Mathl Modelling, Vol. 9, No. 6, pp. 397–417, 1987
Printed in Great Britain. All rights reserved

0270-0255/87 $3.00 + 0.00

DIRECT AND INDIRECT SCALING OF MEMBERSHIP FUNCTIONS OF PROBABILITY PHRASES†

A. RAPOPORT, T. S. WALLSTEN and J. A. COX
Department of Psychology, University of North Carolina, Davie Hall 013-A, Chapel Hill, NC 27514, U.S.A.

Abstract—A crucial issue in the empirical measurement of membership functions is whether the degree of fuzziness is invariant under different scaling procedures. In this paper a direct and an indirect procedure, magnitude estimation and graded pair-comparison, are compared in the context of establishing membership functions for probability phrases such as *probable*, *rather likely*, *very unlikely*, and so forth. Analyses at the level of individual respondents indicate that: (a) membership functions are stable over time; (b) functions for each phrase differ substantially over people; (c) the two procedures yield similarly shaped functions for a given person–phrase combination; (d) the functions from the two procedures differ systematically, in that those obtained directly dominate, or indicate greater fuzziness than do those obtained indirectly; and (e) where the two differ the indirectly obtained function may be the more accurate one. A secondary purpose of the paper is to evaluate the effects of the modifiers *very* and *rather*. *Very* has a general intensifying effect that is described by Zadeh's concentration model for 7 subjects and by a shift model for no one. The effects of *rather* are unsystematic and not described by any available model.

INTRODUCTION

The most fundamental concept in the theory of fuzzy sets is that of a fuzzy subset A of a universe of discourse U, with A characterized by a membership function $\mu_A(x)$ that associates with each point $x \in U$ its "grade of membership" in A. Usually, but not necessarily, $\mu_A(x)$ is assumed to range in the interval $[0, 1]$. The numbers 0 and 1 correspond then, as assumed in this paper, to non-membership and full membership, respectively.

The primary goal of the present paper is to compare membership functions constructed by two alternative scaling procedures, and the secondary goal is to evaluate the effects of certain modifiers on the membership functions. However, it is necessary first to address four issues and their experimental implications—the effects of context, the subjective nature of membership functions, variability in membership judgments and the problem of scale type. We discuss the first two issues only briefly but the latter two in more detail.

It is generally recognized that membership functions depend, at least to some degree, on the context in which they are measured. For example, the grade of membership of a 60-year-old woman in the class of *old women* may vary from country to country depending on life expectancy. And the grade of membership of an event whose probability of occurrence is 0.2 in the class of *unlikely* events probably depends on whether the event gives rise to favorable or unfavorable outcomes. Without an exact specification of the context and an experimental investigation of the effects of context on vagueness and fuzziness, comparisons of membership functions constructed under different experimental setups may be grossly misleading. Context is held fixed throughout the present study.

The second issue stems from the fact that grades of membership are generally subjective. An important experimental implication of this subjectivity, which has not always been followed, is that within- rather than between-person designs should be implemented because averaging over membership functions from different individuals, even when the context is fixed, is meaningless.

But even if these two relatively simple issues are properly addressed, a question arises as to whether the value of $\mu_A(x)$ can be determined accurately. The obvious paradoxical conclusion that the nature of fuzziness does not permit precise measurement of grades of membership led to the solution of creating a "type-2" fuzzy set by laying a second level of membership over the original "type-1" membership function, then creating a "type-3" fuzzy set by laying a third level

† The views, opinions and findings contained in this paper are those of the authors and should not be construed as official Department of the Army position, policy or decision.

of membership over the "type-2" membership, and so on. As noted by Norwich and Turksen,

> "The resulting infinite regress led to discouragement about whether a membership function could ever be meaningfully constructed" [1, p. 68].

The resolution of this paradox proposed by Norwich and Turksen is similar in spirit to the philosophy underlying the experimental testing of algebraic models in decision making and other areas of cognitive psychology. It is based on the recognition that the infinite regress model outlined above treats the grade of membership on every level as deterministic, whereas in any actual scaling of subjective characteristics, determinism is impossible. As argued by Norwich and Turksen [2] and demonstrated experimentally by them [3] and Wallsten *et al.* [4], the repeated elicitation of grades of membership for a given subject in a well-defined context by some psychophysical scaling procedure (e.g. magnitude estimation, paired-comparison) yields variability in measurement. Norwich and Turksen claim that this variability "embodies all the uncertainty or imprecision in this value (and, equivalently, all the information about this value) which exists in the subject's mind" [1, p. 69]. One may, therefore, use the mean of a set of non-deterministic numerical responses as the grade of membership of a type-1 fuzzy set with no need to amass higher levels of deterministic membership to model the subject's fuzziness.

The logic used by Norwich and Turksen to resolve the infinite regress paradox may be used against the form of their argument leading to their claim that the membership function is measurable at most on an interval scale. Norwich and Turksen have correctly contended that since a subject is not precisely sure of the meaning of a subjective concept, it is contradictory to the notion of fuzziness to assert that this concept partitions the universe of discourse precisely into three regions where $\mu_A(x) = 0$, $\mu_A(x) = 1$ and $0 < \mu_A(x) < 1$, respectively, the boundaries of which can be precisely determined (to within the limits of physical discriminability) [2, 5]. Norwich and Turksen base this claim on the observation that the condition for a natural zero for a membership structure will be most likely not be met by a subject. Their claim, however, presupposes deterministic responses, whereas in practice the boundaries of the three regions determined by the membership function are determined statistically, just as are absolute thresholds and difference thresholds in psychophysical scaling. Precisely as one may employ the mean of a set of non-deterministic responses as the grade of membership of a particular element, one may use the mean of non-deterministic responses to determine the region's boundaries.

Our argument above illustrates the point that questions regarding response variability are logically distinct from those regarding scale type. With respect to this final issue, Gougen [6] has stated that the membership function can be no stronger than an ordinal scale, Norwich and Turksen [2, 5] have claimed interval scale strength for the membership function, Saaty [7] has espoused the ratio scale, whereas Thole *et al.* [8] have asserted that grades of membership are measurable on an absolute scale. To resolve this controversy, it must be recognized that scale type is not arbitrary, but rather it minimally depends on properties of the particular measurement procedure utilized. Additional assumptions, ideally ones that are testable, then may be evoked on theoretical or pragmatic grounds to yield stronger scales.

Experimental measurement of membership functions

Although much has been written about the measurement of fuzziness and vagueness, empirical work has been relatively sparse [e.g. 3–5, 8–15]. Moreover, with the exception of an experiment by Norwich and Turksen [3, 5], the empirical studies have each employed only a single scaling procedure to construct membership functions. They have not determined whether the resulting membership functions are invariant under various scaling procedures and if not, what are the relationships between membership functions obtained from the same subject by different procedures. The investigation of the invariance of fuzziness under different scaling procedures is as important to fuzzy set theory as is the study of invariance of other subjective perceptions to psychophysical theories. For example, discussing the theory of signal detection and commenting on three different procedures, the yes–no, rating and forced-choice tasks, for measuring the detection of weak signals in noise, Green and Swets wrote:

> "Comparing the results obtained from these different procedures is extremely important because such comparisons provide the major test of the validity of the decision-theory analysis. If the analysis is verified, it yields measures of the detectability

> of the signal that are independent of the procedure used to estimate these measures. Therefore, this analysis holds forth the possibility of psychophysical relations which are independent of procedure, a goal more often hoped for than achieved." [16, p. 31]

Once "fuzzy set theory" is substituted for the "decision-theory analysis" and "fuzziness of the concept" for "detectability of the signal", this statement holds true also for the measurement of imprecise concepts by fuzzy set theory.

The exception mentioned above is a study by Norwich and Turksen [3], who employed two scaling procedures, direct rating (also known as magnitude estimation) and reverse rating. For the direct rating procedure, the subject was presented one at a time with a life-sized wooden male figure of adjustable height or a cardhouse with adjustable dimensions. The subject was then asked to rate the tallness of the man or the aesthetic pleasantness of the house by adjusting the position of a pointer on a horizontal line segment. The left end-point of the segment corresponded to all men/houses which he or she felt were definitely not tall/pleasing, the right end-point to all those that were definitely tall/pleasing and the line segment between these two end-points was interpreted as representing how strongly the subject agreed that the man/house was tall/pleasing. For the reverse rating procedure, the pointer was randomly set to some position in the segment and the subject adjusted the height/dimensions of the man/house to correspond to the degree of membership indicated by the position of the pointer. In all cases, any particular stimulus was direct or reverse rated a minimum of nine times by the subject and the average was then used to assess the membership functions of "tall" and "aesthetically pleasing". Norwich and Turksen report that

> "It has been found for all subjects that direct and reverse ratings appear to yield the same membership function (this conclusion will be checked by statistical hypothesis testing) and that the axioms of the algebraic-difference structure are obeyed." [5, p. 12]

Detailed comparisons of the outomes of the two scaling procedures or of the axiom tests have not been provided. Furthermore, the algebraic-difference axioms refer to a comparison procedure, and therefore their satisfaction does not validate the very different direct or reverse rating procedures.

Major goals of the present study

Both the direct rating and reverse rating are examples of direct psychophysical scaling techniques, which have yielded remarkably stable relations within psychology between physical scale values and estimates of psychological magnitudes over a wide variety of sensory continua. However, as also noted by Thole *et al.* [8], psychological scales developed with direct methods may be distorted by a number of response biases. Because of the difficulties that have plagued many attempts to measure psychological magnitudes directly, some psychophysicists, notably Fechner [17] and Thurstone [18], have turned to discriminability or confusability between stimuli as a method for inferring psychological magnitudes indirectly [19]. Thurstone's model, which was originally formulated as a psychophysical theory, and signal detection theory [e.g. 16], which emphasizes both judgmental as well as sensory determinants, are examples of indirect psychophysical analyses. Recognizing some of the major advantages of indirect scaling procedures, Thole *et al.* remarked that

> "As yet, however, no practical indirect technique, the result of which is more than an interval scale, is available." [8, p. 170]

However, this situation was addressed at least in part, by the work of Wallsten *et al.* [4], which utilized a graded pair-comparison procedure in two experiments to test the algebraic-difference axioms and to measure membership on an interval scale. The functions had interpretable shapes and predicted an independent set of judgments.

A major purpose of the present study is to compare two scaling procedures for measuring membership functions, a direct rating and the indirect graded pair-comparison procedure developed by Wallsten *et al.* [4]. The theoretical significance of investigating the invariance of fuzziness—or lack of it—under different scaling procedures has already been mentioned above. But there is also a practical reason for the comparison. Indirect scaling procedures are notoriously more laborious than direct ones [8], requiring a multiple of $n(n - 1)/2$ rather than of n judgments. Our comparison of the two scaling procedures should show whether the membership function remains invariant when the less taxing procedure is used.

The imprecise concepts whose membership functions were to be measured consisted of probability phrases such as *probable*, *unlikely*, *possible*, and so forth, with which most people, including experts in medical diagnosis, military intelligence and weather forecasting, generally prefer communicating their uncertain opinions. Wallsten *et al.* [4] also utilized probability phrases. A second purpose of the present study is to examine the effects of modifiers like *very* and *rather* on the shape and interpretation of membership functions for probability phrases.

METHOD

Four groups of 5 subjects each were employed, each responding to a different set of five probability phrases. Table 1 shows the probability phrases for each group. Altogether the membership functions of 13 probability phrases were scaled. The phrase *possible* was presented to all 20 subjects; the 4 phrases *probable*, *improbable*, *likely* and *unlikely* were each presented to 10 subjects; and the same 4 phrases each modified by *rather* or *very* were each presented to 5 subjects. In an attempt to minimize context effects, each group was presented with the same mix of "high" and "low" phrases. Each had a high probability term (*probable* or *likely*), a low probability term (*improbable* or *unlikely*) as well as one of those terms modified by *rather* and the other by *very*, plus a single "neutral" phrase (*possible*). Antonyms were not presented in the same group.

Subjects

Subjects were social science and business graduate students at the University of North Carolina at Chapel Hill. They were recruited by placing notices in graduate student mailboxes in the School of Business and the Departments of Political Science, Economics, Library Science, Psychology and Sociology. None of the subjects had participated in any similar experiments on the measurement of membership functions. The notices described the general nature of the study and promised the subjects $25 each for three sessions of approx. 45 min each. Twenty native speakers of English were randomly assigned to Groups 1, 2, 3 or 4. As explained above in conjunction with Table 1, the groups differed only in terms of the phrases they judged.

General procedure

Subjects were run individually for a practice session followed by two data sessions, with the sessions scheduled generally 1–2 days apart. The experiment was controlled by an IBM PC with stimuli presented on a color monitor and responses made with a joystick. During the practice session all the subjects judged the probability phrases *tossup*, *good chance* and *poor chance*, whereas during Sessions 2 and 3 they judged the phrases indicated in Table 1.

Each session consisted of three parts, all of which employed spinners drawn on the computer monitor. Part 1 employed one spinner, Part 2 employed two spinners and Part 3 employed six spinners, as described below. Each spinner was divided radially into two sectors, one red and the other yellow. The subjects were instructed to imagine a pointer over each spinner that could be spun so that it randomly lands over either the red or yellow sectors. Thus, without ever using numbers, each spinner displayed a probability of the spinner landing on yellow.

At the beginning of the experiment, subjects were instructed that the purpose of the study was "to determine how people use and understand non-numerical probability phrases, such as *good chance*, *poor chance* and *tossup*, for communicating judgments about uncertainty". Subjects were told that there were no right or wrong answers, just individual judgments.

Table 1. Probability phrases used in the experiment

Group 1	Group 2	Group 3	Group 4
Very probable	*Probable*	*Very likely*	*Rather likely*
Probable	*Rather probable*	*Likely*	*Likely*
Possible	*Possible*	*Possible*	*Possible*
Unlikely	*Unlikely*	*Improbable*	*Improbable*
Rather unlikely	*Very unlikely*	*Rather improbable*	*Very improbable*

Subjects then were instructed how to use a joystick to move an arrow to the left or right on a straight line and how to push buttons on the joystick assembly to register their responses. Several practice trials followed to provide the subjects experience with the joystick controls.

The subsequent instructions were divided into three parts, each corresponding to a different part of the session. The instructions for Part 1 were intended to elicit membership functions for the same phrases by a direct scaling procedure referred to hereafter as *direct estimation* (DE). The instructions for Part 2 involved the assessment of membership functions of the five probability phrases by the indirect scaling procedure employed by Wallsten *et al.* [4], referred to hereafter as *pair-comparison* (PC). The instructions for Part 3, which was designed to allow a comparison between the DE and PC procedures, involved a measurement of membership functions of the same five probability phrases on an ordinal scale only. A *rank ordering* (RO) scaling procedure was used for this purpose. Because the results depend strongly on characteristics of the experimental procedures, each part will now be described in more detail.

Part 1

Five probability phrases were presented in this and the subsequent two parts (cf. Table 1). Associated with each phrase were six probabilities displayed as relative areas of yellow on a spinner. The probabilities associated with the "low" terms *(improbable, rather improbable, very improbable, unlikely, rather unlikely, very unlikely)* were 0.05, 0.14, 0.23, 0.32, 0.41 and 0.50; the probabilities associated with each of the "high" terms *(probable, rather probable, very probable, likely, rather likely, very likely)* were 0.50, 0.59, 0.68, 0.77, 0.86 and 0.95; and the probabilities associated with the single "neutral" phrase *(possible)* were 0.32, 0.41, 0.50, 0.59, 0.68 and 0.77. Thus, each membership function was approximated by six points with a difference of 0.09 between adjacent points on the probability continuum. Practical considerations determined that six points be used to approximate what are essentially continuous membership functions, and the previous results of Wallsten *et al.* [4] determined the probability ranges for the "low" [0.05, 0.50], "high" [0.50, 0.95] and "neutral" [0.32, 0.77] probability phrases.

Each probability phrase was presented twice. Thus, Part 1 consisted of 5 phrases × 6 probabilities per phrase × 2 replications for a total of 60 trials. The probability phrase was always displayed at the top of the screen and a spinner divided into red and yellow sectors was drawn below it. The 60 phrase × probability combinations were presented in a random order.

The instructions for the DE procedure read in part:

> "On each trial in Part 1, a probability phrase will appear at the top of the computer screen. One spinner will be drawn directly below this phrase. You are to indicate how well the probability phrase describes the probability of landing on yellow for the displayed spinner. If you think that the spinner probability is very well described by the phrase, move the arrow to the right. If the spinner probability is not at all well described by that phrase move the arrow far to the left. The relative location of the arrow on the line should correspond to how well (right) or how poorly (left) the phrase describes the probability."

The joystick was used to move the arrow on the screen, and a button on the joystick box was used to register the response when the arrow was suitably placed. The arrow could be positioned at any of 200 equally spaced locations on the line.

Part 2

On each trial a single probability phrase was written at the top of the screen, and *two* spinners displaying different probabilities of yellow were drawn below it, one on the left and one on the right. The instructions for the PC procedure read in part:

> "In this part of the experiment you are to move the arrow along the line towards the spinner which is best described by the phrase at the top of the screen. The distance you move the arrow toward one of the spinners should reflect your confidence in that judgment. So if you think that the probability of yellow on one of the spinners is very much better described by the phrase than is the probability of yellow on the

> other spinner, move the arrow all the way to that end of the line. If the phrase describes the probability on one spinner only slightly better than the other, move the arrow just slightly off center."

Each phrase was associated with the same six probabilities as described in Part 1. However, rather than presenting each phrase–probability combination twice as in Part 1, each phrase was now presented once with each of the $6 \times 5/2 = 15$ spinner pairs, for a total of 75 trials. Phrases and spinner pairs were presented in a random, not a blocked order.

Part 3

On each trial a single probability phrase was printed at the top of the screen and *six* spinners each displaying a different probability of yellow were drawn below it. The subject's task on each trial was "to rank order the six spinners according to how well they are described by that phrase." This was done by moving a cursor to the spinner best described by the phrase and then pressing a button to register the response. Once the response was registered, the designated spinner vanished from the screen. The same procedure was repeated six times until all six probabilities were rank ordered. Each of the five phrases was presented twice.

For 10 of the 20 subjects (2, 3, 2 and 3 subjects in Groups 1, 2, 3 and 4, respectively) the DE procedure preceded the PC procedure in Session 2, whereas for 10 other subjects the order of Parts 1 and 2 was reversed in that session. Part 3 always followed Parts 1 and 2. Sessions 2 and 3 were identical, except that the order of Parts 1 and 2 were reversed.

RESULTS

This section begins with an examination of the psychometric properties of the separate scales derived from the DE and PC procedures, following which the membership functions established by the two scaling procedures are compared. Effects of the modifiers *rather* and *very* are examined in the last part of this section.

The PC procedure

To describe the scaling of the responses in Part 2, several terms must first be defined. Recall that the probabilities on the left and right spinners were changed from trial to trial such that, ignoring order, all probability pairs were presented once according to a left side × right side, $P \times P$, factorial design in which $P = (p_1, \ldots, p_n)$, where the p_i $(i = 1, \ldots, n)$ denote specific probabilities of the spinners landing on yellow. We shall consider the bounded response line on the CRT to extend from 1 on the left to 0 on the right, and let $R_W(i, j)$ denote the response when probability p_i is represented by the left spinner, p_j is represented by the right spinner and the phrase W is displayed above them.

By entering the response $R_W(i, j)$ in cell (i, j) and $1 - R_W(i, j)$ in cell (j, i), an ordering is induced on the factorial design according to the degree that the left-hand probability is better described by the phrase than is the right-hand probability. If this ordering satisfies the axioms of an algebraic difference structure [20], 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 phrase W [4].

More specifically, let (p_i, p_j) be a cell in the $P \times P$ factorial design. Denote the rank ordering between any pair of cells by $\gtrsim_W$, where the subscript indicates that the ordering is for phrase W. Recall that the ordering is induced by placing an arrow on the response line, so that the further to the left an arrow is for a pair of probabilities (spinners), the higher in the rank ordering is the pair. Formally,

$$(p_i, p_j) \gtrsim_W (p_k, p_l) \qquad \text{iff } R_W(i, j) \geq R_W(k, l).$$

Krantz *et al.* [20] proved that if the ordered matrix $(P \times P, \gtrsim_W)$ satisfies several plausible axioms, then there exists a mapping μ_W from P into the real numbers such that

$$(p_i, p_j) \gtrsim_W (p_k, p_l) \qquad \text{iff } \mu_W(p_i) - \mu_W(p_j) \geq \mu_W(p_k) - \mu_W(p_l)$$

or, equivalently, such that

$$(p_i, p_j) \gtrsim_W (p_k, p_l) \qquad \text{iff } \mu_W(p_i)/\mu_W(p_j) \geq \mu_W(p_k)/\mu_W(p_l).$$

These two equations state that 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 derived scale values are unique up to a linear (for the difference representation) or a power (for the ratio representation) transformation. Normalized to be non-negative with an arbitrary maximum of 1, these scale values can be taken as the (discrete) membership function representing the degree to which each probability belongs to the vague concept defined by the probability phrase.

Computationally, the scale values under the difference model are obtained by taking arithmetic means of the rows of the $n \times n$ matrix $\mathbb{D} = \{D_W(i,j)\}$, where

$$D_W(i,j) = [R_W(i,j) - 0] - [1 - R_W(i,j)] = 2R_W(i,j) - 1, \tag{1}$$

whereas the scale values under the ratio model are obtained by taking the geometric means of the rows of the $n \times n$ matrix $\mathbb{S} = \{S_W(i,j)\}$, where

$$S_W(i,j) = [R_W(i,j) - 0]/[1 - R_W(i,j)] = R_W(i,j)/[1 - R_W(i,j)]. \tag{2}$$

To avoid division by zero in the latter equation, the responses $R_W(i,j) = 0, 1$ were set equal to 0.00215 and 0.9975, respectively.

For more details regarding these two methods of scaling see Wallsten *et al.* [4], and for a discussion of alternative ratio scaling procedures see De Jong [21], Jensen [22], Saaty [23, 24], Saaty and Vargas [25] and Crawford and Williams [26].

It should be noted that at an axiomatic level no distinction can be made between the difference and ratio representations unless different orderings appear under difference- and ratio-inducing conditions [27, 28]. This is so because any set of differences can be mapped into a set of ratios by taking logarithms, and conversely, any set of ratios can be mapped into a set of differences by exponentiating. However, the representations can be compared to each other in terms of the correlations computed separately for each model between the observed responses and the responses predicted from the derived scale values [4].

To compute these correlations, three steps were taken. First, we computed the measures $D_w(i,j)$ and $S_W(i,j)$ from equations (1) and (2) and subsequently completed the matrices $\mathbb{D}$ and $\mathbb{S}$ by inserting the complementary measures $D_W(j,i) = 1 - D_W(i,j)$ and $S_W(j,i) = 1/S_W(i,j)$ and placing 0s and 1s in the main diagonals of $\mathbb{D}$ and $\mathbb{S}$, respectively. Second, the matrices $\mathbb{D}$ and $\mathbb{S}$, computed separately for each subject, session and phrase, were each scaled according to the difference model and ratio model, respectively. Finally, the derived scale values were used to compute for each model separately the predicted responses, $R^*_W(i,j)$, and the correlations between $R_W(i,j)$ and $R^*_W(i,j)$ were then computed for each subject, session and phrase separately (each based on 15 pairs of observations).

The mean correlations between $R_W(i,j)$ and $R^*_W(i,j)$ over all 20 subjects are shown in Table 2. To maintain an equal number of cases for each mean, the phrases *probable* and *improbable* were pooled together (column 2) as were the two phrases *likely* and *unlikely* (column 3), the four phrases modified by *rather* (column 4), and the four phrases modified by *very* (column 5). Each mean in Table 2 is, therefore, based on 20 correlations, 1 per subject.

A 2 × 2 × 5, model (ratio, difference) × session (2, 3) × probability phrase (*possible*, *probable/improbable*, *likely/unlikely*, *rather* ( ), *very* ( )), ANOVA with repeated measures on all three factors was conducted on the correlations between $R_W(i,j)$ and $R^*_W(i,j)$. All three factors were highly significant ($F = 17.15$, $p < 0.01$, $F = 8.86$, $p < 0.01$, and $F = 4.08$, $p < 0.01$, for model, session and phrase type, respectively). The same ANOVA conducted on the z-transforms of the correlations yielded similar results. As shown in Table 2, the correlations between observed and predicted responses for the difference model exceeded on the average those for the ratio model; the correlations

Table 2. Mean correlations between observed and predicted responses by scaling model. probability phrase and session

Scaling model	Session	Probability phrase					Over phrases
		Possible	*Probable/improbable*	*Unlikely/likely*	*Rather* (·)	*Very* (·)	
Ratio	1	0.572	0.747	0.619	0.691	0.762	0.619
	2	0.710	0.731	0.700	0.757	0.751	0.653
Difference	1	0.691	0.799	0.760	0.753	0.752	0.763
	2	0.727	0.855	0.847	0.846	0.775	0.835

in Session 3 were higher on the average than those in Session 2; and the correlations for the probability phrase *possible* were lower on the average than the correlations for all other probability phrases.

The significant session effect may be due either to changes in the shapes of the membership functions from Session 2 to 3 or to fewer response errors due to learning. To choose between these two hypotheses, the reliability of the scaled values was assessed by computing the correlation between the derived scale values in Sessions 2 and 3 for each subject and model separately over the probability phrases. With 5 different phrases and 6 scale values per phrase, each correlation was based on 30 pairs of observations. Of the 20 correlations computed under the ratio model all but 1 were highly significant ($p < 0.01$). Similarly, all 20 correlations computed under the difference model were highly significant ($p < 0.01$). The mean correlations over subjects were 0.75 and 0.86 for the ratio and difference models, respectively, demonstrating high reliability of the derived scale values and providing additional evidence for the superiority of the difference model over the ratio model.

Based on the test results reported above, we decided to use the Session 3 judgments only, to take the scale values derived from the difference rather than the ratio model as the PC membership functions, and to analyze the various membership functions for each subject separately.

The DE procedure

For each session, the responses of the two replications were averaged and the five membership functions with six points per function were determined directly from the mean responses. Reliability of the responses was assessed by computing the correlation between the mean responses in Sessions 2 and 3 over the five phrases for each subject separately. As in the PC procedure, each correlation was based on 30 pairs of mean responses. All 20 correlations were highly significant ($p < 0.01$). The mean correlation was 0.86, equal to the mean correlation computed for the difference model under the PC procedure. Because of the evidence for session-to-session learning reported in the preceding section and the high correlation of the responses between sessions, it was decided to take the scale values from Session 3 as the DE membership functions.

The RO procedure

As in the DE procedure, the two rankings in each session were first averaged and then five membership functions (unique up to an order-preserving transformation) with six points per function were determined directly for each subject and session. Reliability of the rankings was assessed again by computing rank order correlations between the averaged responses in Sessions 2 and 3. All 20 correlations, each based on 30 pairs of observations, were highly significant ($p < 0.01$). The correlations ranged between 0.66 and 0.99 with a mean of 0.86, which, surprisingly, is exactly equal to the two previous reliability correlations.

Reasonableness of the membership functions

Concluding that the three scaling procedures are equally reliable, we next turn to the scale values to consider how reasonable they are as membership functions. For this purpose, the derived scale values from Session 3 were each normalized to have a maximum at 1, and were plotted separately for each phrase and each subject as a function of the spinner probabilities of landing on yellow. The solid lines in Figs 1–5 represent the membership functions elicited by the DE procedure and the dashed lines those elicited by the PC procedure. Figure 1 shows 20 pairs of membership functions for the phrase *possible*. Figure 2 portrays 10 pairs of functions for the phrase *probable* (subjects 1–10) and 10 for *improbable* (subjects 11–20). The membership functions for the pair of

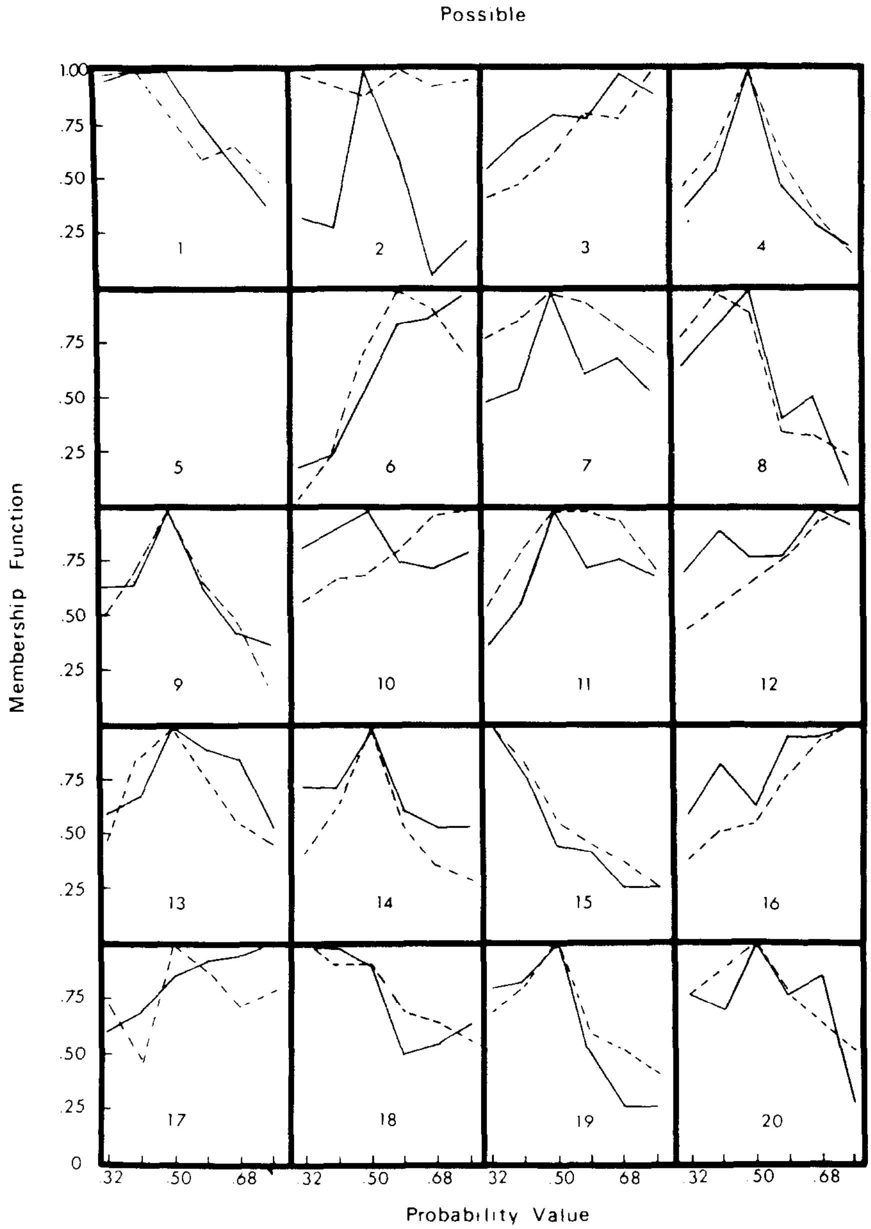

Fig. 1. Individual membership functions for *possible*. DE functions are indicated by solid lines and PC functions by dashed lines. Numbers in the panels refer to individual subjects.

antonyms *unlikely* (subjects 1–10) and *likely* (subjects 11–20) are shown in Figure 3. Figure 4 shows pairs of functions for the probability phrases *rather unlikely* (subjects 1–5), *rather probable* (subjects 6–10), *rather improbable* (subjects 11–15) and *rather likely* (subjects 16–20). Finally, Fig. 5 displays the membership functions for the phrases *very probable* (subjects 1–5), *very unlikely* (subjects 6–10), *very likely* (subjects 11–15) and *very improbable* (subjects 16–20). Recall that the probability values on the abscissas vary between probability phrases as indicated in the figures.

Table 3 summarizes the types of membership functions found over all subjects for low, neutral and high probability phrases. The two low phrases *improbable* and *unlikely* are grouped together as are the two high phrases *probable* and *likely*. The bottom four rows of Table 3 show the latter two pairs of phrases modified by *rather* and *very*. Membership functions can be characterized as either monotonic increasing, monotonic decreasing, single-peaked or other. In classifying the 100 pairs of functions we allowed one inversion per function provided its magnitude did not exceed

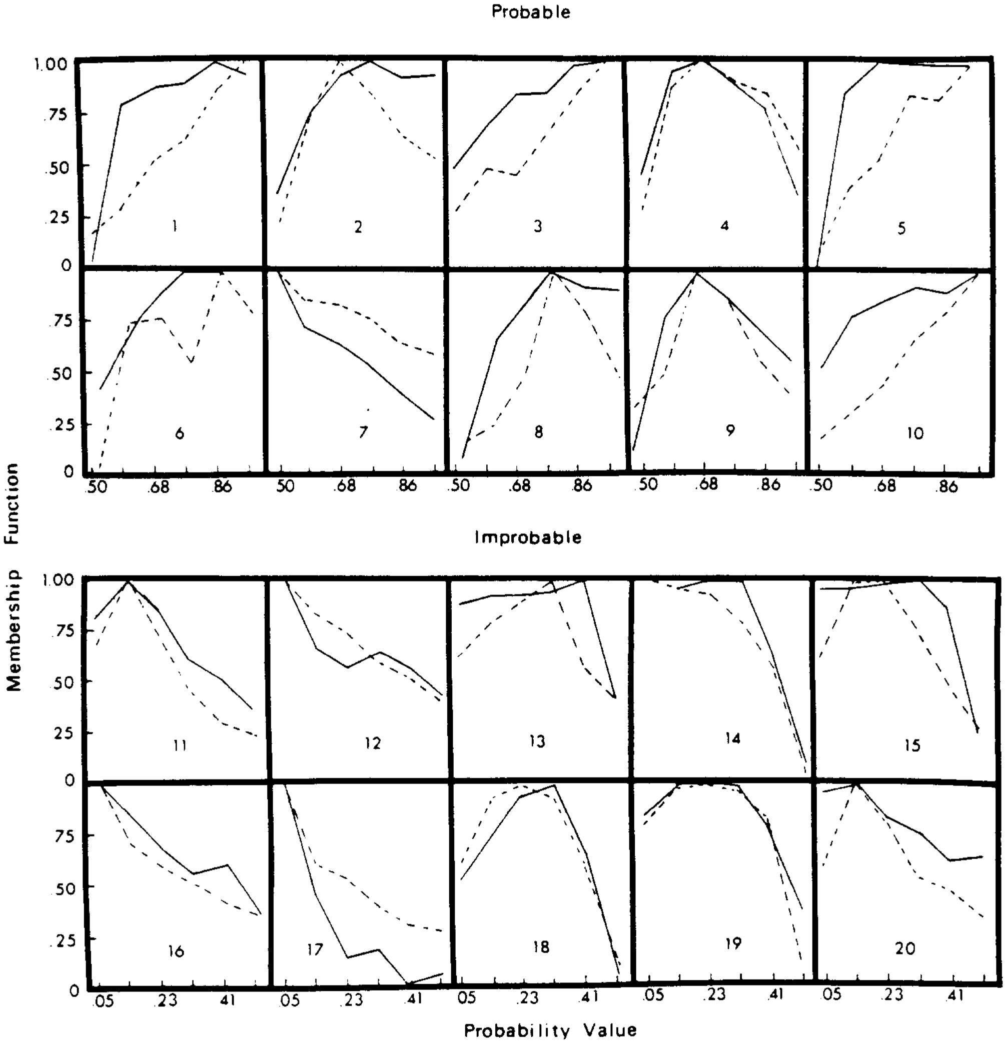

Fig. 2. Individual membership functions for *probable* (subjects 1–10) and *improbable* (subjects 11–20). DE functions are indicated by solid lines and PC functions by dashed lines.

0.10. For example, the PC-elicited function for subject 3 in Fig. 1 is classified as "monotonic increasing", whereas the DE-elicited function for subject 16 in Fig. 1 is classified as "other".

The monotonic and single-peaked functions might all be considered reasonable in terms of the underlying semantics, whereas those classified as "other" cannot easily be so considered. Table 3 shows that overall (allowing for no more than one minor inversion, which occurred in about 20% of all cases), 90% of the functions are reasonable by this criterion. Arguing again in terms of the underlying semantics, one would expect that monotonic functions for low phrases would be decreasing, whereas monotonic functions for high phrases would be increasing. Table 3 confirms this expectation remarkably well. Excluding the neutral phrase *possible*, there are altogether 46 monotonic functions for the high phrases, 44 of which are increasing, and there are 42 monotonic functions for the low phrases, all of which are decreasing.

One would also expect that phrases closer to the two extremes of the probability range [0, 1] will tend to have more monotonic than single-peaked functions, whereas phrases near the middle of the probability range will tend to have more single-peaked than monotonic functions [4]. The two modifiers *very* and *rather* can be represented as operators acting on membership functions [29]. Membership functions of phrases modified by *very* are expected to be more frequently monotonic than functions of phrases modified by *rather*. Table 3 confirms this prediction, too: of

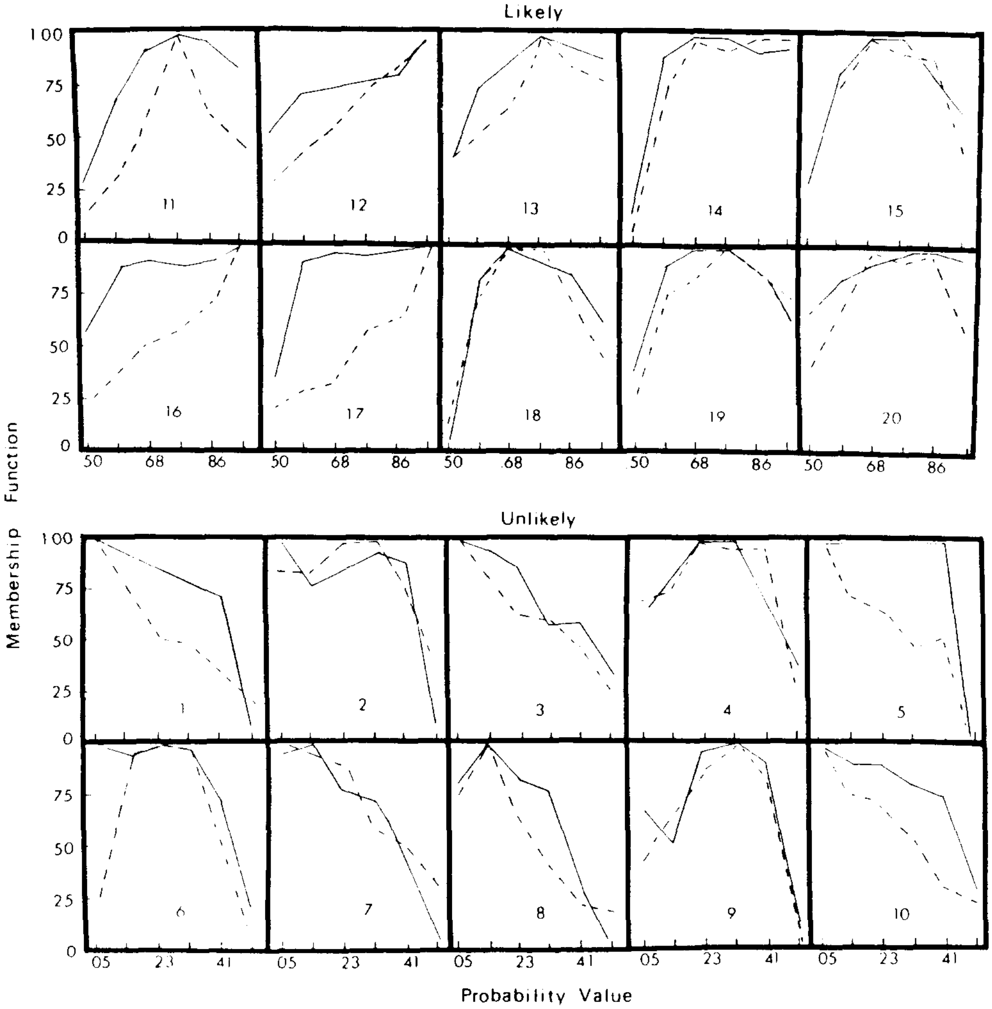

Fig. 3. Individual membership functions for *likely* (subjects 11–20) and *unlikely* (subjects 1–10). DE functions are indicated by solid lines and PC functions by dashed lines.

the 40 functions of phrases modified by *very*, 36 (90%) are monotonic; whereas of the 40 functions of phrases modified by *rather*, only 17 (42.5%) are monotonic.

Comparison of the PC and DE procedures

Concluding that most of the membership functions are not only stable over time but also reasonable and semantically interpretable, we turn next to a comparison of the functions elicited by the PC and DE procedures. An inspection of Figs 1–5 reveals that in virtually all cases the two procedures yield the same shape function. However, with the exception of the phrase *possible* in Fig. 1, the functions elicited by the DE procedure tend to lie above, or dominate, the functions elicited by the PC procedure. If two membership functions A and B are defined over the same support and function A has larger grades of membership than B, then the concept giving rise to function A is more fuzzy or vague than the one giving rise to function B.

The impression regarding dominance gained from inspecting Figs 1–5 is substantiated by the frequencies presented in Table 4. For the functions displayed in Figs 1–5, we say that function A dominates function B if $\mu_A(x) > \mu_B(x)$ for at least five of the six probability phrases in the common support of both functions, with equality holding only if $\mu_A(x) = \mu_B(x) = 1$. Table 4 shows that the functions elicited by the DE procedure dominate those elicited by the PC procedure in 52 of the 100 cases, that the reverse occurs in 12 cases only and that in 37 cases neither member of the pair of functions dominates the other. The null hypothesis of equal proportions of domination is rejected

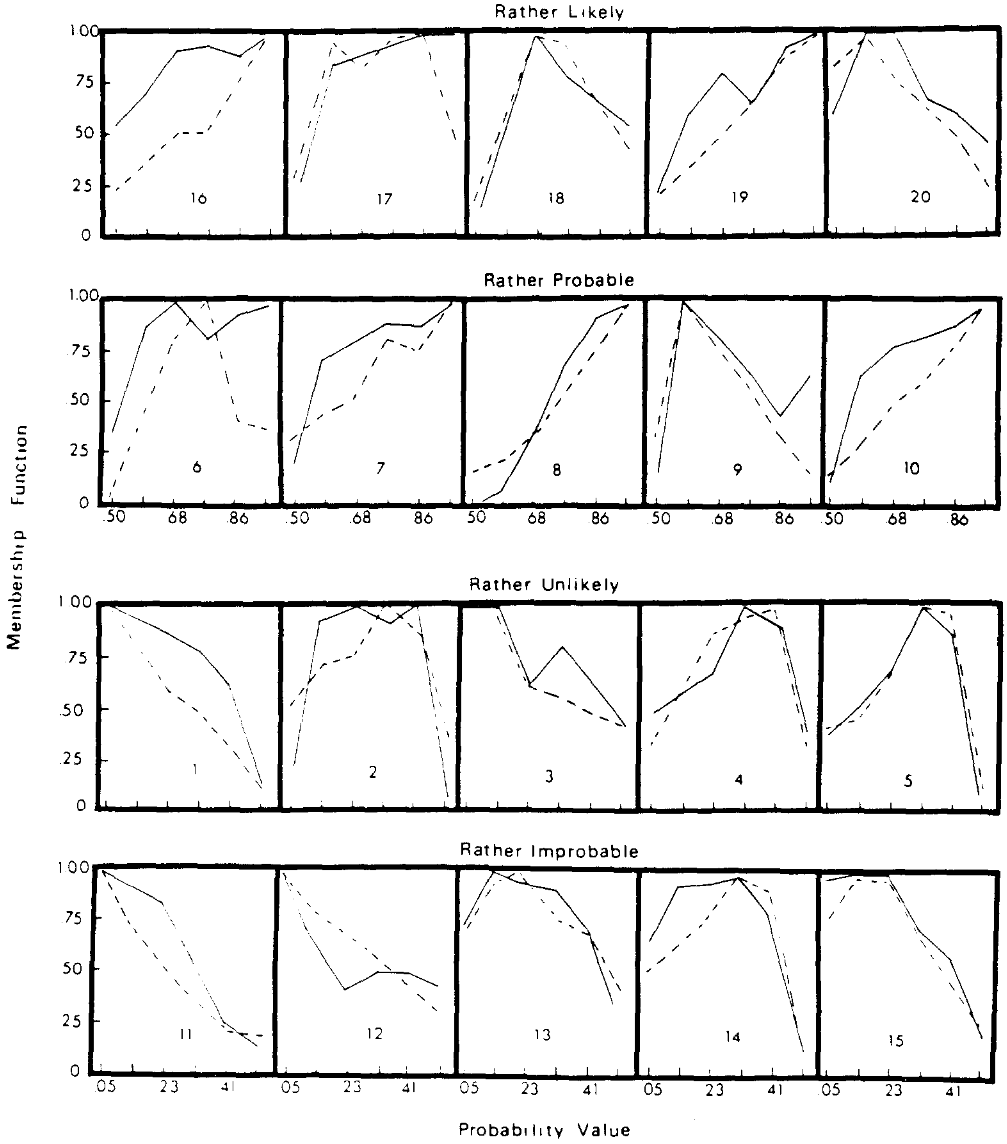

Fig. 4. Individual membership functions for *rather likely* (subjects 16–20), *rather probable* (subjects 6–10), *rather unlikely* (subjects 1–5) and *rather improbable* (subjects 11–15). DE functions are indicated by solid lines and PC functions by dashed lines.

for the total frequencies over phrases (right-hand column of Table 4) by a sign test ($p < 0.01$). It is also rejected for each of the five phrase types in Table 4 except *possible*.

Table 4 shows that the DE and PC procedures do not elicit the same membership functions for the five phrases under consideration. Although the functions are equally stable over time, equally reasonable and of roughly the same shape, those elicited by the DE procedure enhance the magnitude of the grade of membership or, equivalently, those elicited by the PC procedure reduce it.

To choose between the two procedures, we invoked the membership functions elicited by the RO procedure, which, as may be recalled, are measurable on an ordinal scale only. For each subject separately, a rank order correlation was computed between the membership functions elicited by the DE and RO procedures in Session 3. The results were pooled over the five phrases. All 20 correlations, each based on 30 pairs of observations, are highly significant ($p < 0.01$), and their mean value is 0.82. Similar rank order correlations were then computed between the functions elicited by the PC and RO procedures. Again, all 20 correlations are highly significant ($p < 0.01$), but now the mean value is slightly greater at 0.88. A sign test was used to test the null hypothesis

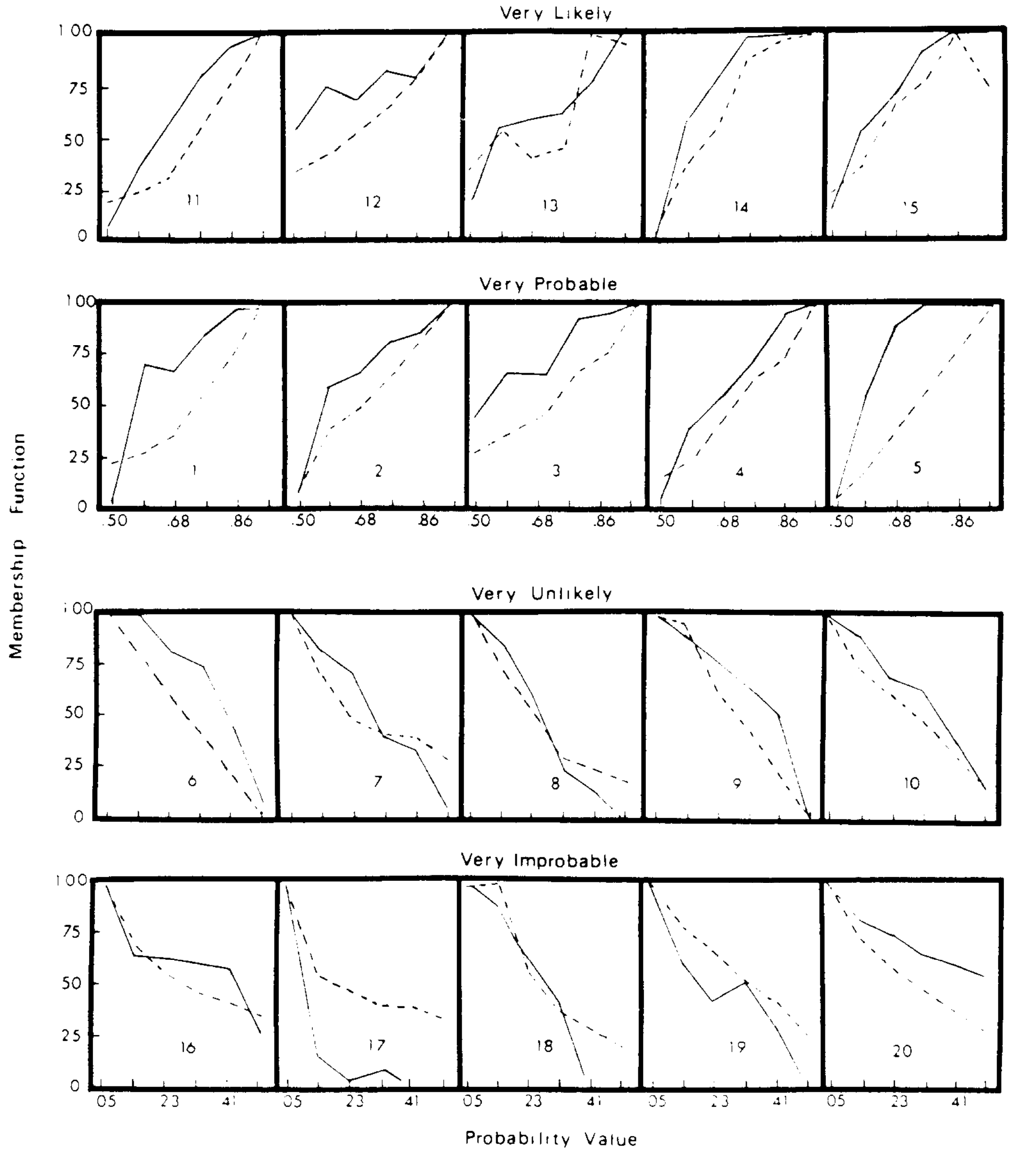

Fig. 5. Individual membership functions for *very likely* (subjects 11–15), *very probable* (subjects 1–5), *very unlikely* (subjects 6–10) and *very improbable* (subjects 16–20). DE functions are indicated by solid lines and PC functions by dashed lines.

Table 3. Classification of the membership functions in Session 3 by shape

Probability phrase	Monotonic increasing	Monotonic decreasing	Single-peaked	Other	Total frequency
Possible	6	3	22	9	40
Probable/likely	17	2	20	1	40
Improbable/unlikely	0	16	21	3	40
Rather (prob./likely)	10	0	6	4	20
Rather (improb./unlikely)	0	7	12	1	20
Very (prob./likely)	17	0	1	2	20
Very (improb./unlikely)	0	19	1	0	20

that the probability is 0.5 that a correlation between DE and RO exceeds in magnitude the corresponding correlation between PC and RO. In 15 of the 20 cases the correlation between PC and RO exceeded the corresponding correlation between DE and RO, leading us to reject the null hypothesis ($p = 0.02$).

A second comparison of the PC and DE procedures consisted of computing the rank order

Table 4. Comparison of membership functions elicited by the DE and PC procedures

	Probability phrase					
	Possible	*Probable/ improbable*	*Likely/ unlikely*	*Rather* (·)	*Very* (·)	Total
DE dominates	4	12	12	12	12	52
PC dominates	5	2	0	2	2	11
Other	11	6	8	6	6	37

correlation between the functions elicited by the DE and PC procedures in Session 3 (as shown in Figs 1–5) for each subject and phrase separately. The mean rank order correlation is 0.85. As each individual correlation was based on six pairs of observations only, the criterion for rejecting the null hypothesis of zero correlation (at $p = 0.05$) in each instance was set at 0.829. In 33 of the 100 cases the rank order correlation did not exceed this very high criterion. For each of these 33 cases, a comparison was made between the rank order correlation between DE and RO and the rank order correlation between PC and RO. Of these 33 comparisons, the rank order correlation between PC and RO exceeded that between DE and RO in 24 cases, the reverse occurred in 7 cases and the rank order correlations were identical in 2 cases. The null hypothesis of equal proportions of cases where one correlation exceeds the other was again rejected by a sign test ($p < 0.01$). Although the RO procedure is not posited as the sole criterion for validating the ordinal properties of membership functions, the results of the last two tests provide convergent evidence which points to the superiority of the PC procedure over the DE procedure.

Effects of very and rather

A basic problem in quantitative fuzzy semantics is to devise an algorithm for the computation of the meaning of a composite term x from the knowledge of the meaning of each of its atomic components. Zadeh [29] addressed a special case of this problem in which the composite term x is of the form $x = mu$, where m is a modifier (e.g. m = *very*, *rather*, *highly*) and u is a primary term (e.g. *likely*, *intelligent*). The fundamental idea is that there is a small number of basic functions that, in combination, produce a wide range of modifiers for fuzzy predicates [30]. For example, *very* typically intensifies the particular predicate it modifies. Thus, any attempt to model the effect of *very* should decrease the degree of membership of those elements in the fuzzy set that represent the modified term whose degree of membership is < 1 in the fuzzy set represented by the unmodified term. A reasonable implementation of the modifier *very* is based on the unary operator called *concentration* (CON). If the result of applying a concentrator to A is denoted by CON(A), then the relationship between the membership functions of A and CON(A) may be given by

$$\mu_{\mathrm{CON}(A)}(x) = \mu_A^r(x), \qquad r > 1, \quad x \in U. \tag{3}$$

Although Zadeh [29] proposed $r = 2$ for *very* (see also, Schmucker [31]), he emphasized that his proposed representation was intended mainly to illustrate his approach toward modeling hedges rather than to provide accurate definitions. The exact form of the function and the values assumed by its parameters can only be determined empirically.

Another unary operation in fuzzy set theory, which also has no counterpart in ordinary set theory, is that of *contrast intensification* or, simply, *intensification* (INT). The effect of INT is to increase values of $\mu_A(x)$ that are above 0.5 and to diminish those that are below this threshold. It may be said, then, that INT heightens the contrast between the elements that are more than half in the set and those that are less than half in it [31]. Zadeh [29] proposed a simple concrete expression for an operation of this type, which we generalize:

$$\mu_{\mathrm{INT}}(x) = \begin{cases} 2r^{-1}\mu_A^r(x), & \text{for } 0 \leqslant \mu_A(x) \leqslant 0.5, \\ 1 - 2^{r-1}[1 - \mu_A(x)]^r, & \text{for } 0.5 \leqslant \mu_A(x) \leqslant 1, \end{cases} \tag{4}$$

where $r > 1$ and $x \in U$. Although Zadeh restricted himself unnecessarily to $r = 2$, other values of r may achieve the same desired effect on $\mu_A(x)$.

The membership functions of the four primary probability phrases that were modified by *very* are shown in Fig. 2 (*probable* and *improbable*) and Fig. 3 (*unlikely* and *likely*). Because of the slight superiority of the PC procedure over the DE procedure, we restrict our attention in the present section to PC-elicited membership functions only. Of the 20 PC-elicited functions, 10 are single-peaked and 10 are monotonic. Inspection of Fig. 5, however, shows that for the terms modified by *very* only 2 functions (subjects 15 and 18) are single-peaked; of the remaining 18 functions, 17 are monotonic and 1 (subject 13) is classified as "other". Although subjects 15 and 18 each yielded single-peaked functions for both the primary and the modified phrase, in neither case does *very* actually intensify the primary phrase it modifies. Similarly, when the primary phrase has a single-peaked function and the modified phrase has a monotonic function, *very* cannot be modeled as a concentrator.

We turn next to test Zadeh's concentration model, equation (3), when the membership functions of both the primary and modified phrases are monotonic. To do so, we first refer to Fig. 6 to illustrate how the CON operator works in this case. Figure 6 consists of three sections each divided into two parts. The solid lines in the upper three parts describe three hypothetical membership functions, which are not unlike those displayed in Figs 2 and 3. All three functions are monotonic increasing and share the same support. The function on the left is convex ($y = ax^4 + b$), the one in the middle is linear ($y = ax + b$) and the one on the right is concave ($y = a \log x + b$). The dashed lines represent the membership functions of the operator CON with $r = 2$. In each case, CON diminishes the degree of membership of those elements whose degree of membership is < 1. Similar functions with other values of r can be readily envisioned.

The three graphs in the lower part of Fig. 6 display the *difference functions* $\mu_A(x) - \mu_{\mathrm{CON}(A)}(x)$. The three difference functions are all single-peaked; they assume zero values at the two extremes of the probability range (0.50 and 0.95 in the present example) and positive values elsewhere. For monotonically increasing functions, the difference function is negatively skewed if $\mu_A(x)$ is convex,

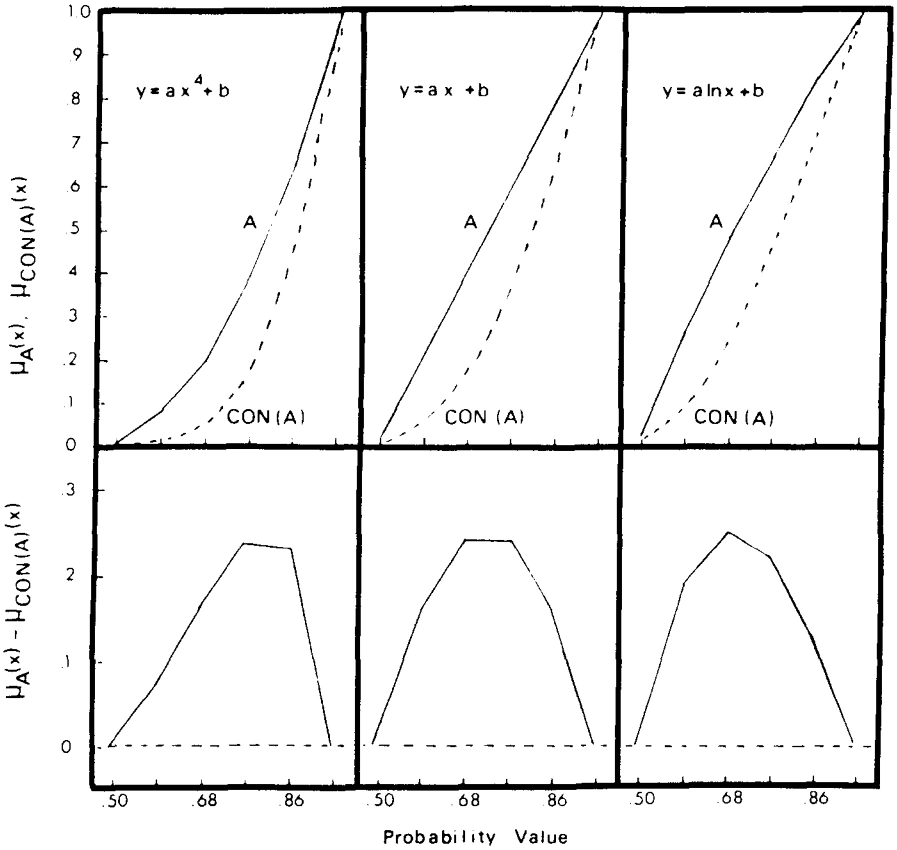

Fig. 6. Illustration of the CON operator (top) and of difference functions $\mu(x) - \mu_{\mathrm{CON}}(x)$ (bottom).

symmetric if $\mu_A(x)$ is linear and positively skewed if $\mu_A(x)$ is concave. For monotonically decreasing functions, positively and negatively skewed functions are interchanged. If *very* operates as a concentrator on monotonic functions, then the observed difference functions should have similar shapes to those displayed in Fig. 6. A very similar approach has been developed by Yager [32].

Experimental evidence concerning the effects of *very* has not supported Zadeh's conjecture. Hersh and Caramazza [9] and MacVicar-Whelan [12] suggested instead that the function shifts by a fixed constant (see also, Lakoff [30]). In other words, rather than modeling *very* by equation (3), they proposed to model it by $\mu_A(x + c)$, where c is positive if the membership function of the primary predicate is monotonic increasing and negative if it is monotonic decreasing. It follows from the latter model that the difference function should have a zero slope for all monotonic functions, a positive intercept for monotonic increasing functions and a negative intercept for monotonic functions.

Empirical functions, $\mu_W(x) - \mu_{very\,W}(x)$, were computed separately for each of the 10 subjects who exhibited monotonic membership functions for *probable*, *improbable*, *unlikely* and *likely*. Figure 7 displays these 10 difference functions. Because the membership functions are normalized to the same maximum, the right-hand ordinates of the top five figures and the left-hand ordinates of the bottom five figures must in general be equal to zero. The other extreme points are not so constrained. Seven of the 10 difference functions (subjects 1, 5, 7, 12, 14, 16 and 20) in Fig. 7 are seen to be in good agreement with Zadeh's [29] concentration model. Three other functions (subjects 3, 10 and 17) support neither of the two competing models.

The membership functions of the four probability phrases modified by *rather* are shown in Fig. 2 (*probable* and *improbable*) and Fig. 3 (*unlikely* and *likely*). Of the 20 PC-elicited functions for the primary phrases, 10 were classified as single-peaked, 9 as monotonic and 1 (subject 6 in Fig. 2) as "other". Empirical difference functions, $\mu_W(x) - \mu_{rather\,W}(x)$ were computed for all 19 subjects with single-peaked or monotonic functions. Figure 8 shows the difference functions for "single-peaked" subjects and Fig. 9 shows those for "monotonic" subjects.

Two models were tested for the effect of *rather* on probability phrases. The first model contends

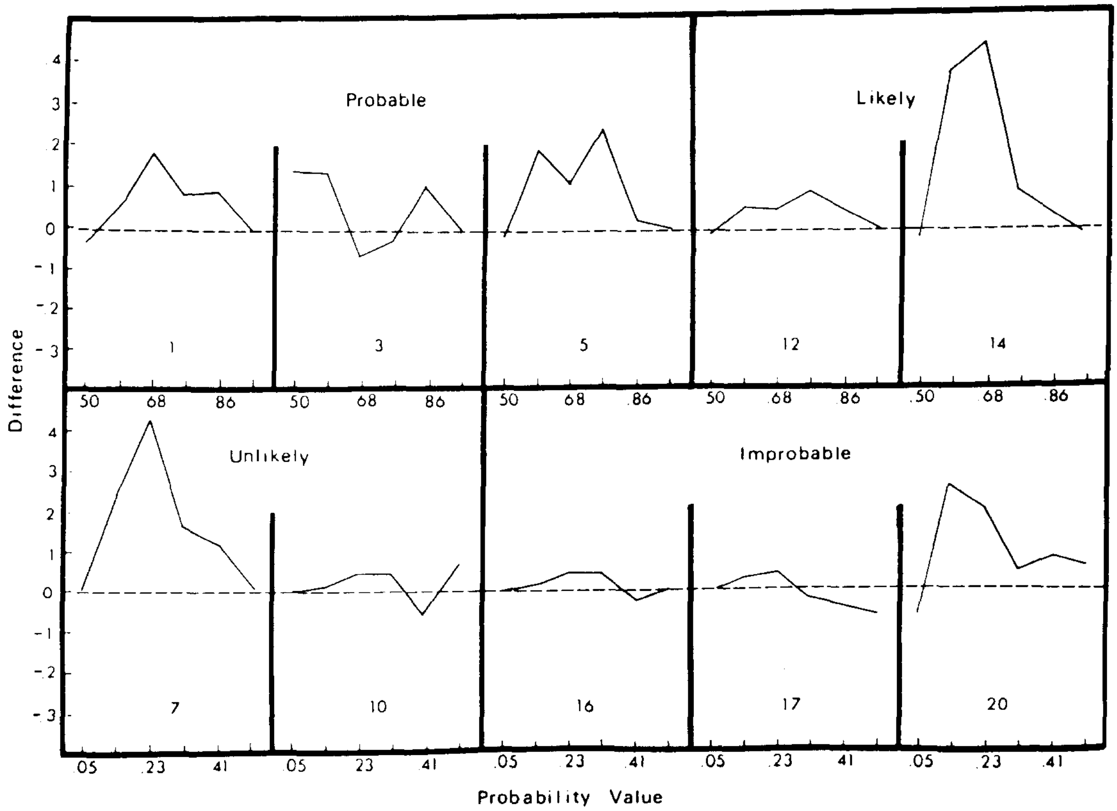

Fig. 7. Empirical difference functions, $\mu_W(x) - \mu_{very\,W}(x)$, for the 10 subjects with monotonic $\mu_W(x)$.

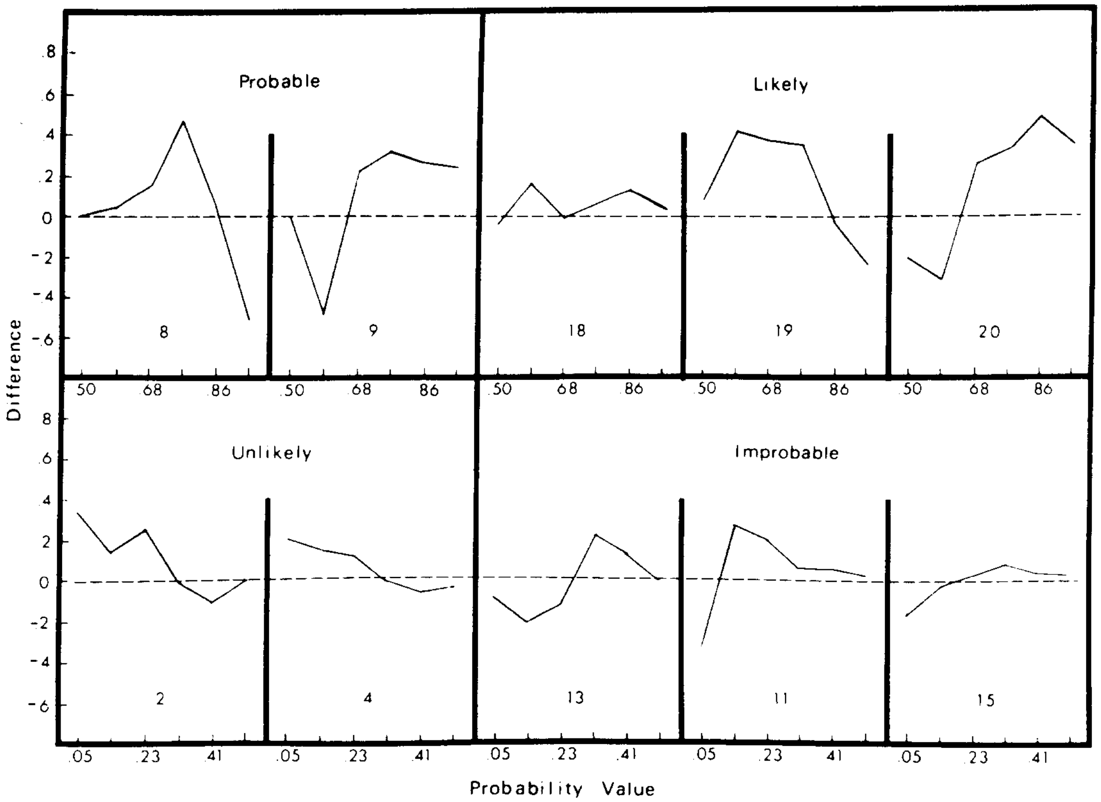

Fig. 8. Empirical difference functions $\mu_W(x) - \mu_{rather\,W}(x)$, for the 10 subjects with single-peaked $\mu_W(x)$.

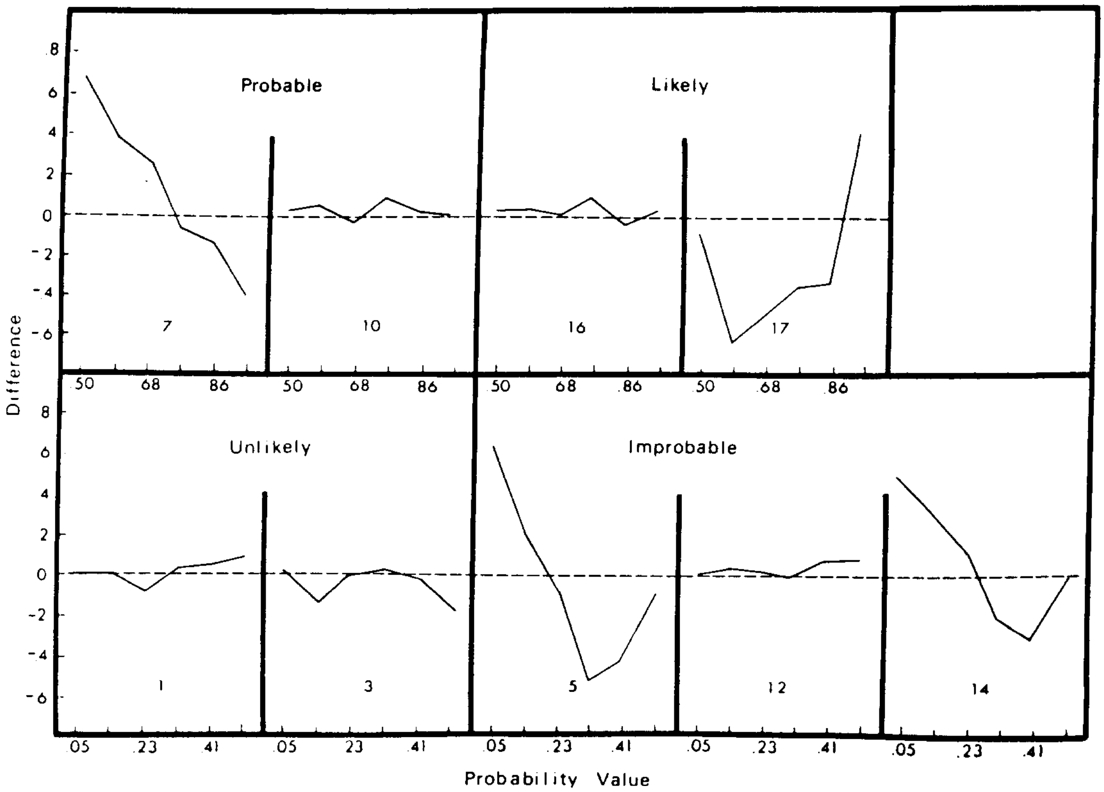

Fig. 9. Empirical difference functions, $\mu_w(x) - \mu_{rather\,W}(x)$, for the 9 subjects with monotonic $\mu_W(x)$.

that *rather* increases the fuzziness of the elements of the fuzzy set that represents the modified term. This effect may be modeled by another unary operation called *dilation* (DIL), which has the opposite effect of concentration [29]. The relationship between the membership functions of A and DIL(A) is given again by equation (3) with $r < 1$ (Zadeh proposed $r = 1/2$). The second model that we tested is due to Lakoff [30], who suggested that the effect of *rather* be modeled by the compound operation INT(CON(x)), where CON is defined by equation (3) and INT by equation (4).

We computed and studied the difference functions for DIL(x), $\mu_W(x) - \mu_{\mathrm{DIL}(W)}(x)$, and for INT(CON($x$)), $\mu_W(x) - \mu_{\mathrm{INT(CON}(W))}(x)$, using several hypothetical single-peaked and monotonic functions. Several tests that we performed on the observed difference functions, the details of which are omitted here, failed to corroborate either of the two models. Inspection of Figs 8 and 9 does not reveal any consistent patterns, suggesting, perhaps, that the use of *rather* to modify probability phrases is highly idiosyncratic.

DISCUSSION

The organization of this section is as follows. We first compare the present results to those obtained by Wallsten *et al.* [4]. This is followed by a comparison of the PC and DE procedures. Next, the effects of the modifiers are considered. Finally, we return to the issue of scale type.

Comparison with previous results

Wallsten *et al.* [4] employed the graded PC procedure in two experiments to establish membership functions for probability phrases similar to those used here. Judgments were highly stable over two sessions and the additive-difference axioms were satisfied to a very high degree. Furthermore, the derived membership functions representing individuals' understandings of a given phrase, varied greatly from one person to another, but were roughly constant over time for each person. Finally, the shapes of the functions were generally semantically reasonable, in that they were primarily single-peaked or monotonic, with single-peaked functions predominating for phrases towards the center of the [0, 1] interval, monotonically decreasing functions predominating near the low end and monotonically increasing functions predominating near the high end of the interval. Because the additive-difference axioms were so well satisfied in the previous study, we did not check them with the PC judgments in the present experiment. However, in all other respects the present results duplicate those of the prior study.

There is one illuminating difference between the present and preceding PC results. As already indicated, one cannot distinguish on ordinal grounds whether subjects are making ratio or difference judgments. In the preceding study it was also not possible to distinguish the two kinds of judgments on numerical grounds, because the two types of scaling models tended to describe the judgments equally well. This was not the case with the present data, however. In the previous experiments, the probability values judged for each phrase were quite closely spaced, whereas in this study they were separated by 0.09. As a result, there was more opportunity for the ratio and difference models to differentially fit the data, and they did. Uniformly, the difference model out-performed the ratio model, suggesting that in the PC situation subjects were judging differences rather than ratios. The same conclusion has been reached in other measurement contexts on the basis of other experimental manipulations [27].

The importance of the extensive individual differences in the present and prior results should not be minimized. It is clear that averaging over subjects would have been meaningless. Furthermore, these results should be troublesome to designers of expert systems who wish to incorporate membership functions representing the "meanings" of particular fuzzy terms.

Comparing the PC and DE procedures

On theoretical grounds alone, the PC procedure is more justifiable than the DE procedure. First, the PC procedure requires comparative rather than absolute judgments, which generally are easier for people to make. Second, the nature of the paradigm imposes sufficient structure on the data that it is possible to reject a set of judgments as being inconsistent with the underlying model. Additional forms of validation such as consideration of the shapes of the membership functions, or use of the membership functions to predict independent sets of judgments, are available, of course, for both procedures. Conversely, on pragmatic grounds the DE procedure is preferable to the PC procedure because it requires so many fewer judgments.

As pointed out earlier, a strong form of joint validation occurs when the two procedures give rise to the same membership functions. Ideally, one would like the resulting functions to be independent of the method of measurement, and when this occurs confidence is increased that the two methods are representing the same underlying construct.

In fact, although the two methods yielded similar functions in virtually all cases, they rarely yielded identical functions. The functions did not differ systematically in the case of *possible*, but for the other phrases the DE functions generally dominated the PC ones. In other words, the DE functions suggested that the phrases were more fuzzy than did the PC functions.

Generally when the PC and DE functions differed, the PC function agreed more strongly with the RO results than did the DE function. Because it involves only comparison judgments without quantitative evaluations, it is reasonable to think that the RO procedure is the simplest of the three, and therefore that its results come closest to representing the underlying ordinal characteristics of the meanings of the phrases. To the degree that this assumption about the RO characteristics is accepted, it can be said that the PC procedure yields more veridical representations than does the DE procedure.

These results can be considered from both theoretical and practical perspectives. Considering the theoretical issues first, Wallsten [33] pointed out that unlike measurement in classical physics, the quantities being measured are not independent of the instruments doing the measuring. Thus, for example, one may compare weights on a pan balance, secure in the knowledge that (unlike in quantum physics) the act of measuring does not affect the weights. However, when two degrees of membership are compared, those degrees are internal to the same organism doing the comparison, and therefore one cannot consider the quantities being measured independently of the method of measurement. Thus, one possibility is that the meaning of a vague concept in a particular situation to a particular individual actually changes with the method of interrogation. A second interpretation is that the meaning stays relatively fixed over measurement procedures, but some procedures lead to less distortion than others. According to this second view, comparative non-quantitative judgments are the simplest to make, and therefore the most accurate (although frequently the least informative), while quantitative judgments involve comparison against an internal number scale which introduces additional measurement error and therefore judgments that are somewhat more similar to each other. These two views cannot be distinguished on the basis of the present data, but the latter seems more sensible to us.

On practical grounds, the outcomes of the two procedures are not so different. Thus, the less taxing DE method may frequently yield measures that are sufficiently good for a particular purpose.

Effects of modifiers

It is frequently difficult to compare which of two or more functions fit erroneous data, particularly when the functions include free parameters. The comparison is easiest when it can be done on the basis of qualitative rather than statistical characteristics of the data. This is precisely what we did in testing models for the effects of *very* and *rather* (cf. Figs 6–9).

Considering *rather* first, the two models tested above are both inconsistent with the responses of our subjects. Furthermore, as the difference functions for *rather* show, the modifier has no consistent effect over subjects. It is possible that the term is completely empty and provides no modifying effect at all, but it is rather more likely that the term has differential effects for different people, depending on their linguistic background. Thus, for some people *rather* might serve as a hedge, i.e. move the meaning of the phrase more toward the center of the [0, 1] interval, while for other people it serves as an intensifier, and for a third group of people it has, in fact, no real meaning. If this is the case, then the modifier *rather* should be eschewed in the development of systems for specific applied purposes.

The effect of *very*, however, is clear. Not surprisingly, *very* acts as an intensifier. It is particularly interesting that 18 of the 20 membership functions for phrases including *very* were monotonic, despite the fact that 10 of the functions for the base term were single-peaked. Thus, *very* not only intensifies, but it also eliminates any hedging quality to the meaning of the phrase (Wallsten *et al.* [4], distinguished between concepts with hedged and unhedged meanings in that the former had membership values of zero at both the upper and lower boundaries of its domain of support, while the latter had membership function values of zero only at one or the other boundary).

No available descriptive model for the effects of *very* can handle this full range of results. However, when attention is restricted only to those phrases for which the membership function of the unmodified term was itself monotonic, i.e. the term already had an unhedged meaning to the subject, then Zadeh's [29] concentration model worked well in 7 of the 10 cases. There was no indication whatsoever that *very* was better represented by a shift than by a concentration function.

Scale of measurement

As already indicated, we can consider the minimum scale of measurement guaranteed by our procedures, and then whether assumptions are justified or required to yield a yet stronger measurement scale. The PC procedure, based on the additive-difference axioms, guarantees interval scale measurement. Because of the lack of internal structure in the DE procedure, its scale properties are not so easily derivable. In order to claim ratio scale measurement we would need empirical grounds for establishing membership functions that are truly zero. While such procedures may be possible in a different study, they were not invoked here, and therefore we have no basis for such a claim. However, our comparisons of the PC and DE procedures, except for the correlations, as well as our examination of the effects of modifiers, implicitly involved an additional assumption regarding the scale properties. Namely, for each subject we assumed that both the DE and the PC functions, as well as functions for different phrases, were all unique up to common linear transformations. Although this assumption remains untestable, it yields a pattern of results that are sensible and interpretable, and on these pragmatic grounds the assumption may be considered reasonable.

Acknowledgement—This research was supported by Contract MDA903-83-K-0347 from the U.S. Army Research Institute.

REFERENCES

1. A. M. Norwich and I. B. Turksen, Meaningfulness in fuzzy set theory. In *Fuzzy Set and Possibility Theory* (Edited by R. R. Yager), pp. 68–74, Pergamon Press, New York (1982).
2. A. M. Norwich and I. B. Turksen, The fundamental measurement of fuzziness. *Ibid.*, pp. 49–60.
3. A. M. Norwich and I. B. Turksen, The construction of membership functions. *Ibid.*, pp. 61–67.
4. T. S. Wallsten, D. V. Budescu, A. Rapoport, R. Zwick and B. Forsyth, Measuring the vague meanings of probability terms. *J. exp. Psychol.: Gen.* **115**, 348–365 (1986).
5. A. M. Norwich and I. B. Turksen, A model for the measurement of membership and consequences of its empirical implementation. *Fuzzy Sets Syst.* **12**, 1–25 (1984).
6. J. A. Gougen, The logic of inexact concepts. *Synthese* **19**, 325–373 (1969).
7. T. L. Saaty, Measuring the fuzziness of sets. *J. Cybern.* **4**, 43–61 (1974).
8. U. Thole, H.-J. Zimmerman and P. Zysno, On the suitability of minimum and product operators for the intersection of fuzzy sets. *Fuzzy Sets Syst.* **2**, 167–180 (1979).
9. H. M. Hersh and A. Caramazza, A fuzzy set approach to modifiers and vagueness in natural language. *J. exp. Psychol.: Gen.* **105**, 254–276 (1976).
10. H. M. Hersh, A. Caramazza and H. H. Brownell, Effects of context on fuzzy membership functions. In *Advances in Fuzzy Set Theory and Applications* (Edited by M. M. Gupta, R. K. Ragade and R. R. Yager). North-Holland, Amsterdam (1979).
11. V. B. Kuz'min, A parametric approach to description of linguistic values of variables and hedges. *Fuzzy Sets Syst.* **6**, 27–41 (1981).
12. P. J. MacVicar-Whelan, Fuzzy sets, the concept of height, and the hedge *very*. *IEEE Trans. Syst. Man Cybern.* **SMC-8**, 507–511 (1978).
13. G. C. Oden, Integration of fuzzy logical information. *J. exp. Psychol.: Hum. Percept. Perform.* **3**, 565–575 (1977).
14. G. C. Oden, Fuzziness in semantic memory: Choosing exemplars of subjective categories. *Memory Cogn* **5**, 198–204 (1977).
15. P. Zysno, Modeling membership functions. In *Empirical Semantics* (Edited by B. B. Rieger). Brockmeyer, Bochum (1981).
16. D. M. Green and J. A. Swets, *Signal Detection Theory and Psychophysics*. Wiley, New York (1966).
17. G. T. Fechner, *Elemente der Psychophysik*. Breitkopf & Hartel, Leipzig (1860). [Translation of Vol. I reprinted as *Elements of Psychophysics*. Holt, Rinehart & Winston, New York (1966).
18. L. L. Thurstone, The law of comparative judgment. *Psychol. Rev.* **34**, 273–286 (1927).
19. R. N. Shepard, Psychological relations and psychophysical scales: on the status of "direct" psychophysical measurement. *J. math. Psychol.* **24**, 21–57 (1981).
20. D. H. Krantz, R. D. Luce, P. Suppes and A. Tversky, *Foundations of Measurement*, Vol. I. Academic Press, New York (1971).
21. P. De Jong, A statistical approach to Saaty's scaling method for priorities. *J. math. Psychol.* **28**, 467–478 (1984).
22. R. E. Jensen, An alternative scaling method for priorities in hierarchical structures. *J. math. Psychol.* **19**, 61–64 (1984).

23. T. L. Saaty, A scaling method for priorities in hierarchical structures. *J. math. Psychol.* **15**, 234–281 (1977).
24. T. L. Saaty, *The Analytic Hierarchy Process.* McGraw-Hill, New York (1980).
25. T. L. Saaty and L. G. Vargas, Inconsistency and rank preservation. *J. math. Psychol.* **28**, 205–214 (1984).
26. G. Crawford and C. Williams, A note on the analysis of subjective judgment matrices. *J. math. Psychol.* **29**, 387–405 (1985).
27. M. H. Birnbaum, Comparison of two theories of "ratio" and "difference" judgements. *J. exp. Psychol.: Gen.* **109**, 304–319 (1980).
28. J. M. Miyamoto, An axiomatization of the ratio/difference representation. *J. math. Psychol.* **27**, 439–455 (1983).
29. L. A. Zadeh, A fuzzy-set-theoretic interpretation of linguistic hedges. *J. Cybern.* **2**, 4–34 (1972).
30. G. Lakoff, Hedges: a study in meaning criteria and the logic of fuzzy concepts. *J. phil. Logic* **2**, 458–508 (1973).
31. K. J. Schmucker, *Fuzzy Sets, Natural Language Computations, and Risk Analysis.* Computer Science Press, Rockville, Md (1984).
32. R. R. Yager, Linguistic hedges: the relation to context and their experimental realization. *Cybern. Syst.* **13**, 357–374 (1982).
33. T. S. Wallsten, The psychological concept of subjective probability: a measurement theoretic view. In *The Concept of Probability in Psychological Experiments* (Edited by C.-A. S. Staël von Holstein), pp. 49–72. Reidel, Dordrecht, The Netherlands (1974).