omitted.

dently for each stimulus. In our extension of the model, we also assumed that the number of repeated observations in each comparison depends on the degree of confusability, or overlap, between the two stimuli. Thus, our test of the model was limited because we had not operationalized this notion of overlap.

The purpose of the second experiment was twofold. The first goal was to replicate the results of the previous study with more data points per comparison, and with numerical representations of the nonnumerical probabilities. The second goal was to test more quantitatively the extended version of the reference point model.

Wallsten, Budescu, Rapoport, et al. (1986) have shown that vague meanings of probability terms can be expressed as membership functions over the {0,1} probability interval. The function takes its minimum value, usually 0, for probabilities not at all in the vague concept represented by the term; its maximal value, generally 1, for probabilities definitely in the concept; and intermediate values otherwise. Derived membership functions have interpretable shapes that in principle can predict the present results. Extreme expressions tend to yield monotonic functions, while the more central ones tend to yield single peaked functions. A pair of functions can be in one of several relationships: they may overlap or be distinct, or one may be enclosed within the other. Furthermore, any function can be vague (the general case for words) or crisp (the general case for numbers). We present, in Figure 4, hypothetical examples of the various relationships for WW and WN pairs (all NN pairs consist of two distinct crisp functions).

The left column displays the various cases for WN pairs: w_1 and p_1 have distinct membership functions, p_2 is enclosed in w_2, and w_3 and p_3 partially overlap. The right column presents similar examples for WW pairs: w_4 and w_5 are distinct words, w_6 is enclosed in w_7, while w_8 and w_9 overlap. This conceptualization of stimuli as membership functions provides natural measures of location and overlap, leading to a set of empirically testable predictions. One of the most commonly used, and

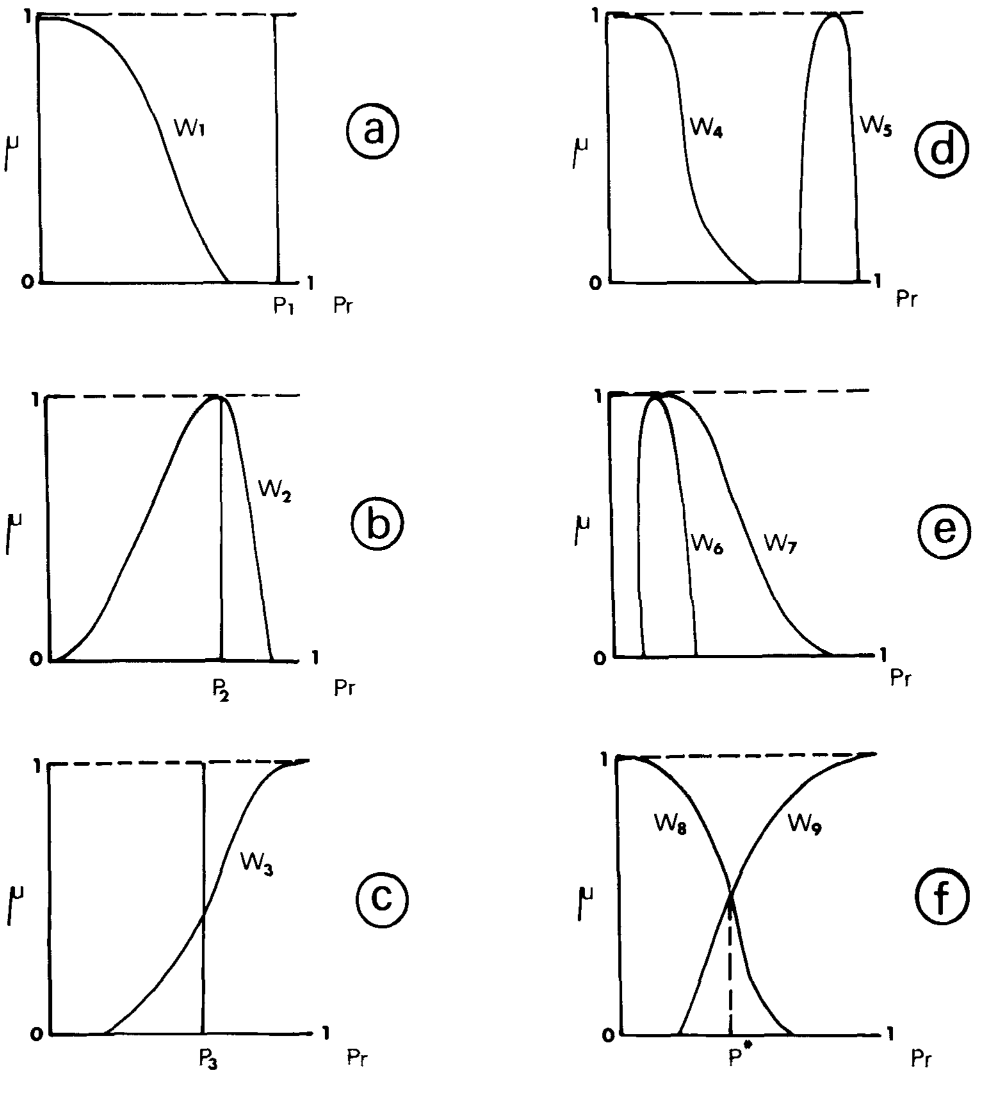

Figure 4. Hypothetical examples of the various relationships between WN and WW pairs.

probably the simplest, ways of ranking elements described by membership functions over the real numbers is to do so according to the functions' central location values, which were defined by Yager (1981) as the mean of the numbers weighted by their normalized memberships. In our case, assume that the membership of probability p in the concept w is $\mu_w(p)$ Then the word's location, I_w, is:

$$I_w = \Sigma\mu_w(Pi)Pi/\Sigma\mu_w(Pi). \qquad (1)$$

A measure of overlap can also be defined for any pair of functions. This measure should range from 0 to 1, equalling 0 when there is no overlap—that is, when the two functions are distinct (as in the NN case)—and 1 when one function entirely encloses the other (as in panels b and e of Figure 4). A convenient measure that satisfies this criterion is [assuming all functions are scaled to have a maximum of $\mu(p) = 1$ for some p]:

$$O_{xy} = \mathrm{Max}\{\mathrm{Min}[\mu_x(p), \mu_y(p)]\}. \qquad (2)$$

Note that for two nonoverlapping functions x and y, $\mathrm{Min}[\mu_x(p), \mu_y(p)] = 0$ for all p, and therefore $O_{xy} = 0$. For two functions μ_x and μ_y where one function completely encloses the other, $\mathrm{Min}[\mu_x(p), \mu_y(p)] = 1$ for some p, and therefore $O_{xy} = 1$. Finally, for two partially overlapping functions μ_x and μ_y, $O_{xy} = \mu_x(p) = \mu_y(p)$ for some p, with $0 < O_{xy} < 1$. Thus, for example, the overlap for the WN case illustrated in panel c of Figure 4 is $O_{w_3p_3} = \mu_{w_3}(p_3)$, and that for the WW case illustrated in panel f of Figure 4 is $O_{w_8w_9} = \mu_{w_8}(p^*) = \mu_{w_9}(p^*)$.

It is important to realize that the distance ratio and overlap parameters are positively correlated—the closer the two stimuli are the more likely they are to overlap. Therefore, it is impossible to fully distinguish between their contributions to the total RT of the comparison. We can, however, hypothesize positive monotonic correlations of RT with both variables.

The type of relationships between the various membership functions and their degrees of overlap cannot be predicted in advance for each subject. Assuming, however, that all types of relationships are obtained, we can predict the following order for the RTs: NN < WN and WW distinct pairs < WN and WW overlapping pairs < WN and WW enclosed pairs. The rationale for this prediction is as follows: NN comparisons depend only on the distance between the terms; distinct pairs also require resolution of the phrase's (or phrases') vagueness; overlapping and enclosed comparisons also depend on the degree of overlap, which is maximal for enclosed pairs. An interesting issue is the ordering of the WW and WN pairs. We have assumed a fixed resolution time and therefore do not expect differences between the two types of distinct pairs. Also, our measure of overlap does not distinguish between the two types of comparisons, so significant differences between the two modes for enclosed pairs are not predicted. Obviously, for overlapping pairs there is no general prediction: one can think of numerous combinations of probabilities and phrases with arbitrarily high or low levels of overlap. However, given our choice of stimuli for this experiment and our expectations regarding the shape of their membership functions, we predict that the average level of overlap in WN pairs will be higher than that in WW pairs. This is because, although both words and numbers were selected to span the probability continuum, we expect high overlap only among those WW pairs made up of nonextreme words that are adjacent to each other, such as UNLIKELY, IMPROBABLE and LIKELY, PROBABLE. We anticipate low overlap for the nonadjacent WW pairs as well as for those pairs involving the end words, CERTAIN and IMPOSSIBLE. In contrast, we expect each word in a WN pair to have high overlap with at least two numerical probabilities. Consequently, on average the WN pair will require longer processing time than the WW pairs.

In the following experiment, we collected, in addition to the RTs, judgments of membership for all words involved, using a method developed by Rapoport et al. (1987), and we used these judgments to test the generalized reference point model at the individual level.

EXPERIMENT 2

Method

Subjects

Six male and 6 female native English speakers who had not participated in Experiment 1 were paid for their involvement in this experiment.

Materials

The stimuli were as in Experiment 1, with the following exceptions: TOSSUP replaced POSSIBLE, and the numerical expressions were presented in one of two formats, spelled out or digital (e.g., FIVE, or 0.05). Six subjects received each numerical format, while all 12 saw the same phrases.

Procedure

The first session started with the elicitation of membership functions for the probability expressions on an IBM Personal Computer, using a procedure similar to that of Rapoport et al. (1987). First the subject defined upper and lower bounds on the probability representations for each verbal expression of uncertainty. Following this, each word was presented three times with each of 11 probabilities, selected by the program for each subject on the basis of the established range. At least one probability at or below the lower bound and at least one probability at or above the upper bound of this range were included. The subjects' task was to indicate to what degree the expression x describes a probability y. This was achieved by moving a cursor by means of a mouse and pad, along a line marking goodness of fit, anchored at one end by "not at all" and at the other end by "perfectly." The original location of the cursor was determined randomly for each judgment. Thus, this task involved 7 words × 11 probabilities × 3 replications = 231 trials. The subjects worked at their own pace, with a break after Trial 115.

Next, the subjects were introduced to the paired-comparisons task, which was essentially the same as in Experiment 1. This first session was intended for practice only, and different stimuli (five nonnumerical and five numerical expressions) were used. Each comparison was replicated three times. The subjects who did not meet the criterion of less than 14 inconsistent judgments were invited to repeat this practice session before commencing the series of four experimental sessions.

Table 4
Shape Classification of 84 Membership Functions (12 Subjects × 7 Phrases)

	Monotonic		Single		
	Decreasing	Increasing	Peaked	Other	Crisp
CERTAIN	0	9	1	0	2
LIKELY	0	5	7	0	0
PROBABLE	0	2	9	1	0
TOSSUP	1	1	10	0	0
IMPROBABLE	2	0	10	0	0
UNLIKELY	2	0	7	3	0
IMPOSSIBLE	3	1	4	1	3
Total	8	18	48	5	5
%	9.5%	21.4%	57.1%	6.0%	6.0%

The four experimental sessions were run on separate days. Two of the sessions employed the instructions "choose higher probability" and two employed "choose lower probability" instructions, with counterbalancing employed to achieve six orders of presentation. Two subjects performed the task in each of these six orders. Each paired comparison was presented three times; the left-right order of presentation was balanced across sessions within instructions. Each term, then, within each pair, could be chosen as the higher 0, 1, 2, or 3 times in each session.

Results

Membership Functions

The reliability of the judgment of membership was measured for each subject (across all phrases) and for each phrase (across all subjects) by the median Pearson correlation among the three replications. The overall reliability across subjects and phrases was 0.79. At the subject level, reliabilities ranged from 0.55 to 0.93 with a median of 0.71, and at the phrase level from 0.64 to 0.86 with a median of 0.75.

Table 4 presents a shape classification of the 84 membership functions. Consistent with previous data (e.g., in Wallsten, Budescu, Rapoport, et al., 1986) most of the functions were single-peaked (57.1%). Monotonically increasing and decreasing functions occurred mostly for high and low probability words, respectively. For some subjects, the end terms IMPOSSIBLE and CERTAIN have crisp functions (6.0%).

Analysis of Response Times

RTs were collected for 13,104 trials. The 170 outliers (1.3%; RTs greater than 6,000 msec or lower than 400 msec) were replaced by RTs sampled from normal distributions, with the means and standard deviations estimated from the remaining data separately for each pair of terms, under each instruction set, for each subject. The effects on the rate of responding of the independent variables—instruction, digital/spelled-out, and type of judgment (WW, NN, or WN)—are shown in Table 5. Replicating the effects in Experiment 1, the following main effects were significant: session [$F(3,30) = 14.22$, $p < .01$], illustrating that practice effects were still not eliminated entirely; instruction [$F(1,10) = 7.73$, $p < .02$], with faster rates of response under "choose higher probability" instructions (0.978 vs. 0.898); and type [$F(2,20) = 136.69$, $p < .01$], with fastest rates of response for NN (1.194 vs. 0.895 for WW and 0.846 for WN).

An interesting interaction between digital/spelled-out × type [$F(2,20) = 4.80$, $p < .02$] was uncovered. Namely, the digital group not only responded faster to NN pairs, as was anticipated, but also outperformed the spelled-out group on WW and WN pairs, to a lesser extent. Tests of simple main effects (Kirk, 1982) revealed a significant advantage to the digital group only for the NN and WN comparisons, but not for the WW pairs. This pattern is perfectly reasonable, since the nonnumerical terms were identical in the two groups.

The distance effect. Ranks were estimated from TVC of paired comparisons per session and were averaged across sessions, since there were high intersession correlations (median WW correlation = 0.985, and median WN correlation = 0.989). Tied ranks were solved by ranking the probabilities obtained from the membership function data for WW, and comparing these W "probabilities" with N probabilities for WN where appropriate.

Distance effects were significant for all three types: NN [$F(5,55) = 25.31$, $p < .01$], WW [$F(5,55) = 45.07$, $p < .01$], and WN [$F(12,117) = 24.36$, $p < .01$]. For WW and WN, there was also an effect of instruction set [NN, $F(1,11) = 1.60$, $p > .05$; WW, $F(1,11) = 9.97$, $p < .01$; WN, $F(1,11) = 7.30$, $p < .05$]; namely, responses were made faster under "choose higher" instructions. In addition, for WW and WN there was a distance × instruction interaction [NN, $F(5,55) = 0.99$, $p > .05$; WW, $F(5,55) = 2.91$, $p < .05$; WN, $F(12,117) = 2.34$, $p < .01$], reflecting the amplified ad-

Table 5
Median Rate of Response and Standard Error of Mean as a Function of Instruction, Type of Judgment, and Group: Experiment 2

	Rate (per second)								
	Digits			Spelled Out			All		
Instruction	NN	WN	WW	NN	WN	WW	NN	WN	WW
Lower Rate	1.321	0.898	0.947	1.026	0.709	0.735	1.173	0.804	0.841
SEM	.014	.011	.017	.012	.007	.010	.010	.007	.010
Higher Rate	1.424	1.023	1.084	1.004	0.754	0.815	1.214	0.889	0.950
SEM	.013	.010	.017	.011	.008	.013	.010	.007	.011
Across Rate	1.373	0.961	1.016	1.015	0.731	0.775	1.194	0.846	0.895
SEM	.009	.007	.012	.008	.005	.008	.007	.005	.008

vantage of "choose higher" instructions at the maximal distance—that is, in pairs including one high term. The median response rates are displayed in Figure 5.

The congruity effect. The interaction of instruction set and pair size is presented graphically in Figure 6, for each type of judgment. The interaction was significant for each type, illustrating congruity effects [NN, $F(1,11) = 5.16$, $p < .05$; WW, $F(1,11) = 32.38$, $p < .01$; and WN, $F(1,11) = 24.89$, $p < .01$]. For WW and WN, there were also main effects of instruction set [NN, $F(1,11) = 0.90$, $p > .05$; WW, $F(1,11) = 10.83$, $p < .01$; and WN, $F(1,11) = 11.29$, $p < .01$] and pair size [NN, $F(1,11) = 2.19$, $p > .05$; WW, $F(1,11) = 30.88$, $p < .01$; and WN, $F(1,11) = 14.57$, $p < .01$].

In all three types, the facilitating effect of "choose higher probability" instructions on high pairs was greater than that of "choose lower probability" instructions on low pairs. All effects were in the expected directions, and in the form of crossovers.

Testing the Model

As in Experiment 1, all model analyses were performed on RTs, not rates, to correspond to the metric of the model. To eliminate session effects and individual differences, RTs were first standardized within each session and each subject. Subsequently, the median standardized RT over sessions for each paired comparison provided the data for model testing. Thus, for each subject there are 182 data points (91 comparisons under two instructions), each based on six judgments. One subject was eliminated from this analysis because her paired comparisons in the third session were inconsistent with those in the other three sessions, suggesting that she misunderstood the nature of the task.

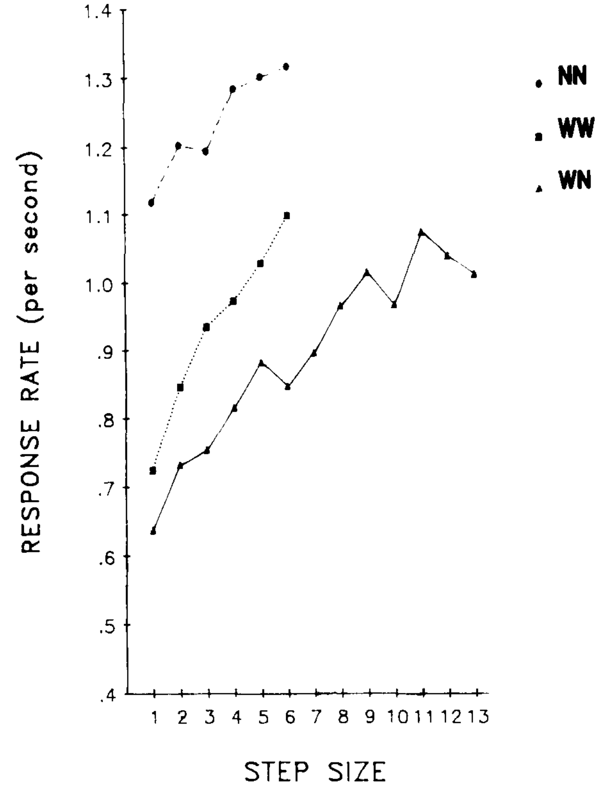

Figure 5. Median response rate as a function of step size (distance) between paired terms: Experiment 2.

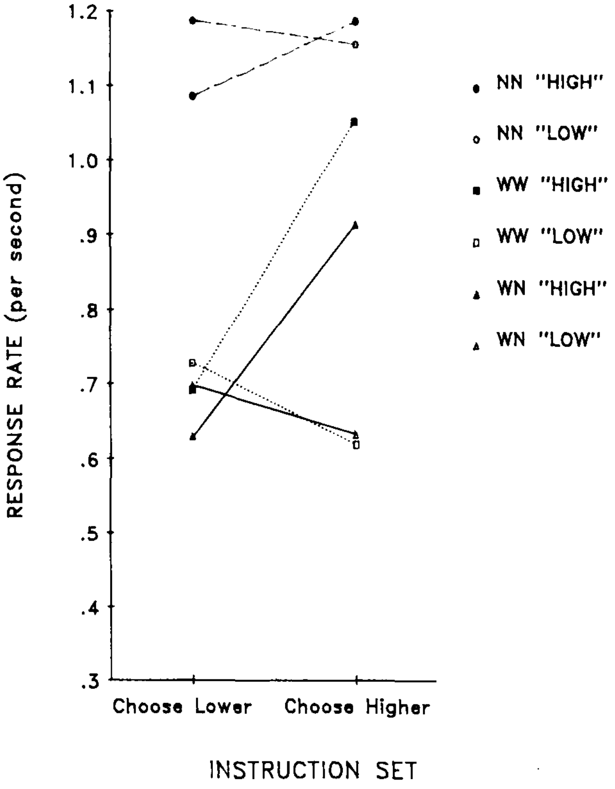

Figure 6. Median response rate as a function of instruction set and pair size: Experiment 2.

Table 6 presents the correlations between median RT and the various components of the generalized model. The first major column in the table, labeled "Total," shows correlations calculated when pair type is ignored, while the remaining major columns show correlations within the separate pair types. The component "Ratio" is the ratio of the distance of each term in the pair from the endpoint implied by the question, where the distance is based on the numerical value for numbers, and on the location value from Equation 1 for phrases. "Overlap" is calculated from Equation 2; "Resolution" equals 0 (for NN) or 1 (for WN and for WW). As expected, all the correlations are positive and statistically significant. Of course, ratio and overlap are highly intercorrelated.

In Table 7 we summarize the quantitative fit of the model for each subject and the group as a whole. For each case we present the R^2 associated with models based on (1) the distance ratio alone (i.e., Holyoak's original model), (2) the overlap alone, (3) the distance and the overlap parameter, and (4) the full model consisting of

Table 6
Correlations of Median Reaction Time with Variables in Model

	Total (N = 182)			NN (n = 42)	WN (n = 98)		WW (n = 42)	
Subject	Ratio	Overlap	Resolution	Ratio	Ratio	Overlap	Ratio	Overlap
02	62	65	45	45	66	63	68	59
03	56	45	42	60	63	41	51	50
04	60	61	40	63	66	65	69	52
05	58	63	45	54	59	59	73	60
06	50	54	42	40	52	45	70	64
07	44	53	52	15	50	41	39	35
08	49	51	35	45	36	37	75	72
09	52	54	38	40	57	51	49	45
11	54	66	43	33	63	70	60	38
12	44	61	42	51	46	60	66	45
13	39	67	60	24	45	65	67	67
Median	52	61	42	45	57	59	67	52

Note—Decimal point omitted.

the ratio, overlap, and resolution parameters. In this case too we avoid tests of significance based on dependent observations. It is clear, however, that the full model outperforms Holyoak's model (by 70% at the individual level and 67% at the group level). It is also interesting to note that the overlap measure alone is a better single predictor of the RT than the distance ratio, but that despite the correlation between the two, their joint effect (column 3) increases the fit of the model for 8 of the 11 subjects. These results are based on the raw judgments of the membership values. In an additional analysis, we fitted a smooth function (a cubic polynomial) to the membership functions for each word and used these fitted values in the test of the model. The results were practically identical to those reported in Table 7.

The model predicts that WN comparisons take longer than WW because the overlap between the former terms is greater than that between the latter. To test this hypothesis we compared the average level of overlap between the terms in the two types of comparisons. In Table 8 we present these values. For all subjects, the direction of this difference conforms with the prediction; and at the group level, a two-way (subject × type) ANOVA reveals that the degree of overlap is significantly higher in WN pairs (56.478 vs. 41.602) [$F(1,10) = 32.74$, $p < .01$].

Finally, we tested the prediction regarding the order of the RTs under the various patterns of relationships between the stimuli. Unfortunately, there were no cases of enclosed membership functions, and only four subjects (6, 11, 12, and 13) had a sufficiently high proportion of distinct functions to apply the test. Thus this analysis is based only on partial data. In Table 9 we present two sets of median response rates and standard errors. The first one is based on the raw data, and the second is "adjusted" for the distance ratio in an Analysis of Covariance (ANACOVA). Obviously, the means are ordered as predicted, and they are significantly different in the ANACOVA [$F(4,12) = 45.38$, $p < .05$]. Multiple comparisons of the means cluster them in three groups: NN are significantly quicker than all others; WN with overlap are significantly slower than all others; and the remaining three conditions do not differ. However, the contrast comparing WW and WN without overlap with their counterparts with overlap is significant. Thus, the ordering hypothesis was supported in a test based only on 4 subjects. We attribute the failure to find a significant difference between WW with overlap and the two conditions without overlap to insufficient power in our test.

Table 7
Goodness of Fit (R^2) of the Class of Models to the Response Time in Experiment 2

Subject	Ratio	Overlap	Ratio + Overlap	Ratio + Overlap + Resolution
02	38	42	50	54
03	31	20	34	41
04	35	37	46	51
05	33	40	45	50
06	25	29	34	46
07	19	28	28	40
08	23	26	32	35
09	27	29	35	37
11	29	44	47	54
12	19	37	38	45
13	15	44	44	64
Median	27	37	38	46

Note—Decimal point omitted.

Table 8
Mean Level of Overlap Across Terms in WN and WW Comparisons

	Mean Level of Overlap	
Subject	WN	WW
02	62.2	51.7
03	80.6	74.8
04	47.4	40.2
05	68.1	58.1
06*	44.1	15.5
07	57.7	43.6
08	51.9	39.0
09	66.5	56.2
11	50.8	35.7
12	43.0	27.4
13*	48.7	15.5
All*	56.5	41.6

*$p < .05$.

Table 9
Median Rate of Response and Standard Error of Mean in Five Classes of Comparisons

	Type of Comparison				
		No Overlap		With Overlap	
	NN	WN	WW	WN	WW
n	336	152	176	276	160
Unadjusted Means	1.14	0.80	0.86	0.62	0.77
SEM	0.01	0.02	0.02	0.02	0.02
Adjusted Means	1.12	0.82	0.84	0.68	0.75
SEM	0.02	0.03	0.04	0.03	0.04

Note—Data from 4 subjects only.

GENERAL DISCUSSION

Response rate was greater, and less variable, in Experiment 2 than in Experiment 1, as a result of the longer practice and increased number of replications, but otherwise, the pattern of results was identical in the two experiments. That is, NN judgments were significantly faster than the WW and WN judgments, which were not significantly different from each other; response rates became faster in later sessions, indicating the continuing effects of practice; responses to "choose higher probability" were significantly faster than to "choose lower probability"; and of particular importance, both the distance effect and the congruity effect were replicated. Specifically, response rate increased monotonically with ordinal distance between the members of the pair being judged, and the rate was greater when the question was congruent with the magnitude of the stimuli (i.e., "select the greater of two high stimuli or the lesser of two low stimuli") than when the converse was true. Finally, in addition, subjects who saw the numbers in digit form rather than spelled out tended generally to respond faster than did the other subjects in the NN and WN conditions.

The membership functions obtained in Experiment 2 were similar to those found in other studies. That is, the level of reliability, the shapes of the functions, and the degree of individual differences in the functions are consistent with those obtained by Rapoport et al. (1987) and by Wallsten, Budescu, Rapoport, et al. (1986).

However, the most important finding in Experiment 2 is the good level of support obtained with the aid of the membership functions for the generalized reference point model at the individual level. The generalized model differs from the original one in that it explicitly acknowledges the possibility that probability expressions (or more generally, magnitude expressions) are differentially vague, whereas numbers are relatively precise. Consequently, the ease with which a comparison is made depends not only on how close two stimuli are but also on the degree to which they overlap. Indeed, Table 7 suggests that although distance and overlap are correlated for most subjects (all but Subject 3), overlap alone is a better predictor of response time than is distance ratio alone. Further, for 8 of the subjects (excluding Subjects 7, 12, and 13) predictability is enhanced when both variables are employed rather than either one alone. Although statistical model comparisons are inappropriate here, due to the nonindependence of the data points, the fact that the conclusions hold descriptively for most of the subjects is noteworthy. The robustness of the model was demonstrated by the secondary analyses, using fitted functions and slightly different parameterizations of location and overlap.

If distance and overlap alone were sufficient to explain the results, then the NN, WN, and WW response times could be handled by a multiple regression model with only those two variables as predictors. However, a third dummy variable, distinguishing whether or not both of the stimuli are numeric (an initial resolution of the vagueness is necessary), improves the fit for all subjects by amounts ranging from 3% to 20%. Thus, we are forced to conclude that the verbal-numerical distinction is not accounted for solely by differential vagueness as indicated by overlap.

In the remainder of this discussion, we first present a process model consistent with the results in Table 7 as well as with Holyoak's (1978) original reference point model and with the ν-μ model of judgment given vague information, described by Budescu et al. (1988) and Wallsten, Budescu, and Erev (in press). Finally, we relate additional aspects of the present results to the generalized reference point model.

According to Holyoak's original reference point model, described earlier in this paper, stimuli are represented internally by distributions. When comparing two stimuli, the subject draws a sample from each distribution, computes the distance of that sample from a reference point implied by the question, calculates the ratio of those two distances, and responds if the ratio exceeds a threshold. Otherwise, the sampling and computation process is repeated and the result is cumulated with that of the first sampling. This process continues until a threshold is crossed and a decision is made.

Wallsten et al. (in press) have proposed a model similar in spirit, but different in detail from that of Holyoak (1978). The model was developed to handle a variety of findings concerning the vague meanings of probability phrases, and it was tested in a series of studies (Wallsten & Erev, 1988) involving untimed choices between linguistic and numerical goals. On each trial subjects were faced with two binary gambles involving identical outcomes, one of which was always zero and the other of which was a positive or a negative amount. In one gamble the probability was represented on a spinner, while in the other gamble it was stated verbally. The choice situation, then, was similar to the WN case here, except that judgments were untimed, they were worth money, and the nature of the judgment was implied by the task rather than specifically asked. The model was reasonably successful in predicting choice probabilities from individual membership functions. According to this model, the vague meaning of a linguistic probability expression to an individual is represented by his or her membership function μ over the probability interval for that phrase. When required to make a choice, the individual chooses in accordance with a specific probability value whose membership is sufficiently high, that is, above a threshold ν. The probability value employed is randomly selected according to a weighting function that depends on the membership values above ν. The mathematical details for this model can be found in Wallsten et al. (in press), or Wallsten and Erev (1988).

Combining essential ideas from the two models, we might assume that internal sampling distributions for verbal probabilities are derived from membership function values above a threshold. Those for probability numbers may also arise from a membership function of sorts, but in any case they are relatively tight or possibly even point distributions. If this interpretation is correct, then the location values (from Equation 1) used to compute the distance ratios represent the expected values of the phrase distributions under the assumption that the threshold is set at zero. (Assuming thresholds at various values between zero and one, we computed distance ratios with essentially no effect on the results.) The greater the degree of overlap between two distributions, the greater would be the need for repeated sampling and, of course, the higher the response time. The dummy variable necessary in the multiple regression may reflect additional time needed to bring the vague representation of the phrases to mind or to set a threshold.

This interpretation of the generalized reference point model allows explanation of two otherwise puzzling results. First, WN judgments were generally (although not significantly) slower than WW judgments. A priori it would seem that judgments would be quicker when at least one of the stimuli is crisp than when both are vague. However, the result can be attributed to the fact, illustrated in Table 8, that WN membership function overlap was greater than WW membership function overlap. This interpretation is substantiated by comparisons within subjects where possible (Table 9). Here WN and WW judgments without overlap are significantly faster than corresponding judgments with overlap.

Thus, the reference point model generalized to include sampling distributions based on membership functions appears to provide a parsimonious explanation of all the present results in both experiments, except for the fact of faster judgments to "choose higher" than to "choose lower." As suggested earlier, that result may be attributed to a congruity effect based on the form in which probability relationships are generally stored.

REFERENCES

Banks, W. P., & Flora, J. (1977). Semantic and perceptual processes in symbolic comparisons. *Journal of Experimental Psychology: Human Perception & Performance*, **3**, 278-290.

Banks, W. P., Fujii, M., & Kayra-Stuart, F. (1976). Semantic congruity effects in comparative judgments on magnitudes of digits. *Journal of Experimental Psychology: Human Perception & Performance*, **2**, 435-447.

Besner, D., & Coltheart, M. (1979). Ideographic and alphabetic processing in skilled reading of English. *Neuropsychologia*, **17**, 467-472.

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

Buckley, P. B., & Gillman, C. B. (1974). Comparisons of digits and dot patterns. *Journal of Experimental Psychology*, **103**, 1131-1136.

Budescu, D. V., & Wallsten, T. S. (1985). Consistency in interpretation of probabilistic phrases. *Organizational Behavior & Human Decision Processes*, **36**, 391-405.

Budescu, D. V., & Wallsten, T. S. (1987). Subjective estimation of precise and vague uncertainties. In G. Wright & P. Ayton (Eds.), *Judgmental Forecasting* (pp. 63-82). Chichester, England: Wiley.

Budescu, D. V., Weinberg, S., & Wallsten, T. S. (1988). Decisions based on numerically and verbally expressed uncertainties. *Journal of Experimental Psychology: Human Perception & Performance*, **14**, 281-294.

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

Clark, H. H. (1969). Linguistic processes in deductive reasoning. *Psychological Review*, **76**, 387-404.

Coombs, C. H. (1964). *A theory of data*. New York: Wiley.

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

Foltz, G., Poltrock, S., & Potts, G. (1984). Mental comparison of size and magnitude: Size congruity effects. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **10**, 442-453.

Holyoak, K. J. (1978). Comparative judgments with numerical reference points. *Cognitive Psychology*, **10**, 203-243.

Holyoak, K. J., & Patterson, K. K. (1981). A positional discriminability model of linear-order judgments. *Journal of Experimental Psychology: Human Perception & Performance*, **7**, 1283-1302.

Holyoak, K. J., & Walker, J. H. (1976). Subjective magnitude information in semantic orderings. *Journal of Verbal Learning & Verbal Behavior*, **15**, 287-299.

Jamieson, D. G., & Petrusic, W. M. (1975). Relational judgments with remembered stimuli. *Perception & Psychophysics*, **18**, 373-378.

Johnson, E. M. (1973). *Numerical encoding of qualitative expressions of uncertainty*. (Technical Paper No. 250). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences*. Monterey, CA: Brooks Cole Publishing.

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

Moyer, R. S., & Bayer, R. H. (1976). Mental comparison and the symbolic distance effect. *Cognitive Psychology*, **8**, 228-246.

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgments of numerical inequality. *Nature*, **215**, 1519-1520.

Paivio, A. (1975). Perceptual comparisons through the mind's eye. *Memory & Cognition*, **3**, 635-647.

Parkman, J. M. (1971). Temporal aspects of digit and letter inequality judgments. *Journal of Experimental Psychology*, **91**, 191-205.

Potts, G. R. (1974). Storing and retrieving information about ordered relationships. *Journal of Experimental Psychology*, **103**, 431-439.

Rapoport, A., Wallsten, T. S., & Cox, J. A. (1987). Direct and indirect scaling of membership functions of probability phrases. *Mathematical Modelling*, **9**, 397-417.

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

Saaty, T. L. (1980). *The analytic hierarchy process*. NY: McGraw-Hill.

Takane, Y., Young, F. W., & de Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika*, **42**, 7-67.

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

Wainer, H. (1977). Speed vs reaction time as a measure of cognitive performance. *Memory & Cognition*, **5**, 278-280.

Wallsten, T. S., Budescu, D. V., & Erev, I. (in press). Understanding and using linguistic uncertainties. *Acta Psychologica*.

Wallsten, T. S., Budescu, D. V., Rapoport, A., Zwick, R., & Forsyth, B. (1986). Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General*, **115**, 348-365.

Wallsten, T. S., Budescu, D. V., & Zwick, R. (1986, November). *On the representation and use of linguistic probabilities in judgment and decision making*. Paper presented at the Annual Meeting of the Judgment/Decision Making Society, New Orleans, LA.

Wallsten, T. S., & Erev, I. (1988). *Choosing between linguistic and precise gambles: A test of a theory of choice given vague information*. Manuscript in preparation.

Wallsten, T. S., Fillenbaum, S., & Cox, J. A. (1986). Base rate effects on the interpretations of probability and frequency expressions. *Journal of Memory & Language*, **25**, 571-587.

Wallsten, T. S., Zwick, R., Kemp, S., & Budescu, D. V. (1988). *Some factors affecting preference for verbal and numerical communication of uncertainty*. Manuscript in preparation.

Yager, R. R. (1981). A procedure for ordering fuzzy sets on the unit interval. *Information Sciences*, **24**, 143-161.

Zimmer, A. C. (1984). A model for the interpretation of verbal predictions. *International Journal of Man-Machine Studies*, **20**, 121-134.

(Manuscript received April 14, 1987;
revision accepted for publication August 23, 1988.)