Journal of Experimental Psychology: General
1986, Vol. 115, No. 4, 348-365

0096-3445/86/$00.7

Measuring the Vague Meanings of Probability Terms

Thomas S. Wallsten
University of North Carolina at Chapel Hill

David V. Budescu and Amnon Rapoport
University of Haifa, Israel

Rami Zwick and Barbara Forsyth
University of North Carolina at Chapel Hill

SUMMARY

Can the vague meanings of probability terms such as *doubtful, probable,* or *likely* be expressed as membership functions over the [0, 1] probability interval? A function for a given term would assign a membership value of zero to probabilities not at all in the vague concept represented by the term, a membership value of one to probabilities definitely in the concept, and intermediate membership values to probabilities represented by the term to some degree. A modified pair-comparison procedure was used in two experiments to empirically establish and assess membership functions for several probability terms. Subjects performed two tasks in both experiments: They judged (a) to what degree one probability rather than another was better described by a given probability term, and (b) to what degree one term rather than another better described a specified probability. Probabilities were displayed as relative areas on spinners. Task a data were analyzed from the perspective of conjoint-measurement theory, and membership function values were obtained for each term according to various scaling models. The conjoint-measurement axioms were well satisfied and goodness-of-fit measures for the scaling procedures were high. Individual differences were large but stable. Furthermore, the derived membership function values satisfactorily predicted the judgments independently obtained in task b. The results support the claim that the scaled values represented the vague meanings of the terms to the individual subjects in the present experimental context. Methodological implications are discussed, as are substantive issues raised by the data regarding the vague meanings of probability terms.

Most people, including expert forecasters, generally prefer communicating their uncertain opinions with nonnumerical terms such as *doubtful, probable, slight chance, very likely,* and so forth, rather than with numerical probabilities. On anecdotal grounds, the imprecision of nonnumerical terms is preferred to the precision of probability numbers for at least two reasons: First, opinions are generally not precise and therefore, the claim goes, it would be misleading to represent them precisely. For example, commenting that numbers denote authority and a precise understanding of relations, a committee of the U.S. National Research Council wrote with regard to formal risk assessments that there is an

> important responsibility not to use numbers, which convey the impression of precision, when the understanding of relationships is indeed less secure. Thus, while quantitative risk assessment facilitates comparison, such comparison may be illusory or misleading if the use of precise numbers is unjustified. (National Research Council Governing Board Committee on the Assessment of Risk, 1981, p. 15)

The second reason frequently suggested for communicating with nonnumerical terms rather than with probability numbers is that most people feel they better understand words than numbers. Zimmer (1983) pointed out that it was not until the 17th century that probability concepts were formally developed, yet expressions for different degrees of uncertainty existed in many languages long before then. He suggested that people generally handle uncertainty by means of verbal expressions and their associated rules of conversation, rather than by means of numbers.

The dual claims that vague opinions are well communicated with probability expressions and that people more naturally think about uncertainty in a verbal than in a numerical manner, can be investigated only after methods have been developed for validly measuring the vague meanings of probability terms.

This research was supported by Contract MDA 903-83-K-0347 from the U.S. Army Research Institute for the Behavioral and Social Sciences to the L. L. Thurstone Psychometric Laboratory, University of North Carolina at Chapel Hill. The views, opinions, and findings contained in this paper are those of the authors and should not be construed as an official Department of the Army position, policy, or decision.

Barbara Forsyth is now at Ohio University. We thank Samuel Fillenbaum for numerous helpful discussions throughout the course of the work, and James Cox, Brent Cohen, Samuel Fillenbaum, and Jaan Valsiner for comments on a previous draft of this article.

Correspondence concerning this article should be addressed to Thomas S. Wallsten, L. L. Thurstone Psychometric Laboratory, Davie Hall 013A, University of North Carolina, Chapel Hill, North Carolina 27514.

Recognizing that the meanings of words are subject to individual differences and numerous context factors, the research presented in this article is primarily methodological, aimed at developing suitable measurement techniques and at making preliminary statements about probability terms. If procedures for validly measuring vague meanings can be established, they can be used to investigate the many substantive issues.

In most of the empirical work to date on the meaning of probability words, subjects have been asked to give numerical equivalents to various probability phrases. The overwhelming result has been that there is great intersubject variability in the numerical values assigned to probability terms and great overlap among terms (Bass, Cascio, & O'Connor, 1974; Beyth-Marom, 1982; Budescu & Wallsten, 1985; Foley, 1959; Johnson, 1973; Lichtenstein & Newman, 1967; Simpson, 1944, 1963). Within-subject variability in the assignment of numbers to probabilistic terms is not minor, but is considerably less than between-subjects variability (Beyth-Marom, 1982; Budescu & Wallsten, 1985; Johnson, 1973). However, neither the within- nor the between-subjects variability alone can be taken as evidence that probability terms have vague meanings. First of all, as pointed out by Budescu and Wallsten (1985), there is no way to determine whether the variability is due to differences between subjects, or within subjects over time, in the use of numbers rather than in the use of words. Second, and more to the present point, as Rubin (1979) noted in a related context, these data can be interpreted either as showing that the meanings of probability terms are not constant over people or times or that the expressions have generally vague meanings. An alternative approach is therefore necessary.

Membership Functions

Several people (e.g., Watson, Weiss, & Donnell, 1979; Zadeh, 1975; Zimmer, 1983) have suggested that the meaning of a probability term can be represented by a function on the [0, 1] probability interval, as illustrated in Figure 1. The function takes its minimum value, generally zero, for probabilities that are not at all in the concept represented by the phrase. It takes its maximum value, which is generally one, for probabilities definitely in the concept, and intermediate values for probabilities with intermediate degrees of memberships in the concept represented by the term. There are no constraints on the shapes such functions can have, nor must they be expressable by equations of any particular sort. Within fuzzy set theory, such a function is called a membership function, but it is not necessary to tie this idea strictly to fuzzy set theory.

Of course, the question of defining and measuring the vague meaning of a term arises in a vast array of semantic domains, and the concept of a membership function has been applied quite broadly within fuzzy set theory (e.g., Norwich & Turksen, 1984; Zadeh, 1975; Zysno, 1981). As a general definition, a membership function is a rule that assigns to each element in the universe of discourse a number in the closed [0, 1] interval indicating the degree to which that element is a member of a particular set or category. If the category is well defined (e.g., male humans beyond their 60th birthday), then all membership functions are either 0 or 1. If the category is not well defined

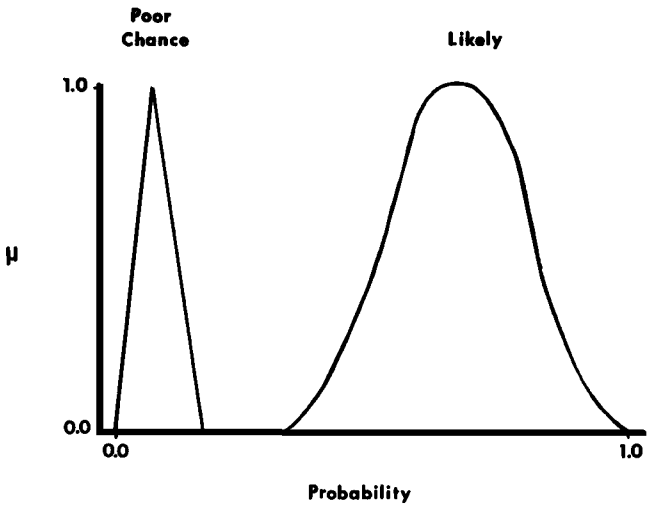

Figure 1. Hypothetical membership functions for two probability terms.

(e.g., middle-aged men), then the membership functions can take on any value in the [0, 1] interval.

Measurement of Vague Meanings

A considerable literature exists on the topic of vagueness (e.g., Ballmer & Pinkal, 1983; Gaines & Kohout, 1977; Goguen, 1969; Hempel, 1939; Hersh & Caramazza, 1976; Labov, 1973; Oden, 1981; Skala, Termini, & Trillas, 1984; Zadeh, 1965). However, although much has been written about the measurement of vagueness or fuzziness, empirical work has been relatively sparse. One method relies on choice probabilities. For example, a stimulus, such as a square, is presented along with a word such as *small* (Hersh & Caramazza, 1976; Hersh, Caramazza, & Brownell, 1979). The subject answers *yes* or *no* according to whether the word describes the stimulus. The fraction of *yes* responses over subjects or within subjects over trials is then taken as the degree of membership for that stimulus in the vague concept represented by the word. Rubin (1979) has criticized this procedure because (a) it confounds measures of fuzziness with response variability that is due to experimental procedures, and (b) it can just as well be interpreted as showing that words have different meanings to different people or at different times as that words have vague or fuzzy meanings.

A second method of obtaining membership functions is direct scaling, in which subjects rate stimuli on a scale from *definitely in the concept* to *definitely not in the concept.* For example, Oden (1977b) had subjects rate propositions on a scale from *absolutely true* to *absolutely false.* Similarly, Zysno (1981) had subjects rate grade of membership on a scale from 0% to 100% of a man X years of age in concepts such as *old man, very young man,* and so forth, for various values of X (see also MacVicar-Whelan, 1978). In other studies (e.g., Kuz'Min, 1981), subjects picked stimuli with specified grades of membership. The direct-scaling methods overcome some of the problems with the choice probabilities, in that the construct of vagueness is directly assessed in individual responses. However, as with all magnitude

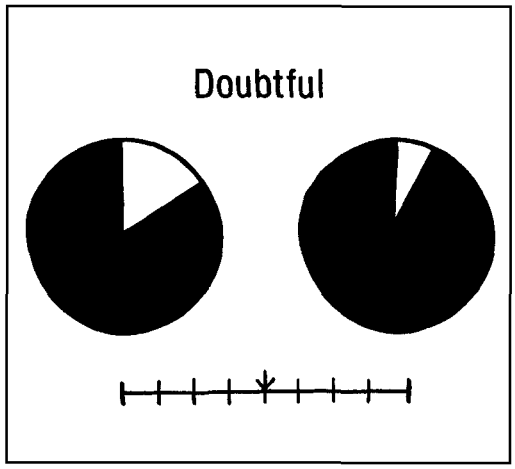

Figure 2. Sample experimental scenario.

estimation procedures, the responses cannot be evaluated unless they are embedded within a theory. Oden used functional measurement techniques to assess his measures; many other authors simply have displayed the estimates after they are obtained (e.g., Norwich & Turksen, 1984) or have fitted them with explicit functions that are evaluated by means of goodness-of-fit measures (e.g., Zysno, 1981).

We use a different approach, which utilizes a modified pair-comparison method for measuring the vague meanings of probability terms. Empirically, the procedure is similar to one utilized by Oden (1977a), but the data are analyzed much differently. The data can first be checked at an ordinal level to determine if they satisfy certain axioms necessary for scaling vagueness according to an algebraic difference (or ratio) model (Krantz, Luce, Suppes, & Tversky, 1971). If the axioms are reasonably well satisfied, then specific difference- or ratio-scaling procedures (Saaty, 1977, 1980; Torgerson, 1958) can be applied to the data for the purpose of deriving the vagueness measure, or membership function, for each expression. Furthermore, goodness-of-fit measures can be calculated to evaluate the quality of the metric scaling.

A Pair-Comparison Method

Consider a sample experimental trial as shown in Figure 2. Two spinners are drawn on a computer monitor. Subjects are told to imagine a pointer over each spinner that can be spun so that it randomly lands over either the white or the dark sector. Thus, each spinner displays a different probability of the pointer landing on white. There is a probability term printed above the spinners and a line with an arrow on it below them. The subject must move the arrow on the line to indicate for which spinner the probability of landing on white is better described by the probability term and how much better it is described. Moving the arrow to the far left indicates that the left spinner is absolutely better described, leaving the arrow in the middle indicates that the two spinners are equally well described, and so forth. The probabilities on the two spinners are changed from trial to trial according to a left side by right side, $P \times P$, factorial design in which $P = \{p_i, \ldots, p_n\}$, where for $i = 1, \ldots, n$, the p_i denotes specific probabilities of the spinners landing on white.

Consider the bounded response line shown in Figure 2 to extend from 1 on the left to 0 on the right and let $R_W(ij)$ be the response when probability p_i is on the left, p_j is on the right, and expression W is displayed above them. The responses $R_w(ij)$ induce an ordering on the factorial design according to the degree that the left hand probability is better described by the term than is the right hand probability. If, as will be described, this ordering satisfies the axioms of an algebraic difference structure (Krantz, et al., 1971), then a suitable transformation of the cell entries can be used in a difference or a ratio scaling model to establish a membership function for the term W, such as is shown in Figure 1.

A bit of notation will aid in making these concepts clear. Let p_ip_j refer to a cell in the $P \times P$ factorial design, or in other words, be an element in the Cartesian product of $P \times P$. The cells of the factorial design are rank ordered according to how much better phrase W describes the left-hand probability than the right-hand probability. The rank ordering between any pair of cells is denoted by $\succsim_W$ where the subscript indicates that the ordering is for the particular phrase (*doubtful* in Figure 2). Stated formally,

$$p_ip_j \succsim_W p_kp_l \quad \text{iff} \quad R_W(ij) \geq R_W(kl). \tag{1}$$

Let $(P \times P, \succsim_W)$ refer to an ordered matrix of the sort just described. Krantz et al. (1971) proved that if $(P \times P, \succsim_W)$ satisfy five axioms, then there exists a mapping μ_W from P into the real numbers such that

$$p_ip_j \succsim_W p_kp_l \quad \text{iff} \quad \mu_W(p_i) - \mu_W(p_j) \geq \mu_W(p_k) - \mu_W(p_l), \tag{2}$$

or, equivalently, such that

$$p_ip_j \succsim_W p_kp_l \quad \text{iff} \quad \mu_W(p_i)/\mu_W(p_j) \geq \mu_W(p_k)/\mu_W(p_l). \tag{3}$$

In other words, scale values can be assigned to these probabilities such that the rank order of differences (or of ratios) in the assigned values matches the rank order of differences (or of ratios) in the degrees to which the left-hand and right-hand probabilities are described by the phrase. The scale values are unique up to a linear (for the difference representation) or a power (for the ratio representation) transformation. These scale values, normalized to be nonnegative with an arbitrary maximum of 1, and plotted as a function of the probabilities (as illustrated in Figure 1) can be taken as the membership function representing the degree to which each probability belongs to the vague concept defined by the expression.

It should be noted that at an axiomatic level, the difference and ratio representations cannot be distinguished unless different orderings appear under difference- and ratio-inducing conditions (see Birnbaum, 1980, and Miyamoto, 1983). This is because any set of differences can be mapped into a set of ratios by taking logs, and conversely, any set of ratios can be mapped into a set of differences by exponentiating.

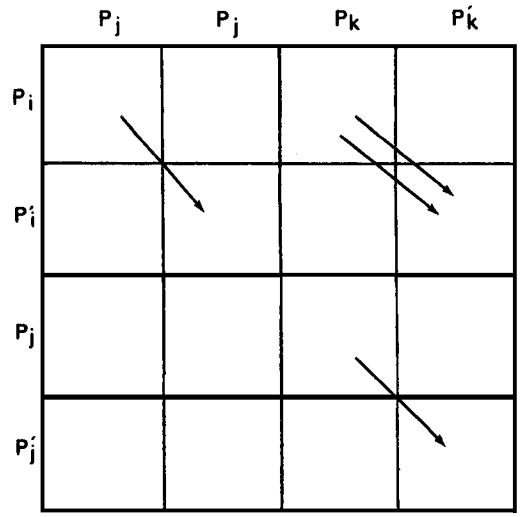

Figure 3. Illustration of the weak monotonicity axiom.

Tests of the Axioms

The five axioms specified by Krantz et al. (1971) include two that are of purely mathematical interest and three that can be subjected to empirical test. One of these, the weak order axiom, states that all the elements of $P \times P$ can be compared to each other and that the ordering is transitive. Our method of using the arrow location to rank order the matrix forces this axiom to be satisfied, and therefore it is not of empirical interest here. However, the remaining two axioms, sign reversal and weak monotonicity, can be evaluated.

The *weak monotonicity axiom* is illustrated in Figure 3. It states that for all $p_i, p_j, p_k, p_{i'}, p_{j'}, p_{k'} \in P$, if $p_ip_j \succeq_W p_{i'}p_{j'}$ and $p_jp_k \succeq_W p_{j'}p_{k'}$, then $p_ip_k \succeq_W p_{i'}p_{k'}$. Single arrows in Figure 3 indicate the antecedent conditions and the double arrow indicates the consequent.

The monotonicity axiom can be evaluated separately within the $P \times P$ matrix associated with each term. This is done by selecting suitable subsets of six cells within the matrix and then for all those subsets for which the antecedent conditions are met, checking to determine whether the consequent condition is also met. The number of subsets available for test depends on the size of the matrix and can be substantial. Of course, there is considerable overlap among the subsets, and therefore the tests are not independent. A convenient summary statistic for each matrix is the percentage of possible tests that are satisfied.

The *sign reversal axiom* states that for all $p_i, p_j, p_k, p_l \in P$, if $p_ip_j \succeq_W p_kp_l$, then $p_lp_k \succeq_W p_jp_i$. The axiom is checked easily on all suitable quadruples of cells.

Norwich and Turksen (1982, 1984) were apparently the first to recognize the close relation between the axiomatic formulation of the algebraic difference structure and the validation of membership functions. They provide an elegant mathematical development of the measurement system just outlined.

It is important to note that a pure pair-comparison procedure will yield ordinal data sufficient for checking the axioms and also for nonmetric scaling, but will not provide data from which membership functions can be derived by means of metric scaling procedures. The present modified or any other graded pair-comparison method (Sjöberg, 1980) does yield data that can be analyzed in terms of both axiomatic and metric models.

Scaling Models

One approach to applying a metric difference model is to assume that for a given expression W and probability pair p_ip_j, the subject places the arrow on the response line such that the difference in the distances of the arrow from the two ends is inversely proportional to the difference in the degrees to which W describes p_i and p_j. Thus, the response R can be converted to a difference score $D = 2R - 1$. Least squares estimates of $\mu_W(p_i)$ are then obtained by taking row means of the full $P \times P$ matrix of difference scores for phrase W. The mathematical details underlying this procedure are given in Appendix A.

Similarly, for the ratio-scaling models, it can be assumed that the arrow is placed on the response line such that the ratio of the distances of the arrow from the two ends is inversely proportional to the ratio of the degrees to which W describes p_i and p_j. Thus, the response R can be converted to a ratio score $S = R/(1 - R)$. One ratio-scaling model then involves taking row geometric means of the full $P \times P$ matrix of ratio scores to obtain log least squares estimates of $\mu_W(p_i)$, assuming the matrix is reciprocal. Three other ratio-scaling models, anticipated by Gulliksen (1959), also require a reciprocal matrix, and yield scale values by means of an eigenvalue–eigenvector decomposition. The models differ in terms of whether the scale values are taken as the normalized right eigenvector (Saaty, 1977, 1980), normalized left eigenvector (Johnson, Beine, & Wang, 1979), or the geometric mean of the two eigenvectors (Budescu, 1984). The mathematical details are given in Appendix A.

The difference and four ratio-scaling models can be applied to the $P \times P$ matrix associated with each phrase W. A goodness-of-fit measure that allows all five models to be evaluated and compared is the linear correlation between the observed and the predicted responses. Note that because the scalings are done independently for the $P \times P$ matrix associated with each W, membership values across phrases are not comparable without specific assumptions.

Cross Validation

It is necessary for the validity of any of these scaling models that the axioms be satisfied within limits of error and that the model's goodness-of-fit measure be high. However, such tests are not sufficient for supporting the more interesting claim that the vague meanings of the expressions are validly represented by the derived scale values. For example, this claim would appear unjustified if the scale values plotted as a function of the probabilities (cf. Figure 1) yielded uninterpretable curves, for example, multipeaked curves. Furthermore, for this claim to be justified, it is necessary that the derived values correctly predict an independent set of judgments based on the presumed vagueness of the terms.

In the present experiments, subjects were also tested on trials that were the converse of that shown in Figure 2: namely, there was one spinner at the top of the screen with two terms below it, one on the left and one on the right. The subject moved the

arrow on the response line to indicate how much better one term rather than the other described the displayed probability of landing on white. Scale values derived from the previous judgments should predict certain properties of these responses.

Three properties are derived in Appendix B. They are presented in terms of the scale values obtained by taking row geometric means of ratio scores, because ultimately those were the values with which they were tested. One prediction applies when a fixed pair of phrases W_i and W_j is considered with various probabilities p. In this case, converting the responses R to ratio scores $S = R/(1 - R)$ yields a measure that should be a linear function of the ratio of the previously derived membership functions, given certain reasonable assumptions about the scaling parameters. This prediction is tested in Experiment 1. Similarly, given slightly stronger assumptions about the parameters, when a fixed probability p is considered with various phrases $W_1, W_2, \ldots, W_m$, the ratio scores S should be linearly related to the ratio of the membership functions. This prediction is tested in Experiment 2.

The third prediction derived in Appendix B applies when for a given p there is a left side by right side, $T \times T$ factorial design, in which T is a vector of probability phrases. When this matrix is scaled in the same manner as is the $P \times P$ matrix for a given phrase W, scale values $\mu_p(W)$ are obtained. If $\mu_p(W)$ and the previously discussed $\mu_w(p)$ both represent the same vagueness construct, then they should be related by a power function. This prediction is also tested in Experiment 2.

Although the various empirical evaluations could be carried on in many domains, the present experiments do so for the vague concepts defined by probability expressions. Specifically, the purposes of the present experiments are (a) to evaluate the measurement models by testing their ordinal and goodness-of-fit predictions, (b) to evaluate the claim that the derived values represent the vague meanings of the phrases both by considering the reasonableness of the resulting scales and by predicting an independent set of judgments, and (c) to make some preliminary statements about meanings of nonnumerical probability expressions.

Experiment 1

Experiment 1 was designed to test the feasibility of the modified method of pair comparison and to evaluate its results according to the three aforementioned criteria.

We considered the experimental task to be a difficult one and therefore made a number of decisions intended to maximize the quality of the data. First, we elected to use social science and business graduate students rather than undergraduates as subjects. We assumed that they would represent a population of people who think seriously about communicating degrees of uncertainty, and who generally do so with nonnumerical phrases.

Second, the probabilities used with each term were determined uniquely for each subject. Furthermore, each probability pair was presented only once with a given term in a session. Thus, if probability p_i was presented on the left and p_j on the right, the arrow location, $R_w(ij)$, expressed as a number from 1 to 0, was entered in cell ij and its complement, $1 - R_w(ij)$, was entered in cell ji. Although this procedure has some drawbacks, it greatly reduced the number of trials and the motivation for subjects to hurry through the session. Of particular interest to this study, the procedure forced the sign reversal axiom to be correct, and also yielded the reciprocal matrix required by Saaty's ratio-scaling technique (see Appendix A).

Method

Subjects. In Experiment 1, 20 subjects were recruited by placing notices in graduate student mailboxes in the business school and in the departments of anthropology, economics, history, psychology, and sociology. The general nature of the study was described and subjects were promised $25 for three sessions of approximately an hour and a half each.

Probability phrases. Session 1 was for practice, and Sessions 2 and 3 were for data. During Session 1, all subjects judged *chance, very likely,* and *slight chance.* Ten probability phrases covering the 0–1 range were selected for presentation during Sessions 2 and 3. All subjects judged *doubtful, tossup,* and *likely* during Session 2, *improbable* and *good chance* in Session 3, and *possible* in both sessions.

In addition to these six core phrases, half of the subjects also judged *almost certain* in Session 2 and *probable* in Session 3, whereas the other half judged *almost impossible* and *unlikely,* respectively. The goal of this particular choice was to allow examination of possible effects of list composition on the subjects' judgments.

General procedure. The practice and two data sessions were scheduled generally two days apart. The experiment was controlled by an IBM PC with stimuli presented on a color monitor and responses made on the keyboard. During Sessions 2 and 3, subjects judged the terms listed above. An index card was continuously in view listing all the expressions that the subject would encounter during the course of the experiment.

Each session consisted of three parts. The purpose of Part 1 was to determine the maximum, p^*, and the minimum, p_*, probability for which the subject would judge a given term to be appropriate. The results of this part were then used to determine the unique probabilities to be used in Parts 2 and 3 for each subject.

The second part of the session involved the presentation of probability terms with pairs of spinners, as already discussed. Part 3 reversed the procedure, as also already discussed. Each part will now be described in more detail.

Part 1. The instructions for this segment read in part:

> In a specific context that we will describe shortly, we are interested in the range of uncertainties for which you think it appropriate to use each of various words or phrases that will be displayed on the screen . . .
>
> The context that we will provide is that of spinning a pointer on a spinner that is radially divided into a red sector and a white sector. The relative areas of each sector are clear to you and you must convey that information to a friend. You want to tell him how likely it is that the pointer will land on white if it is fairly spun and randomly stops at some position. However, you are not allowed to tell the person the actual probability of landing on white. Rather, you are forced to use a nonnumerical descriptive phrase . . . We want to know the range of probabilities in this specific spinner context for which you would consider (each term) to be appropriate . . .

The terms scheduled for a given session were presented in random order. On each trial a phrase was written at the top of the screen and a spinner divided vertically into equal areas of red and white was drawn below it. The subject then increased the proportion of white by pressing the *I* key and decreased it by pressing the *D* key. The relative area of white first was adjusted to indicate the lowest probability for which the

subject would conceivably use the displayed term. This value was then registered by pressing the L key.

After the lower limit was indicated, the subject then adjusted the spinner to display the highest probability for which he or she might use the term, which was registered by pressing the U key. The upper limit could not be set below the lower limit.

Instructions for this part ended with three reminders: (a) to consider the use of the expression only in terms of describing the chances of the pointer landing on white for the particular spinner displayed on the screen, not how it might be used in other contexts; (b) not to decide whether the particular term is the best of all possible terms for a given probability, but only whether it conceivably could apply to the displayed relative area; and (c) to select the lowest and highest probabilities carefully, because they were to be used to determine the range of probabilities used with each expression in the subsequent parts of the experiment.

Immediately following Part 1, the interval from p_* to p^* for each term was divided online into n equally spaced probability values for use in Part 2. For each term, n was set at the largest integer between 0 and 8, inclusive, such that the spacing of adjacent probability values was not less than 0.02.

Part 2. Depending on the Part 1 results, the number of probabilities, n, associated with each term ranged from 0 to 8. Terms were presented in this part only if $n \geq 2$. Probabilities were displayed as the relative areas of white on a spinner. Each phrase was presented once with each of the $n(n-1)/2$ pairs of spinners. Phrases and spinner pairs were presented in a random order.

A single trial appeared as shown in Figure 2. The subject moved the arrow on the line to indicate for which spinner the probability of landing on white was better described by the expression and how much better it was described.

The instructions said in part:

> If you had to assign the phrase at the top of the screen to one of the two spinners, to describe the probability of landing on white, to which spinner is it more appropriately assigned and how much more appropriate is the assignment of the phrase to that spinner than to the other one? . . . If you believe the two probabilities are equally well described by the phrase, leave the arrow in the middle. If the probability on one spinner is better described by (the term) than is the other, move the arrow closer to that spinner. The greater the relative appropriateness of the phrase for one probability than for the other, the closer the arrow should be moved to the corresponding spinner. In other words, place the arrow so that its relative distance between the two spinners represents its relative appropriateness for the two probabilities.

The < and > keys on the keyboard were used to move the arrow, which could be positioned at any of 17 equally spaced locations on the line, consistent with response procedures normally used for Saaty's (1977, 1980) ratio-scaling techniques. The R key was used to register the response when the arrow was suitably placed.

Part 3. This was the converse of Part 2. A pair of terms was presented only if the Part 1 estimates for the two terms overlapped. During Session 2, pairs were selected only from terms that were used in Parts 1 and 2 of that session. Pairs were selected the same way in Session 3, but in addition, pairs were formed with one member from Session 2 and one from Session 3 if their Part 2 estimates overlapped sufficiently. The number of probabilities presented with a pair ranged from 1 to 8, with adjacent probabilities differing by at least 0.02. Because of a programming error, the Session 2 and 3 presentations of *possible* were treated separately. Thus, in Session 3, *possible* may have been paired with other phrases up to 16 times each. Spinner and phrase pairs were presented in a random order.

On a trial, a spinner representing a particular probability was presented at the top of the screen; two terms were written below it, and a marked line segment with a centered arrow was below them. In the same manner as in Part 2, the subject moved the arrow on the line segment to indicate which of the two terms better described the probability of the spinner landing on white and how much better the description was.

The instructions read in part:

> If you had to select one of the two phrases to describe the displayed probability of landing on white, which of the two is better, and relatively how much better is it? . . . The relative distance you place the arrow between the two phrases should represent relatively how much better one phrase is for the displayed probability than is the other.

Results

No apparent differences emerged between the two lists of words, so this distinction will be disregarded. Data will be presented separately for the three parts of the experiment.

Part 1. Each subject set upper and lower limits for the range of probabilities that could be associated with each expression. A summary over subjects of these estimates is shown in Figure 4. For each term, the lower left-hand bar shows the interquartile range of the lower limit determinations. Similarly, the lower right-hand bar indicates the 25th and 75th percentiles of the upper limit determinations. The medians of the lower and the upper limit determinations are connected by the top bar for each term. Note (a) the considerable variability over subjects, (b) that even the word *tossup* has a range of meanings from about 0.4 to 0.6 for most subjects, and (c) the enormous differences over subjects in the upper limit of values suitable for the word *possible.*

Despite the considerable between-subject variability in the upper limits for *possible,* individual subjects were reasonably stable over sessions. The correlation between the first and second determinations of the lower limit for *possible* was 0.94 ($p < 0.0001$), and for the upper limit, it was 0.69 ($p < 0.001$).

Part 2. Data from this task were analyzed at the individual level. Each of 20 subjects set upper and lower limits for nine expressions (counting *possible* separately for Sessions 2 and 3), for a total of 180 determinations. The width of each interval determined the number of probabilities to be associated with the corresponding term in this part. At most, eight probabilities were selected to be equally spaced within the interval such that adjacent values differed by at least 0.02. In fact, on 144 occasions (80%), eight probability values were associated with terms in Part 2, on 14 occasions (7.8%), six or seven probabilities were associated with terms, and on 14 occasions (7.8%), the number of probability values used for each terms was three, four, or five. On 8 of the 180 determinations (4.4%), the upper and lower probability limits coincided, and therefore, that term never appeared in Parts 2 or 3.

Considering the ordinal data properties first, judgments were collected in this experiment in a manner such that both the weak ordering and the sign reversal axioms were forced to be satisfied. However, the weak monotonicity axiom could be tested.

Evaluation of the axiom required a matrix of size $n > 4$. Because only one of each reciprocal pair of cells in the $P \times P$ matrix for a phrase was responded to, the number of subsets of six cells for which the axiom could be tested equaled $\binom{n}{3}^2 - \binom{n}{3}$. A

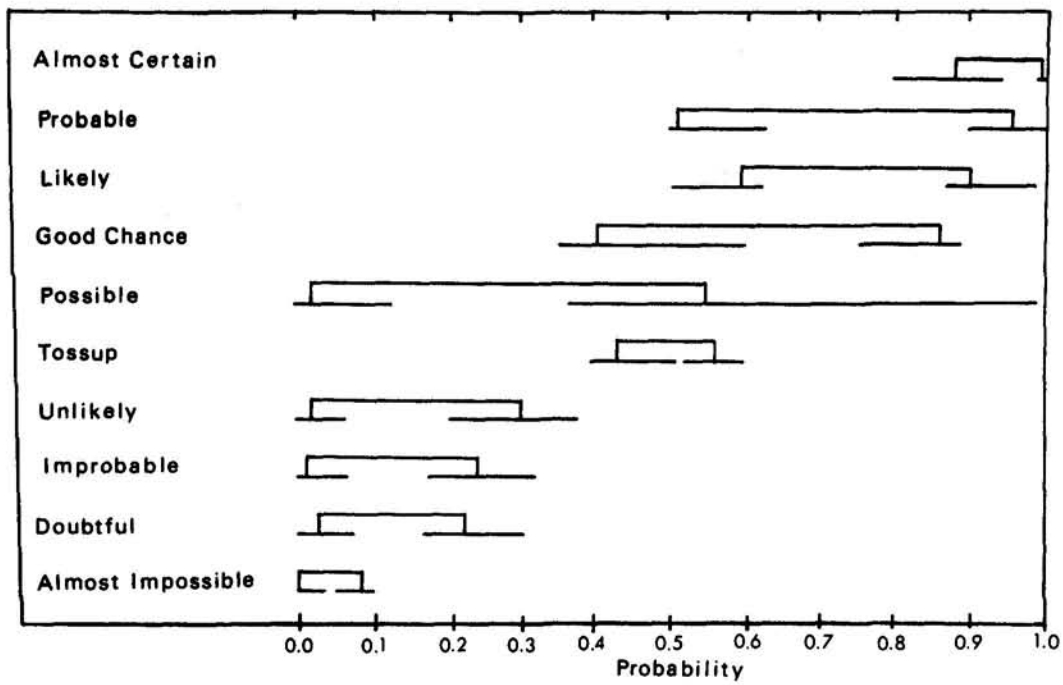

Figure 4. First, second, and third quartiles over subjects of the upper and lower probability limits for each phrase in Experiment 1.

satisfaction index, defined as the percentage of subsets satisfying the consequent condition that also satisfied the antecedent conditions, was determined for each phrase for which $n \geq 4$ for each subject.

The results of the weak monotonicity test are summarized in Table 1, as a function of matrix size. The table shows the three quartiles of the empirical distributions of the satisfaction indices. Because the distribution of this statistic is not known, 400 random matrices were generated for each matrix size encountered in our experiment, and the mean and variance of these null distributions were calculated. The last row in the table shows the percentage of matrices at each size that had satisfaction indices exceeding the mean value for random data by at least three standard deviations. It seems reasonable to conclude that weak monotonicity was well satisfied.

We now turn to the metric scaling. For this analysis the 17 equally spaced response locations were assigned values from left to right of 1, 0.9375, . . . , 0.0625, 0. Because subjects responded to only one member of each pair of reciprocal cells in a matrix, the complementary response was computed. That is, if $R_W(ij)$ was the response to $p_i p_j$ for phrase W, $R_W(ji) = 1 - R_W(ij)$ was entered in cell $p_j p_i$.

Each matrix was scaled according to the models in Appendix A. Scale values from the difference models were obtained through application of Equations A1 and A3. In order to transform responses by Equation A4 for ratio scaling, responses, $R_W(ij)$, of 0 and 1 were first set equal to 0.0156 and 0.9844, respectively (i.e., one fourth of the distance between the most extreme and the immediately adjacent responses), to avoid division by 0. Then the geometric-mean ratio scaling was accomplished via Equation A6. Ratio-scaling solutions were also obtained by a right eigenvector–eigenvalue decomposition, a left eigenvector–eigenvalue decomposition, and by taking the geometric mean of the two eigenvectors.

The mean linear correlations between observed and predicted responses over all subjects and phrases were .75, .77, .75, .75, and .76 for the difference, geometric mean, right eigenvector, left eigenvector, and mean eigenvector models, respectively. Thus, all the models scaled the data about equally well, with a slight superiority for the geometric-mean model. Detailed results will be presented only for the geometric-mean model; the others show similar patterns.

Recall that on eight occasions, the upper and lower probability limits from Part 1 coincided so that the phrases did not ap-

Table 1
Summary of Satisfaction Indexes for Weak Monotonicity in Experiment 1

Index	Matrix size			Total
	4 or 5	6 or 7	8	
Number of matrices	9	14	144	167
25th percentile	80	77	75	75
50th percentile	87	83	82	82
75th percentile	92	91	89	89
% for which $z > 3.0$[a]	0	100	91	87

Note. The satisfaction index is the percentage of submatrices for which the antecedent conditions are met that also satisfy the consequent condition.

[a] This figure is based on simulated data.

Table 2
Summary of Linear Correlations Between Observed and Predicted Responses for the Geometric-Mean Scaling Model in Experiment 1

Index	Matrix size 3–5	6 or 7	8
Number of matrices	14	7	144
25th percentile	.70	.57	.64
50th percentile	.85	.74	.79
75th percentile	.91	.85	.87
% for which $p < .01$	64	79	83

pear in Part 2. Thus, 172 matrices were scaled, and for each a linear correlation was calculated between observed responses and those predicted by the geometric-mean scaling model. The distribution of correlations is summarized in Table 2 as a function of matrix size. The last row in the table shows the percentage of correlations that are significantly different from zero at each matrix size. It can be seen that the model reproduces the data to a reasonably good degree. For example, at matrix size 8, the model accounts for at least 62% of the response variance (0.79^2) in 50% of the cases, and for at least 41% (0.64^2) of the response variance in 75% of the cases.

One may ask whether subjects judged some expressions with more internal consistency than others, so that the scaling model provided a better fit in those cases. The top part of Table 3 shows the mean linear correlation between observed and predicted responses for the geometric mean model separately for each expression. The Session 2 and Session 3 presentations of *possible* are combined, because they were not different. Note that *tossup* is fitted considerably better than the other expressions on the average, but that otherwise there are no substantial differences among the terms.

We now turn to the scale values to consider how reasonable they are as membership functions. For this analysis, the derived values from each matrix were multiplied by a suitable constant so that the maximum value equaled one. The normalized values were plotted separately for each subject and each expression as a function of the spinner probabilities of landing on white. We will use the term *membership function* for the resulting graphs. Figure 5 illustrates the membership functions from 3 different subjects to show the range of results obtained.

Subject 1 has monotonic membership functions with the exception of that for *tossup*. The remaining terms each span a range of probabilities, and the probability best described by each term is at one end of the range. Because Subject 1 set the upper and lower probability limits for *tossup* equal to each other, its membership function is a point.

Subject 23 has membership functions that tend to be single peaked. Thus, for this subject each expression spans a range of probabilities and the probability best described by that expression is somewhere in the center of the range.

Subject 6 has both kinds of membership functions. This subject also illustrates functions that are not quite as well behaved, having two or even three peaks.

Recall that the functions for each expression were arbitrarily adjusted to have a maximum of one, so that comparisons of ordinate values over terms is not meaningful. Also, ordinate values do not extend quite to zero, because the method of selecting probabilities based on Part 1 judgments purposely omitted probabilities with such membership values.

The various functions can be characterized as either point (4%), flat (2%), monotonic increasing or monotonic decreasing (30%), single peaked (31%), or as having two (26%), or up to four peaks (7%). The point, flat, monotonic, and single-peaked functions might all be considered reasonable, in terms of the supposed underlying semantics, whereas the others cannot easily be so considered. Overall, 67% of the functions were reasonable by this criterion. If these double-peaked functions in which one peak is minor are also included, then about 75% of the functions are reasonable and interpretable.

The bottom part of Table 3 shows the percentage of types of membership functions obtained for each term. For these purposes it was assumed that the multipeaked functions contained noise, and they were classified in with the flat, monotone increasing, single peaked, or monotone decreasing functions, as appropriate. It can be seen first that there was no expression for which all subjects had the same shape function. Second, terms closer to the extremes tended to have more monotonic than sin-

Table 3
Mean Goodness-of-Fit Correlations and Percentages of Different Shapes of Membership Functions for Each Term in Experiment 1

Shape	Almost certain	Probable	Likely	Good chance	Possible	Tossup	Unlikely	Improbable	Doubtful	Almost impossible
					Goodness-of-fit correlations					
All	.76	.73	.75[a]	.73	.73[a]	.93	.70	.80	.76	.85
					Percentage of different membership function shapes					
Point						25				20
Flat			5		12					
Monotonic increasing	90	60	45	45	12					
1 peak	10	40	50	45	58	75	50	20	25	20
Monotonic decreasing				10	18		50	80	75	60

[a] Excluding solutions with equal scale values.

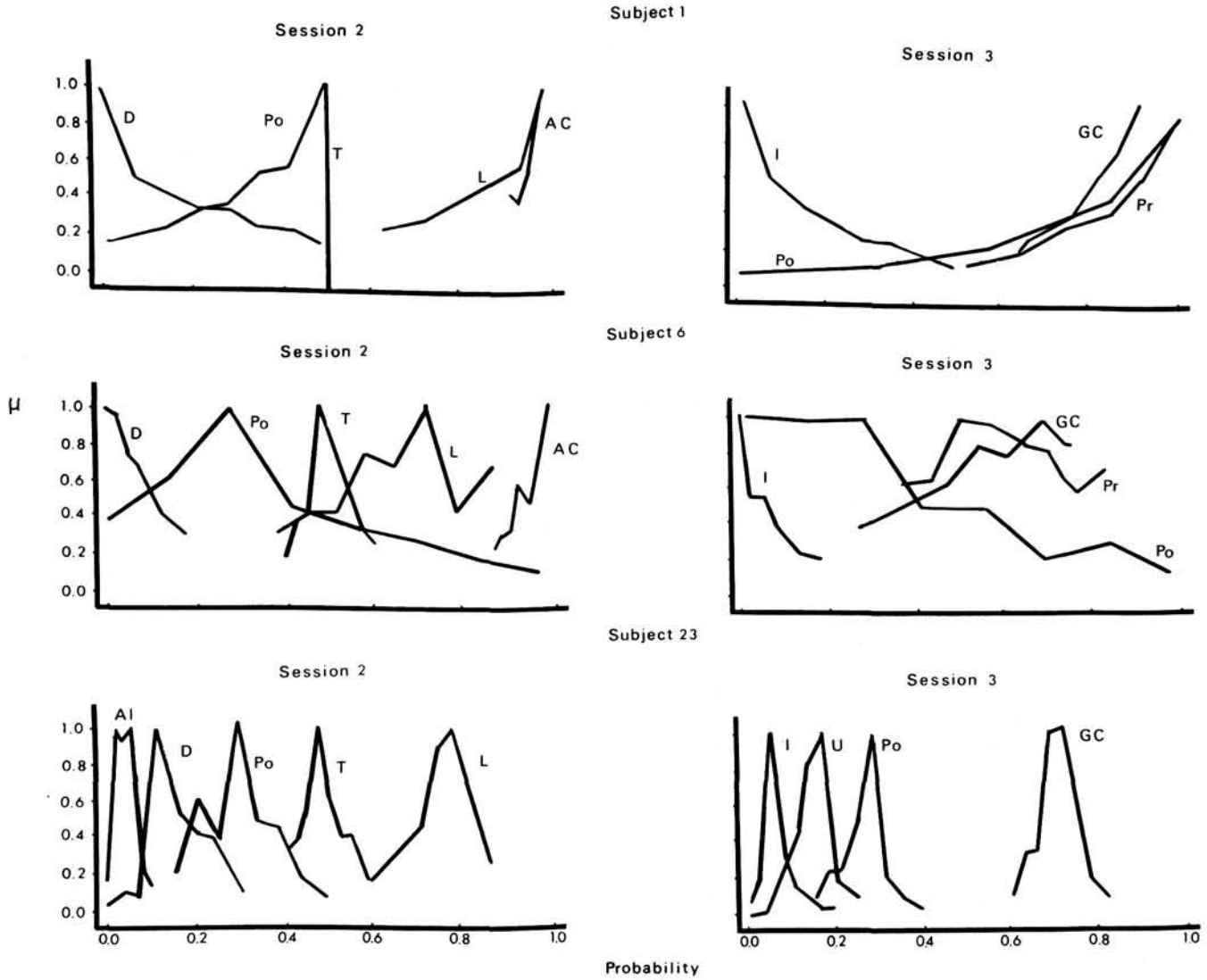

Figure 5. Derived membership functions for three subjects in Experiment 1. (The functions are coded as follows: AC = *almost certain;* AI = *almost impossible;* D = *doubtful;* GC = *good chance;* I = *improbable;* L = *likely;* Po = *possible;* Pr = *probable;* T = *tossup;* U = *unlikely.*)

gle-peaked functions, whereas terms near the middle of the probability range tended to have more single-peaked than monotonic functions. *Tossup* and *almost impossible* had point meanings for a few people. Finally, all forms of functions except point were obtained for *possible.*

However, even membership functions of the same type for a term did not look the same over subjects. The three expressions for which the highest agreement on meaning was obtained are *almost impossible, almost certain,* and *tossup.* Their membership functions from all subjects are shown in Figure 6. Five subjects have point functions for *tossup,* 2 have single-peaked functions that look different from the others, and the remaining 13 subjects show very similar functions. *Almost impossible* and *almost certain* show more variability over subjects.

Expressions that are not near the anchor points of 0, 0.5, and 1 show even greater individual differences. As one example, the membership functions for the word *doubtful* are shown in Figure 7. For purposes of clarity only, the monotonic functions are shown on the top half of the figure and the single-peaked functions are shown on the bottom half. Note that some functions cover a large range and some a much smaller one. The peaks of the functions range from probability values close to zero to approximately 0.17. Analogous results hold for the other terms as well.

Part 3. The number of pairs of expressions and the number of probabilities per pair that a subject judged depended on the upper and lower limits set in Part 1. Combining over both sessions, the number of pairs judged per subject ranged from 1 to 18 with a mean of 11.5 and a standard deviation of 4.4.

The number of probabilities judged per pair of expressions ranged from 1 to 8, except that up to 16 probabilities were judged with pairs including *possible* in Session 3. Combining

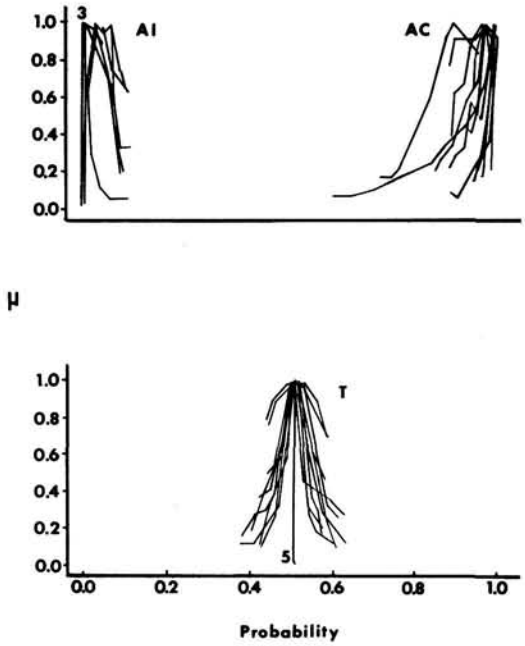

Figure 6. Membership functions from all subjects in Experiment 1 for *almost certain* (AC), *almost impossible* (AI), and *tossup* (T).

over sessions, the mean number of probabilities judged per pair was 8.4 with a standard deviation of 1.4.

The analysis involved evaluating the predicted linear relation between Part 3 judgments converted to ratio scores and ratios of the membership function values derived in Part 2. Because the same probabilities were not generally presented with a term in the two parts, membership function values for probabilities used in Part 3 were estimated by linearly interpolating between the values derived for the two adjacent probabilities that were used in Part 2. The ratios of the estimated values were then used in Equation B1 (see Appendix B) to predict the judgments, R, converted to ratios of distances, $S = R/(1 - R)$, where, as before, $R = 0$ and 1 were converted to 0.0156 and 0.9844, respectively.

If β_j, $\beta_i = 1$ for all W_i and W_j in Equation B1, then within a pair of phrases the ratio of distances should be a linear function of the ratio of membership function values. This prediction was evaluated by means of a simple linear correlation pooled over phrase pairs for each subject in order to increase power. By pooling over expression pairs, the number of observations per correlation ranged from 7 to 146 over subjects ($M = 83.2$, $SD = 37.7$). The mean pooled correlation over subjects was 0.37, with a standard deviation of 0.23. Thirteen of the 20 correlations were significantly different from zero at $p < 0.05$.

Discussion

The results of Experiment 1 are quite encouraging overall, although in hindsight some design features were problematic. Part 1 provides the sole point of comparison between this study and others that have used a more traditional method to assess the meanings of probability phrases. The usual finding when subjects are asked to give numerical equivalents to probability phrases is considerable between-subjects variability that is inversely related to distance from the center of the scale. This is precisely the pattern we obtained for the judgments of upper and lower probability limits.

The data of primary interest, of course, are from Part 2. Despite the lack of good inferential statistics, it seems justifiable to say that the weak monotonicity axiom was well satisfied in the vast majority of cases. This, in conjunction with the fact that the other necessary conditions were forced to be satisfied by the data collection procedure, provided justification for applying the metric models to the data. The scaling models fit well, accounting on the average for about 56% of the variance in the observed judgments without fitting a single free parameter. Nonmetric scaling procedures or procedures involving the estimation of free parameters might have done even better. Nevertheless, the derived scale values were generally of reasonable shape, and predicted the Part 3 responses to a relatively high degree. Thus, it appears justifiable to conclude that subjects can compare degrees of membership in a way that leads to consistent, meaningful, and interpretable scaling of vague meanings according to either a ratio or a difference model. However, it must be emphasized that nothing in the data allows us to determine whether subjects are more likely to judge ratios or differences. Another conclusion is that even in the context of the present experiment, in which the probabilities are well defined, there are large individual differences in the vague meanings of probability phrases.

We allowed unique probability values to be associated with each expression over subjects because we expected considerable individual differences, and because we were uncertain as to what

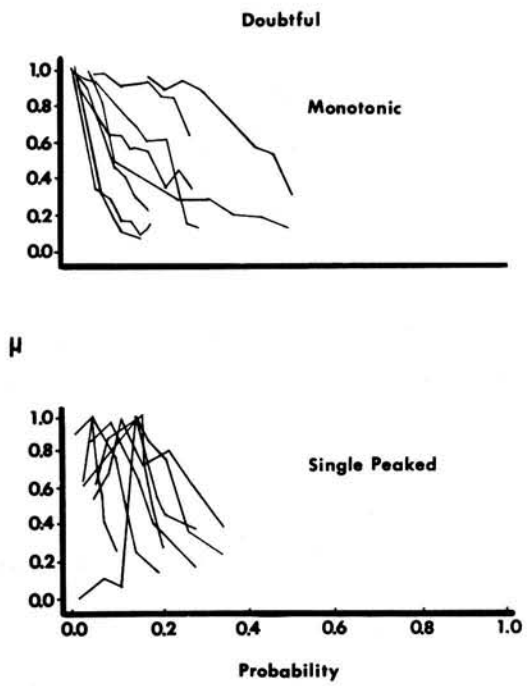

Figure 7. Membership functions from all subjects in Experiment 1 for *doubtful.*

Table 4
Expressions and Probability Values Used in Experiment 2

	Probabilities (×100)												
	5	20	30	40	45	47	50	53	55	60	75	85	95
						List 1							
Probable				X	X		X			X	X	X	X
Good chance			X	X	X		X		X		X		X
Tossup				X	X	X	X	X	X	X			
Doubtful	X	X	X	X	X	X	X						
						List 2							
Likely				X	X		X			X	X	X	X
Good chance			X	X	X		X		X		X		X
Tossup				X	X	X	X	X	X	X			
Improbable	X	X	X	X	X	X	X						

probability range would be appropriate for a large group of people. However, problems emerged as a result of individualizing the stimuli for each subject. First, all the trials in Parts 2 and 3 depended on a single determination of an upper and lower probability per term. If a subject made an error in Part 1 by setting a limit too high or too low, that error affected all the subsequent results. Note in Figure 6, for example, the two single-peaked functions for *tossup* that are different from the others. One would expect the derived membership function to extend closer to zero. If in Part 1 those two subjects had provided lesser lower bounds and greater upper bounds, then a larger range of probabilities would have been presented to them in Part 2 and presumably more complete functions would have been derived.

As a further result of the strong reliance on the Part 1 judgments, there was no good way to evaluate the stability over time of the membership functions for *possible*. This is because subjects tended to set different upper and lower probability limits for this term in the two sessions, resulting in different pair comparisons. We originally selected *possible* on the assumption that it would cover the broadest probability range and therefore provide the most sensitive test. In retrospect, the semantics of *possible* are very complex and subjects probably attributed different meanings to the phrase in the two sessions.

There are yet two more consequences to having determined the Part 2 and 3 stimuli uniquely for each subject and for each part. One is that it was difficult to compare a phrase's membership functions over subjects. Membership functions that differ in shape are obviously distinct. However, two functions of the same shape may have distinct values at a given point, which is due only to the particular probabilities presented to the subject.

The other consequence is that predictions from Part 2 to Part 3 were weakened because it was necessary to base them on linear interpolations. Predictions would have been much more direct had they involved the same probabilities appearing with the phrase in Parts 2 and 3.

Finally, subjects did not report the task to be as difficult as we originally had envisioned they would. Thus, it might not have been necessary to have presented each probability pair only once for an expression in Part 2 and each expression pair only once for a probability in Part 3. Had each combination been presented at least twice (with the left–right ordering of the pairs reversed in half the trials) it would have been possible to have checked the sign reversal axiom, to have obtained a more thorough test of weak monotonicity (because more cells would have been involved), and to have determined empirically whether the response matrix was reciprocal. Thus, a second experiment was performed to provide more reliable membership functions and more complete tests of the predictions.

Experiment 2

Method

Subjects. From each of the two phrase-list groups in Experiment 1, the 4 subjects with the highest mean goodness-of-fit correlations for the geometric mean scaling model were invited to take part in this study. Each of the 8 was promised $15 for two sessions of approximately an hour and a quarter each.

Procedure. There was no Part 1. (However, for the sake of continuity, we will continue to denote the other two parts of the sessions as Part 2 and Part 3, respectively.) Rather, probabilities were selected on the basis of results from Experiment 1 and the same values were used for all subjects.

Table 4 shows the expressions and associated probability values that were used. List 1 consisted of *doubtful, good chance, tossup,* and *probable,* whereas List 2 included *improbable, good chance, tossup* and *likely.* Part 2 used all possible pairs of the seven probabilities indicated for each term in the table. Part 3 used all possible pairs of the terms indicated for each probability. Full left side by right side factorial designs were run within each session. That is to say, in Part 2, each distinct pair of probabilities was presented twice with each expression in each session, once in one left–right orientation and once in the reverse orientation. Similarly, in Part 3 each expression pair was presented twice with each probability in each session, once in each orientation. Within each part, presentation order was random.

The response procedure was also changed, so that subjects used a joystick to move the arrow on the response line. When the arrow was located in the desired position, the subject registered that response by pressing a button on the joystick assembly. Whereas in the previous experiment the arrow could be located at any one of 17 discrete locations, the response line was essentially continuous in this study, limited only by the resolution of the screen.

Table 5
Reliability Correlations for Parts 2 and 3 and Satisfaction Indexes for the Weak Monotonicity Axiom: Experiment 2

Subject	Reliability correlations		Satisfaction index
	Part 2	Part 3	
1	.89	.93	90
4	.93	.83	90
8	.96	.81	91
9	.75	.89	86
14	.78	.61	90
16	.97	.98	91
17	.89	.93	90
20	.88	.95	91
M	.90	.90	90

Each subject was tested on the full design within each of two sessions with approximately two days intervening.

Results

Reliability. Linear correlations were used to assess reliability separately for Parts 2 and 3. The results are shown by subject in the first two columns of Table 5. (In this and subsequent tables, marginal mean correlations are based on Fisher's *r*-to-*z* transformation.) All subjects demonstrated quite high reliability, with Subject 9 showing the lowest Part 2 correlation and Subject 14 showing the lowest Part 3 correlation. Considering this result, most subsequent analyses were done over the two sessions combined.

Weak monotonicity. By using mean responses over the two sessions, this axiom was checked in the same manner as in Experiment 1. However, because the full $P \times P$ matrix was responded to for each phrase, $[n!/(n-3)!]^2 - [n!/(n-3)!]$ subsets of cells are available for test in each $P \times P$ matrix. For $n = 7$, a total of 43,890 subsets of cells can be tested for each phrase.

The last column in Table 5 shows the mean percentage of monotonicity tests that were satisfied. It can be seen that the axiom is extremely well satisfied for all subjects. The mean satisfactions for the terms *doubtful* and *improbable* are 75% and 83%, respectively, but those for the other four terms vary between 92% and 94%.

Sign reversal and reciprocity. If for a given $P \times P$ matrix the entry in cell p_ip_j is the complement of that in cell p_jp_i, then the axiom of sign reversal is satisfied. In addition, the matrix for ratio scaling obtained by the transformation in Equation A4 will be reciprocal. An evaluation of complementarity is obtained by calculating the correlation for responses in cell p_ip_j as a function of those in cell p_jp_i, as well as by fitting a linear structural model to these values (Isaac, 1970). Ideally, the correlation and the slope of the best fitting line will both be -1. A linear structural model differs from a regression model in that it allows random error in both coordinates, not in just one. These analyses were applied to the response matrices for both Part 2 and Part 3. Mean slopes and pooled correlations for each subject are shown in Table 6, where it can be seen that the slopes and the correlations are very close to -1 for all subjects.

Ratio scaling and membership functions. Part 2 responses were transformed according to Equation A4 (setting R = 0 and 1 equal to 0.004 and 0.9996, respectively), and the geometric mean scaling model was applied to them. Goodness-of-fit correlations are shown in Table 7 separately for each subject and phrase, but averaged over sessions. It can be seen that goodness of fit is excellent, with the lowest correlation being 0.81 for *likely* for Subject 16.

Normalizing the scale values from the separate matrices to have a maximum of 1 and plotting them as a function of the probabilities demonstrates that the Session 1 and Session 2 membership functions for each subject are quite similar, as would be expected given the high reliability. Thus, average membership functions were obtained by applying the geometric-mean model to the mean responses of Sessions 1 and 2. The results are shown in Figures 8 and 9, respectively, with a separate panel for each subject. As in Experiment 1, all subjects demonstrate similar membership functions for the word *tossup*. The functions for *doubtful* are also quite similar in shape, but those for the remaining expressions show remarkable differences over subjects. Application of nonmetric scaling procedures resulted in functions substantially similar to those shown in Figures 8 and 9.

Predicting Part 3 responses. Note in Table 4 that only three probability values (0.40, 0.45, and 0.50) were each associated with all four terms, whereas the remaining probabilities that appeared in Part 3 were associated with only two terms each. Thus, predictions are possible only for trials that included these three values.

For each of these values, there was a left side by right side, term by term factorial design, except of course omitting the diagonal cells. Scale values derived in Part 2 were used in Equation B1 to predict Part 3 responses transformed to a ratio of distances and combined over reciprocal cells for increased stability. On the assumption that all constants in the equation are equal to 1, the prediction is evaluated by calculating the linear correlation between predicted and observed values at each of the three probabilities. Results by subject are shown in Table 8, where it can be seen that the prediction is quite well sustained for all except Subjects 16 and 20.

Finally, the factorial design at each of the three probabilities

Table 6
Mean Slopes from the Linear Structural Model and Pooled Correlations for Responses in Cell (j, i) as a Function of Cell (i, j), Over Terms and Sessions in Experiment 2

Subject	Part 2		Part 3	
	Slope	Correlation	Slope	Correlation
1	−1.00	−.92	−1.01	−.96
4	−1.00	−.93	−.97	−.97
8	−1.05	−.94	−.87	−.86
9	−.98	−.85	−.88	−.95
14	−.94	−.88	−1.08	−.85
16	−1.03	−.98	−.98	−.99
17	−1.00	−.88	−.91	−.96
20	−.92	−.92	−1.02	−.95
M	−.99	−.92	−.97	−.95

Table 7
Mean Linear Correlations Between Observed and Predicted Responses for the Geometric-Mean Model: Experiment 2

Subject	Probable	Likely	Good chance	Tossup	Improbable	Doubtful	$\bar{r}$
1	.93		.88	.93		.97	.95
4	.97		.96	.96		.98	.97
8	.98		.99	.96		.98	.98
9	.78		.91	.83		.96	.89
14		.91	.89	.97	.97		.95
16		.81	.84	1.00	.97		.95
17		.98	.97	.88	.96		.96
20		.95	.92	.94	.98		.95
$\bar{r}$	.96	.93	.94	.95	.97	.97	.95

allows a geometric-mean ratio scaling of the response matrices by using equations analogous to Equations A4, A5, and A6. As shown in Equation B3, the resulting membership function values, $\mu_p(W)$, should be a power function of those derived in Part 2, $\mu_W(p)$, if they both represent the same vagueness construct.

Power functions were fitted to the scatter plot of $\mu_p(W)$ versus $\mu_W(p)$ for each subject, and were assessed by means of F ratios. The F ratios ranged from 43.2 to 7,461 over subjects, with a median value of 143. Although inferential statistics are not appropriate (because the data points are not independent), it is descriptively clear that the functions fit very well.

Discussion

Experiment 2 seems to have overcome the problems of Experiment 1 while substantiating its main results. Because subjects were selected for inclusion in this study on the basis of their scaling results in Experiment 1, it is perhaps not surprising that the algebraic difference structure axioms were well satisfied and that the geometric-mean ratio-scaling model described the judgments to a high degree in each case. However, it was necessary to obtain the good fits in order to test properly the other predictions.

The first notable result is that judgments were very stable over the two sessions, but differed considerably over subjects for all terms except *tossup*. As a consequence, membership functions for all the other terms varied widely and reliably over subjects. *Tossup* yielded similar single-peaked functions for all eight subjects. *Doubtful* yielded different monotonic decreasing functions for the four subjects who judged it, and the remaining phrases resulted in both monotonic and single-peaked functions. Furthermore, in these cases, same-shaped functions did

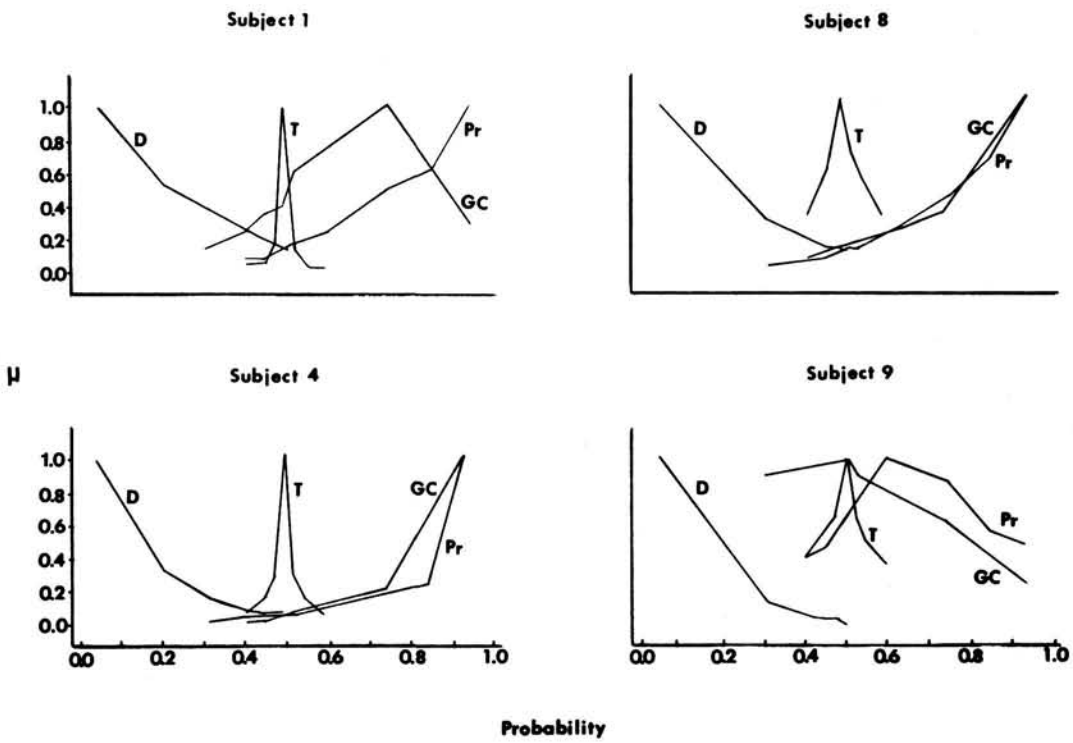

Figure 8. Membership functions for Subjects 1, 4, 8, and 9 in Experiment 2. (The functions are coded as follows: D = *doubtful;* GC = *good chance;* Pr = *probable;* T = *tossup.*)

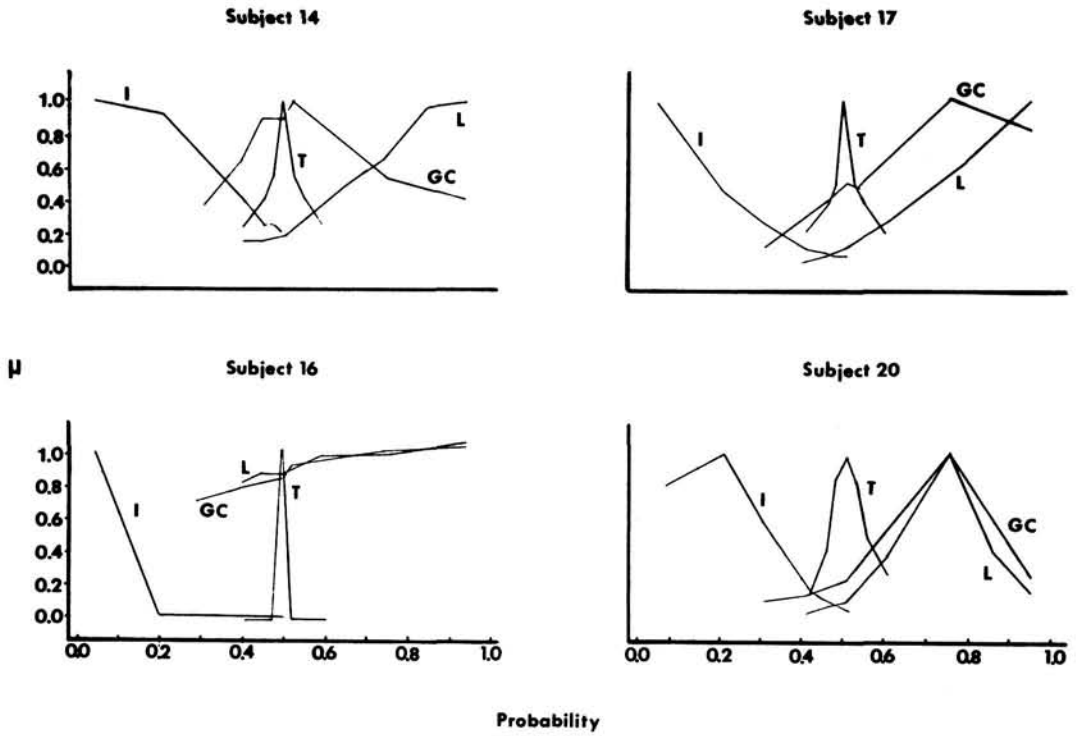

Figure 9. Membership functions for Subjects 14, 16, 17, and 20 in Experiment 2. (The functions are coded as follows: GC = *good chance;* I = *improbable;* L = *likely;* T = *tossup.*)

not take on similar values, so that none of the remaining terms had precisely the same functions for any 2 subjects.

It is of interest to compare these membership functions to their counterparts in Experiment 1. Recall that the subjects in this experiment also participated in the first one, and that they had judged the same expressions (among others) at that time. Because different probability values were used in the two studies, the only possible comparisons are in terms of membership function shapes. Of the 32 comparisons (8 subjects × 4 expressions each), derived membership functions were similar in shape in 25 cases. Of the remaining 7 cases, 6 changed from point or monotonic to single peaked, and 1 changed from single peaked to monotonic decreasing.

On two grounds, it is reasonable to assume that the membership functions in this experiment in fact represented the vague meanings of the phrases to the subjects in this context. First, they all had sensible shapes. But of greater importance, they predicted independent judgments in Part 3 very well. Freed from the necessity of interpolation, ratios of membership function values derived in Part 2 correlated very highly with ratios of Part 3 responses converted to distances. In addition, membership function values independently derived from judgments in Parts 2 and 3 were related by a power function, as they were predicted to on the assumption that they were both measures of the same construct.

Table 8
Linear Correlations Between Observed and Predicted Part 3 Responses Transformed to Distance Ratios in Experiment 2

	Probability			
Subject	.40	.45	.50	$\bar{r}$
1	.71	.77	.99	.91
4	.85	.95	.95	.93
8	.87	.74	.89	.84
9	.73	.97	.85	.89
14	.60	.86	.98	.89
16	.43	.57	.88	.68
17	.93	.78	.82	.86
20	.76	.27	.55	.56
$\bar{r}$	.77	.82	.92	.85

General Discussion

Methodological Issues

We have demonstrated that in a specific context an individual's understanding of the vague meaning of a nonnumerical probability expression can be measured in a valid and reliable way. Previous studies in which membership functions have been constructed from choice probabilities have been criticized as doing little more than relabeling measurement and sampling error as construct vagueness. Studies that used magnitude estimation procedures have addressed vagueness directly, but frequently without a way to assess the meaningfulness or validity of the resulting scales. The procedures used in the present ex-

periments avoided these problems. Subjects directly compared degrees of membership of a stimulus in two ill-defined categories within an experimental design that yielded three converging means for assessing the quality of the judgments.

First, conjoint measurement provided the theoretical rationale for numerically scaling the judgments. Therefore, evaluation of the necessary conjoint-measurement axioms provided a means of evaluating the internal consistency of the judgments prior to numerical scaling. If the axioms had failed empirically, then we would have concluded that the subjects were not judging degrees of membership according to the difference or ratio rule that was to underlie the numerical scaling. Consequently, such scaling would have been inadmissible.

Because the axioms were generally well satisfied, the numerical scaling procedures were applied to the judgments. Goodness-of-fit measures, namely, the correlations between observed and predicted responses, provided a second validity check. If the fits had been poor, then we would have concluded that the judgments were not represented well by the scales. We used metric scaling procedures that utilized no free parameters, and, particularly in Experiment 2, achieved excellent fits. Had that not been the case, nonmetric methods with parameters fit to data could have been used. Goodness of fit would have improved, but not necessarily to an acceptable level.

Although the two checks on the validity of the measurement procedures were passed, it is not necessary to conclude that subjects were judging the semantic vagueness of the term. They could have been consistently judging some other quality instead. The third validity assessment, in the spirit of construct validity, was achieved by using the derived membership function values to predict independent judgments that were presumed to be based on the underlying vagueness dimension. The predictions generally were borne out, and consequently it appears justifiable to claim that the vague meanings of the terms were measured.

From the usual perspective of test theory, reliability is logically prior to validity and therefore must be established first. Judgments were reliable in Experiment 2 by the usual criteria, as were Part 1 upper and lower probabilities for *possible* in Experiment 1. However beyond the high test–retest correlations, the derived membership functions for each subject in Experiment 2 were very similar over the two sessions, and indeed, generally reproduced the membership function shapes derived some 10 weeks earlier for corresponding terms in Experiment 1. This can be taken as further evidence that the subjects were judging an enduring property of the expressions.

On all the above grounds, we believe that the methodological aims of the study have been satisfied, and that we have established a means for validly measuring the vague meanings of nonnumerical probability expressions. The results of Experiment 1, in which 20 subjects judged 10 phrases, were substantiated and refined in Experiment 2, in which 8 experienced subjects judged 6 phrases. Although we have not done so, there is no reason to think that the procedures could not be applied to other linguistic variables or vague categories as well.

Substantive Issues

Numerous questions of substantive interest are raised by these results. First, it must be emphasized that the data clearly support the claim that nonnumerical probability expressions convey vague uncertainties. It is noteworthy that such results were obtained despite the fact that the probabilistic events (spinner pointers landing on white) were exactly specified and easily judged numerically. To check the truth of this latter statement, three subjects subsequently provided numerical judgments of these spinner probabilities with essentially no error. Also, subjects gave virtually errorless probability estimates of physical spinners in a study by Wallsten (1971). Thus, the vagueness can be attributed to the verbal expressions, and not to the perceived uncertainty.

Individual differences. Of course, it is just when the uncertainty and the events are ill defined that nonnumerical expressions are normally used. On the basis of Beyth-Marom's (1982) results, we expect that individual differences in understanding these expressions would be even greater in such ill-defined situations than in the present context. Alternatively, it might be argued that the large individual differences emerged because each person developed his or her own strategy for coping with the unnatural task of using nonnumerical probability expressions in a situation involving precise probabilities. Consequently, individual differences would be less in more natural situations. The claim strikes us as unlikely, but it cannot be dismissed at this point. However, the methodology can be extended easily to ill-defined uncertainties where the competing claims can be investigated. Should consistent individual differences remain in the location and shapes of membership functions under more natural conditions, it would become necessary to identify characteristics, such as experience, training, or linguistic background, that would correlate with them.

Context effects. The meanings of nonnumerical probability phrases, even to an individual, are almost assuredly not fixed over contexts. Thus, the precise numerical aspects of the present results should not be taken too seriously. Various systematic context effects may be identified.

For example, Pepper and Prytulak (1974) have shown that the interpretations of relative quantifiers such as *frequently* or *sometimes* depend on the expected frequency of the event being described; Cohen, Dearnley, and Hansel (1958) have shown that the interpretations of quantifiers of amount such as *some* or *several* depend on the available quantity; and Wallsten, Fillenbaum, and Cox (1986) have shown that the interpretations of probability expressions depend on the base rate of the event in question. In addition, we may speculate that event importance and desirability also affect the meanings of probability expressions.

Zimmer (1984) has suggested that the interpretations of probability expressions vary over knowledge domains. He has proposed a model in which each phrase has a basic meaning represented by a membership function, which is then operated on by a context-specific "scope function" to yield the phrase's context-bound meaning. Whether or not this intriguing model is correct, it is reasonable to expect changes in the probability ranges covered by an expression, or changes in the relative magnitudes of membership values, as a function of knowledge domain.

Semantics. In Experiment 1, 55% of the membership functions were monotonic, 39% were single peaked, and 6% were point or flat. In Experiment 2, 56% of the functions were mono-

tonic and 44% were single peaked. Furthermore, the extreme expressions tended to yield monotonic functions, whereas the more central ones tended to yield single-peaked functions.

It is noteworthy that similar distributions of function shapes occurred in other studies using other domains. Hersh and Caramazza (1976) considered the terms *small* and *large* along with the modifiers *not, very, very very, not very, not very very,* and *sort of,* as applied to squares of different areas. Hersh et al. (1979) investigated the terms *short* and *long* alone and with the modifiers *very* and *sort of* as applied to line lengths. Norwich and Turksen (1984) investigated *tall, very tall, not tall,* and *short,* and MacVicar-Whelan (1978) looked at *tall* and *short,* alone and in combination with *very,* all with regard to men's heights. Kuz'Min (1981) considered *cold* and *warm* along with numerous modifiers as applied to water temperature for swimming, as well as *obsolete* and *up to date* with and without modifiers as applied to age of journal articles with regard to relevance. Finally, Zysno (1981) considered *old* and *young,* alone and with *very,* applied to men's ages. The studies used various empirical procedures, some of which we have taken issue with earlier, but generalizations do emerge. In particular, the majority of the membership functions were monotonic. Single-peaked functions occurred more frequently with hedged expressions (e.g., *sort of large*), with expressions that naturally occupy a mid-range on a continuum (e.g., *warm*), and occasionally with certain expressions when a clearly more extreme one was also being considered (e.g., *large* in the presence of *very large*).

Given the similarity in distribution of function shapes between the present experiments and the others, it is plausible to assume that this aspect of the results is not artifactually due to the spinner context. Thus, although probability values represented by phrases change with context, perhaps the shapes of the functions for an individual do not. If this were so, it might be possible to relate an individual's use of specific expressions to his or her membership functions for them.

Undoubtedly, people select and understand probability phrases not only as representing amounts of uncertainty, but also as representing degrees of confidence in that uncertainty, expectation that the uncertainty may change with information, as well as other factors. An interesting possibility is that some of these factors may be captured by the function shape. For example, perhaps an expression selected by an individual to represent a firmly established level of uncertainty following receipt of information would be either monotonic or sharply single peaked for this person, whereas another expression selected to represent more diffuse uncertainty would be broadly single peaked.

Another possibility is that selection or understanding of phrases may relate to their relative membership values for an individual. Recall that in the present derivation of membership functions, the maximum membership value for each expression was arbitrarily set to one. However, direct comparisons of relative membership values could be elicited. Thus, some terms such as *possible* might have uniformly low membership function values, whereas others such as *tossup* or *almost certain* would have high values for some probabilities.

The procedures developed in our experiments provide some insight into individuals' use of probability terms. More important, however, they provide the means for investigating questions of the sort raised in the latter part of this discussion regarding how people form and communicate vague opinions.

References

Ballmer, T. T., & Pinkal, M. (Eds.) (1983). *Approaching vagueness.* Amsterdam, the Netherlands: North-Holland.

Bass, B. M., Cascio, W. F., & O'Connor, E. J. (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.

Birnbaum, M. H. (1980). Comparison of two theories of "ratio" and "difference" judgments. *Journal of Experimental Psychology: General, 109,* 304–319.

Budescu, D. V. (1984). Scaling binary comparison matrices: A comment on Narasimhan's proposal and other methods. *Fuzzy Sets and Systems, 14,* 187–192.

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

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

Cohen, J., Dearnley, E. J., and Hansel, C. E. M. (1958). A quantitative study of meaning. *British Journal of Educational Psychology, 28,* 141–148.

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

De Jong, P. (1984). A statistical approach to Saaty's scaling method for priorities. *Journal of Mathematical Psychology, 28,* 467–478.

Foley, B. J. (1959). The expression of certainty. *American Journal of Psychology, 72,* 614–615.

Gaines, B. R., & Kohout, L. J. (1977). The fuzzy decade: A bibliography of fuzzy systems and closely related topics. *International Journal of Man–Machine Studies, 9,* 1–68.

Goguen, J. A. (1969). The logic of inexact concepts. *Synthese, 19,* 325–373.

Gulliksen, H. (1959). Mathematical solutions for psychological problems. *American Scientist, 47,* 178–201.

Hempel, C. G. (1939). Vagueness and logic. *Philosophy of Science, 6,* 163–180.

Hersh, H. M., & Caramazza, A. (1976). A fuzzy set approach to modifiers and vagueness in natural language. *Journal of Experimental Psychology: General, 105,* 254–276.

Hersh, H. M., Caramazza, A., & Brownell, H. H. (1979). Effects of context on fuzzy membership functions. In M. M. Gupta, R. K. Ragade, & R. R. Yager (Eds.), *Advances in fuzzy set theory and application* (pp. 389–408). Amsterdam, The Netherlands: North-Holland.

Isaac, P. D. (1970). Linear regression, structural relations, and measurement error. *Psychological Bulletin, 74,* 213–218.

Jensen, R. E. (1984). An alternative scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology, 28,* 317–332.

Johnson, C. R., Beine, W. B., & Wang, T. Y. (1979). Right–left asymmetry in an eigenvector ranking procedure. *Journal of Mathematical Psychology, 19,* 61–64.

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

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement: Vol. I.* New York: Academic Press.

Kuz'Min, V. B. (1981). A parametric approach to description of linguistic values of variables and hedges. *Fuzzy Sets and Systems, 6,* 27–41.

Labov, W. (1973). The boundaries of words and their meanings. In C.-J. N. Bailey & R. W. Shuy (Eds.), *New ways of analyzing variation in English* (pp. 340–373). Washington, DC: Georgetown University Press.

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

MacVicar-Whelan, P. J. (1978). Fuzzy sets, the concept of height, and the hedge *very. IEEE Transactions on Systems, Man, and Cybernetics, SMC-8,* 507–511.

Miyamoto, J. M. (1983). An axiomatization of the ratio/difference representation. *Journal of Mathematical Psychology, 27,* 439–455.

National Research Council Governing Board Committee on the Assessment of Risk. (1981). *The handling of risk assessments in NRC Reports.* Washington, DC: U.S. National Research Council.

Norwich, A. M., & Turksen, I. B. (1982). The fundamental measurement of fuzziness. In R. R. Yager (Ed.), *Fuzzy set and possibility theory* (pp. 49–60). New York: Pergamon Press.

Norwich, A. M., & Turksen, I. B. (1984). A model for the measurement of membership and the consequences of its empirical implementation. *Fuzzy Sets and Systems, 12,* 1–25.

Oden, G. C. (1977a). Fuzziness in semantic memory: Choosing exemplars of subjective categories. *Memory & Cognition, 5,* 198–204.

Oden, G. C. (1977b). Integration and fuzzy logical information. *Journal of Experimental Psychology: Human Perception and Performance, 3,* 565–575.

Oden, G. C. (1981). A fuzzy propositional model of concept structure and use: A case study in object identification. In G. W. Lasker (Ed.), *Applied systems research and cybernetics* (Vol. 6, pp. 2890–2897). Elmsford, NY: Pergamon Press.

Pepper, S., & Prytulak, L. S. (1974). Sometimes frequently means seldom: Context effects in the interpretations of quantitative expressions. *Journal of Research in Personality, 8,* 95–101.

Rubin, D. C. (1979). On measuring fuzziness: A comment on "A fuzzy set approach to modifiers and vagueness in natural language." *Journal of Experimental Psychology: General, 108,* 486–489.

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.* New York: McGraw-Hill.

Saaty, T. L., & Vargas, L. G. (1984). Inconsistency and rank preservation. *Journal of Mathematical Psychology, 28,* 205–214.

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.

Sjöberg, L. (1980). Similarity and correlation. In E. J. Lantermann & H. Feger (Eds.), *Similarity and choice* (pp. 70–87). Bern, Switzerland: Hans Huber.

Skala, H. J., Termini, S., & Trillas, E. (1984). *Aspects of vagueness.* Dordrecht, The Netherlands: Reidel.

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

Wallsten, T. S. (1971). Subjectively expected utility theory and subjects' probability estimates: Use of measurement-free techniques. *Journal of Experimental Psychology, 88,* 31–40.

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

Watson, S. R., Weiss, J. J., & Donnell, M. L. (1979). Fuzzy decision analysis. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-9,* 1–9.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8,* 338–353.

Zadeh, L. A. (1975). The concept of a linguistic variable and is application to approximate reasoning (II). *Information Sciences, 8,* 301–357.

Zimmer, A. C. (1983). Verbal vs. numerical processing of subjective probabilities. In R. W. Scholz (Ed.), *Decision making under uncertainty* (pp. 159–182). Amsterdam, The Netherlands: North-Holland.

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

Zysno, P. (1981). Modeling membership functions. In B. B. Rieger (Ed.), *Empirical semantics* (pp. 350–375). Bochum, West Germany: Brockmeyer.

Appendix A

Scaling Models

One approach to applying the metric scaling models proceeds as follows. Consider the difference model first. Assume that for a given expression W and probability pair $p_i p_j$, the subject places the arrow on the response line such that the difference in the distances of the arrow from the two ends is inversely proportional to the difference in the degrees to which W describes p_i and p_j. Thus, the response $R_W(ij)$ can be converted to a difference score D for purposes of scaling:

$$D_W(ij) = 2R_W(ij) - 1. \quad \text{(A1)}$$

The proportionality assumption plus an assumed error component yield

$$D_W(ij) = \alpha_W[\mu_W(p_i) - \mu_W(p_j)] + \epsilon_{Wij}, \quad \text{(A2)}$$

with $\alpha_W > 0$. Considering the full matrix of difference scores for phase W, a least squares estimate of $\mu_W(p_i)$ is obtained by taking row means. In other words, from Equation A2,

$$\hat{\mu}_W(p_i) = \sum_j D_W(ij)/n, \quad \text{(A3)}$$

where n is the size of the matrix and $\alpha_W = 1$. The scale values, of course, are unique up to the transformation, $\mu'_W(p_i) = \alpha_W\mu_W(p_i) + \beta_W$ and can easily be rescaled to be positive with a maximum at 1.

For the ratio-scaling models, it is assumed that the arrow is placed on the response line such that the ratio of the distances of the arrow from the two ends is inversely proportional to the ratio of the degrees to which W describes p_i and p_j. Thus, the response R is converted to a ratio score S:

$$S_W(ij) = R_W(ij)/[1 - R_W(ij)], \quad \text{(A4)}$$

with $R_W(ij) \neq 0, 1$. Now the proportionality assumption plus assumed error yields

$$S_W(ij) = \alpha_W\epsilon_{Wij}\mu_W(p_i)/\mu_W(p_j), \quad \text{(A5)}$$

with $\alpha_W > 0$. The geometric means (GMs)

$$\hat{\mu}_W(p_i) = [\prod_j S_W(ij)]^{1/n}, \quad \text{(A6)}$$

with $\alpha_W = 1$, are least squares estimates of the logarithms of the scale values (Torgerson, 1958), assuming that the matrix is reciprocal, that is, $S_W(ij) = 1/S_W(ji)$. The resulting scale values are unique up to the transformation, $\mu'_W(p_i) = \alpha_W\mu_W(p_i)^{\beta_W}$, with $\alpha_W, \beta_W > 0$.

An alternative ratio-scaling procedure, anticipated by Gulliksen (1958), also requires a reciprocal matrix. Scale values can be obtained from the matrix by an eigenvalue–eigenvector decomposition, obtaining either a normalized right eigenvector (RE) (Saaty, 1977, 1980), a normalized left eigenvector (LE) (Johnson et al., 1979), or the geometric mean of the two eigenvectors (ME) (Budescu, 1984). If a reciprocal matrix is consistent that is, for any three entries, $S(ij)$, $S(jk)$, and $S(ik)$, $S(ik) = S(ij)S(jk)$, then GM, RE, LE, and ME all yield the same scales. Otherwise they do not, and there is currently some controversy concerning the merits of each solution. Properties of the various solutions have been investigated mathematically (e.g., De Jong, 1984; Jensen, 1984; Saaty & Vargas, 1984) and with Monte Carlo procedures (e.g., Budescu, Zwick, & Rapoport, 1986; Johnson et al., 1979; Crawford & Williams, 1985).

Appendix B

Cross Validation

The predictions are derived here only in terms of scales obtained from Equation A6, because ultimately those were the values with which they were tested. Consider an experimental trial with probability p and terms W_i and W_j, for which the subject sets the arrow at location $R_p(ij)$. (Note the shift in notation to correspond with the change in the structure of a trial. We are now assuming a fixed p and a set of terms $T = \{W_1, \ldots, W_m\}$.) The response value $R_p(ij)$ is transformed to $S_p(ij)$ by Equation A4 (with indices suitably changed). If the previously derived scale values $\mu_{W_i}(p)$ and $\mu_{W_j}(p)$ represent the degree to which p is a member of W_i and W_j, respectively, then it should be the case that

$$S_p(ij) = \delta\mu_{W_i}(p)^{\beta_i}/\mu_{W_j}(p)^{\beta_j}, \quad \text{(B1)}$$

where $\delta, \beta_i, \beta_j > 0$. For clarity, the scaling parameters are not fully subscripted. But they have been included in Equation B1, because it is important to note what assumptions are being made about them.

Consider first a fixed pair of phrases W_i and W_j and various p, all of which have nonzero membership functions in W_i and W_j. If it is assumed that $\beta_i = \beta_j = 1$, then from Equation B1, the $S_p(ij)$ should be a linear function of the ratios of the derived membership functions. This prediction was tested in Experiment 1.

Now consider a fixed probability p with various phrases $W_1, W_2, \ldots, W_m$. In this case, $S_p(ij)$ is a linear function of the ratio of the derived membership functions only if it is assumed that $\delta = \beta_i = \beta_j = 1$ for all W_i and W_j. This prediction is tested in Experiment 2.

A very strong prediction emerges if for a given p there is a left side × right side, $T \times T$ factorial design, in which T is the vector of probability terms. The data matrix for each p can be scaled in a manner analogous to that described with Equations A4, A5, and A6. The resulting scale values, $\alpha_p(W)$, are unique up to a power transformation, $\mu'_p(W) = \alpha_p\mu_p(W)^{\beta_p}$, with $\alpha_p, \beta_p > 0$. Omitting subscripts, on the reasonable assumption that $\mu_p(W)$ and $\mu_W(P)$ both represent the same vagueness construct, it is easy to show that the two sets of derived values should be related by a power function. This is done by setting the generalized forms of the scale values equal to each other,

$$\alpha_W\mu_W(p)^{\beta_W} = \alpha_p\mu_p(W)^{\beta_p}, \quad \text{(B2)}$$

and solving for $\mu_W(p)$. Setting $\alpha = \alpha_p/\alpha_W$ and $\beta = \beta_p/\beta_W$, the result is

$$\mu_p(W) = \alpha\mu_W(p)^{\beta}, \quad \text{(B3)}$$

with $\alpha, \beta > 0$. Equation B3 is tested in Experiment 2.

Received February 3, 1986
Revision received April 22, 1986 ■