*Int. J. Man–Machine Studies* (1989) **30**, 69–111

# Combining stochastic uncertainty and linguistic inexactness: theory and experimental evaluation of four fuzzy probability models

RAMI ZWICK

*Department of Marketing, 701 Business Administration Building, The Pennsylvania State University, University Park, PA 16802, USA*

THOMAS S. WALLSTEN

*Psychology Department, University of North Carolina, Chapel Hill, NC 27599, USA*

*(Received 22 January 1988)*

Two major sources of imprecision in human knowledge, linguistic inexactness and stochastic uncertainty, are identified in this study. It is argued that since in most realistic situations these two types exist simultaneously, it is necessary to combine them in a formal framework to yield realistic solutions. This study presents such a framework by combining concepts from probability and fuzzy set theories. In this framework four models (Kwakernaak, 1978; Yager, 1979; 1984*b*; Zadeh, 1968; 1975) that attempt to account for the numeric or linguistic responses in various probability elicitation tasks were tested. The linguistic models were relatively effective in predicting subjects' responses compared to a random choice model. The numeric model (Zadeh, 1968) proved to be insufficient. These results and others suggest that subjects are unable to represent the full complexity of a problem. Instead they adopt a simplified view of the problem by representing vague linguistic concepts by multiple-crisp representations (the $\alpha$-level sets). All of the mental computation is done at these surrogate levels.

## 1. Introduction

### 1.1. BACKGROUND

Most real world decision-making takes place in an environment in which the states of nature, feasible actions and outcomes, and available information are only imprecisely known. Imprecision has been quantified primarily by means of probability theory, a practice which tacitly implies that imprecision of any sort can be equated to randomness. Several authors (Bellman & Zadeh, 1970; Dutta, 1985) have emphasized the need for differentiating among the sources of imprecision underlying particular assumptions or items of evidence. Imprecision can arise from a variety of sources (Dutta, 1985): incomplete knowledge, inexact language, ambiguous definitions, inherent stochastic characteristics, measurement problems, etc. This work deals with two major sources of imprecision in human knowledge: *stochastic uncertainty*, and *linguistic inexactness*. Specifically, it investigates the combined effects of stochastic uncertainty and linguistic inexactness in a probability encoding task.

The plan for the rest of the Introduction section is as follows. In the remainder of this section we briefly discuss stochastic uncertainty and linguistic inexactness. The



0020-7373/89/010001 + 22$03.00/0 © 1989 Academic Press Limited

next section discusses relationships between the two concepts. In it we emphasize the inability of classical probability theory to cope with the multiplicity of kinds of imprecision that one encounters in decision analysis, and the need to independently account for the vague meanings of linguistics terms. Section 1.3 discusses the widespread use of linguistic terms in decision analysis and the apparent advantage of *thinking* about uncertainty in a linguistic rather than numeric mode. We conclude that there is a need for a general framework for combining stochastic uncertainty and linguistic inexactness to deal with situations in which both sources of imprecision exist. Section 1.4 presents such a framework. Section 1.5 reviews four models based on fuzzy set theory for representing the probability of a fuzzy or non-fuzzy event in a fuzzy or non-fuzzy environment, and locates these models in the framework developed in Section 1.4. Finally, Section 1.6 presents the purpose of the experimental work.

At least three levels of subjective stochastic uncertainty can be differentiated. On the first level it is assumed that beliefs can be represented by a single probability measure defined over the states of nature. This assumption is very strong since it amounts to the agent having complete information in the sense that he or she is certain of the probabilities of the possible states of nature. On the second level it is assumed that the beliefs can be represented by a second order probability distribution over probability values. The third level refers to the case where none of the above assumptions can be made. Whereas these levels have received little attention among subjective statisticians (de Finetti, 1977), their distinction for understanding the psychology of choice and inference behaviour has been demonstrated experimentally (Becker & Brownson, 1964; Ellsberg, 1961; Yates & Zukowski, 1976).

The second source of imprecision is linguistic inexactness. Black (1937) distinguished three kinds of inexactness in natural language. The first is *generality*, in which a word applies to a multiplicity of objects in the field of reference. For example, the word *chair* can apply *perfectly well* to objects differing in size, shape and material. The second kind of linguistic inexactness is *ambiguity*, which occurs when a finite number of alternative meanings have the same phonetic form (e.g., *bank*), and the third is *vagueness*, in which there are no precise boundaries to the meaning of a word (e.g., *young, rich*).

In general, a proposition is *uncertain* if it involves a stochastic process; a proposition, whose contents state the value of some variable, is linguistic inexact if this value is not sufficiently determined with respect to a given scale (Dutta, 1985). Note that an exact proposition may be uncertain ("it will be 4°C tomorrow"), and a proposition which is completely certain may be linguistically inexact ("it is *warm* now").

### 1.2. THE RELATIONSHIP BETWEEN STOCHASTIC UNCERTAINTY AND LINGUISTIC INEXACTNESS

Central to the standard axiomatic model of subjective probability (Ramsey/de Finetti/Savage) is the idea that probability, which is a measure of one's degree of belief, can be operationalized via choices among gambles (Savage, 1954). This model encounters difficulties even in an idealized paradigm, such as Ellsberg's (1961) classical paradox of "ambiguous-probabilities," or Gardenfors & Sahlin's

(1982) example of "unreliable" probabilities. It is clear in these examples that the ordinary additive probability representation does not capture the entire psychological reality of uncertainty. Demonstrations such as those by Ellsberg (1961) and Gardenfors & Sahlin (1982) suggest that the notion of probability refers in natural language to several distinct states of mind, to which the rules of the standard calculus of probability may not be equally applicable (Kahneman & Tversky, 1982).

Numerous attempts have been made to develop a more general model of subjective uncertainty (Budescu & Wallsten, 1987). Two principle alternatives that have been suggested within a probability framework are: (i) *second order* probabilities, described by a probability density function over probabilities; and (ii) *interval* or lower-and-upper probabilities, described by a rectangular indicator function. In both models the probability of an event is represented by some sort of *function* defined on the unit interval rather than a single point in the interval. In this work we are interested in cases where neither of the above alternatives are justifiable on the basis of the available information. Such is the case, for example, whenever the event itself is loosely defined, or whenever the available information with regard to the probabilities of possible states of nature is given in linguistic rather than numeric terms.

Consider for example the following problem, in which the emphasized words have vague meanings: An urn contains *approximately* 100 balls of various sizes, of which *several* are *large*. What is the probability that a ball drawn at random is *not large*? (Zadeh, 1984).

The main limitation of classical probability theory in coping with this problem is that it is based on two-valued logic. This means that all predicates and concepts in probability theory have crisp denotations, a restriction that rules out events defined by linguistic predicates like *warm*, *young*, *short*, and/or linguistic quantifiers like *most*, *several*, *few*, and/or linguistic probabilities like *probably*, *likely*, or *doubtful* (Zadeh, 1986). What is needed is a general computational system for representing the meaning of inexact propositions.

Within the theory of fuzzy sets the vague meaning of a linguistic term or phrase is represented by a *membership function* from the universe of discourse to the $[0, 1]$ interval. The function assigns the value zero to elements that are not in the concept represented by the phrase. The value one is assigned to elements that are definitely in the concept, and intermediate values are assigned to elements with intermediate degrees of membership in the concept represented by the phrase. If the linguistic term is well defined then the membership function can take on only 0 or 1. If the concept is not well defined (e.g., *young*), then the membership function can take on any value in the $[0, 1]$ interval. Note that all three types of linguistic inexactness distinguished by Black (1937) can be represented by the membership function; generality occurs when the portion of the universe of discourse where the membership value equals one is not just one point; ambiguity occurs when there is more than one local maximum of the membership function; and vagueness occurs when the function takes values other than just 0 and 1 (Goguen, 1968–69).

There are both fundamental differences and similarities between fuzzy set theory and probability theory, as well as relationships between the two (Gaines, 1978). In this work we advocate combining both fuzzy set and probability theories in order to deal with situations for which each theory alone is inadequate. Fuzzy set theory and

probability theory should be viewed not as rivals, but rather as similar logical systems, having a common core that is adequate for many aspects of decision analysis, and differing in certain well-defined features that may or may not be relevant in particular applications.

### 1.3. THE USE OF LINGUISTIC TERMS IN DECISION ANALYSIS

Forecasting is essential for decisions that involve possible future events. Sometimes the forecaster and the decision maker are the same person, but frequently, however, one person forecasts, while another makes the decisions. A necessary condition for good decision processes is good *communication* between these two persons (Beth-Marom, 1982).

Poor communication resulting from the use of ambiguous expressions can lead to a disaster. Behn & Vaupel (1982) offer the following example:

> In early 1961, President Kennedy ordered the Joint Chiefs of Staff to study the Central Intelligence Agency's plan for an invasion of Cuba by expatriates. The general in charge of the evaluation concluded that its chances of overall success were "fair," by which he meant that they were 30%. Yet, when the Joint Chiefs sent their report to the president, no probabilities were included; instead the report stated. "This plan has a fair chance of ultimate success," the rest is history. Years later the general felt that the misinterpretation of the word "fair" had been one of the central misunderstandings of the Bay of Pigs fiasco, and he was still unhappy with himself for not insisting that a specific, numerical assessment be used. Recalled the general, "We thought other people would think that "a fair chance" would mean "not too good."
>
> (Quoted by Peter Wyden, 1979, pp. 89–90.)

A second recent example is taken from the report of the Rogers Committee on the space shuttle disaster. This example demonstrates how widespread the practice of qualitatively estimating frequencies and probabilities is. This is how Milton Silveira, a NASA official, described the way NASA estimated the risk involved in launching the space shuttle:

> They get all the top engineers together down at Marshall Space Flight Center and ask them to give their best judgement of the reliability of all the components involved." The engineers' *adjectival descriptions* are then converted to numbers. For example, Silveira says, "frequent" equals 1 in 1000; "occasional" equals 1 in 10 000; and "remote" equals 1 in 100,000.
>
> When all the judgements were summed up and averaged, the risk of a shuttle booster explosion was found to be 1 in 100 000. That number was then handed over to DOE.
>
> (*Science*, July 1986)

These kinds of examples have led some researchers to conclude that we need an agreed-upon set of rules for translating linguistic probabilities into numbers (Kent, 1964), and others to conclude that forecasting organizations should change their policy and use numerical expressions of probability rather than verbal ones (Beyth-Marom, 1982; Bryant & Norman, 1980; Nakao & Axelrod, 1983). Behn & Vaupel (1982) arrive at the same conclusion. They write:

> "for decision makers, ambiguous probability statements are useless. They simply do not provide the information necessary to analyse a decision." Behn & Vaupel (1982), p. 78.

The emphasis on the exclusive use of the numerical representation of uncertainty even when the data are imprecise is based on the current inability to effectively *communicate* in linguistic terms, and ignores the following points:

(1) Linguistic communication is common practice. In addition to the above examples, Szolovits & Pauker (1978) found that while experts seem quite prepared to give qualitative estimates of likelihood, they often refuse to give precise numerical estimates of outcomes.

(2) It may be possible to measure the vague meanings of linguistic terms, and hence eliminate the communication objection. A method, based on fuzzy set concepts, to measure the vague meaning of linguistic terms has been developed by Wallsten and his associates (Wallsten, Budescu, Rapoport, Zwick & Forsyth, 1986*a*) and by others (Norwich & Turksen, 1984; Zimmermann, 1987; Zwick, 1987*a*; Zysno, 1981) and has been used successfully with linguistic probabilities. The meanings of linguistic terms such as *doubtful, probable* or *likely* are expressed as membership functions over the [0, 1] probability interval. Wallsten *et al.* (1986*a*) have demonstrated that the derived membership functions have interpretable shapes and, more importantly, predict the judgements well for each subject in an independent task. Subjects' membership functions were stable over time, although considerable between-subject variability was observed. The reader is referred to Wallsten *et al.* (1986*a*) and to Rapoport, Wallsten & Cox (1987) for a full discussion of this topic.

(3) Thinking about uncertainty may be facilitated by the use of linguistic terms. For example, Zimmer (1983) found that subjects are better able to consider complex dependencies in a problem if they are to make verbal judgements than if they are to provide numerical judgements. Kochen (1979), based on findings in perception and problem solving, concluded that excessive precision and (apparent) clarity may be as ineffective as excessive vagueness and confusion. The attempt to perceive a complex situation with an "inappropriately high" degree of precision may fail, with costly consequences. Fox, Barber & Bardhan (1980) have designed a rule-based system for the diagnosis of indigestion in which hypotheses formed from data were marked with strings like "possible", "maybe", etc. They claim that the imprecise approach was viable partially because of the ability to exploit patterning in the data which the Bayesian method could not. The exploitation of the qualitative features of the problem could therefore compensate for the loss of precision normally provided by numerical procedures. Nagy & Hoffman (1981) conducted a preliminary study of the performance difference between subjects using natural language estimates and those using numerical estimates in assessing the security risks of various computer installation configurations. Though the study used few subjects, it indicated that the use of natural language rather than numbers was associated with an increase of accuracy due to the elimination of extremely inaccurate estimates.

(4) For many decisions precise quantitation is unnecessary and may in practice be an illusion when the data are unreliable. Commenting that numbers denote authority and a precise understanding of relationships, a committee of the US

National Research council wrote 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).

All of the above points suggest that in order to take advantage of the possible superiority of thinking about uncertainty, in linguistic terms, in line with common practice, a general framework is needed that will combine stochastic uncertainty and linguistic inexactness. Such a framework is developed in the next section.

### 1.4. A FRAMEWORK FOR COMBINING STOCHASTIC UNCERTAINTY AND LINGUISTIC INEXACTNESS

Probability theory and the theory of fuzzy sets are both used to study imprecision. Probability theory deals with imprecision due to the occurrence of random events, while fuzzy set theory deals with imprecision inherent in the use of natural language. Consider once again the following problem in which the emphasized words have a vague meaning: An urn contains *approximately* 100 balls of various sizes, of which *several* are *large*. What is the probability that a ball drawn at random is *not large*? (Zadeh, 1984).

One way to deal with such a combination of randomness and natural language concepts, is to fuzzify the concept of a *random variable*. A random variable consists of two components, a real valued function from the sample space to the real line (variable), and a probability structure defined over the sample space (random). The concept of a random variable can be fuzzified through each of its components.

*Definition 1* (Zadeh, 1965). Let $U$ be the universe of discourse. A *fuzzy subset* $F$ of $U$ is characterized by a membership function $\mu_F : U \rightarrow [0, 1]$, which associates with each element $u$ of $U$ a number $\mu_F(u)$ representing the *grade of membership* of $u$ in $F$.

*Definition 2* (Dubois & Prade, 1982*a*). A *fuzzy function* $\tilde{f}$ from a set $U$ to a set $V$ is a function from $U$ to the set of nonempty fuzzy subsets of $V$, namely $\tilde{P}(V) - \{\phi\}$.

In other words, each element $u \in U$ corresponds to a fuzzy set $\tilde{f}(u)$ defined on $V$, whose membership function is $\mu_{\tilde{f}(u)}$, and $\tilde{f}(u)$ is nonempty. Other definitions of fuzzy functions exist in the literature (Chang & Zadeh, 1972; Negoita & Ralescu, 1975). In this study however, we will adopt the above definition. Intuitively, a fuzzy function is an ill-defined function in the sense that to a given precise element $u \in U$, there correspond several more or less possible images. For example, a function that assigns a linguistic age (represented by a fuzzy subset of the real line) to each member of a set of students is a fuzzy function.

Dubois & Prade (1978*a*; 1982*a,b,c*) have systematically studied the fundamental theory of fuzzy set-valued functions and its application.

*Definition 3* (Zadeh, 1975). Let $X$ be a variable, whose universe of discourse, $\Omega$, is a finite set. $\Omega = \{u_1, u_2, \ldots, u_n\}$. With each $u_i$, $i = 1, \ldots, n$, we associate a linguistic probability. The $n$-ary linguistic vector $(\tilde{P}_1, \ldots, \tilde{P}_n)$ constitutes a *linguistic probability assignment list*, and will be referred to as a *linguistic probability distribution* or a *fuzzy probability structure*.

For example, by assigning the phrase "very likely" to the event "rain tomorrow", and the phrase "very unlikely" to the event "no rain tomorrow", we define a fuzzy probability structure over the sample space $\Omega = \{\text{rain, no-rain}\}$.

Several authors (Adamo, 1980; Nguyen, 1977; Zadeh, 1975) have studied the properties of a fuzzy probability structure.

*Definition 4* (Zadeh, 1968). Let $\Omega$ be a sample space. A *fuzzy event* is a fuzzy subset of $\Omega$.

In what follows the term *crisp* will refer to the classical definitions of function, probability structure, and event.

Using the above definitions, we propose combining stochastic uncertainty and linguistic inexactness by defining four different kinds of random variables.

Let $\Omega$ be a sample space $\mathcal{R}$ the real line, and $\tilde{P}(\mathcal{R})$ the set of all fuzzy subsets of $\mathcal{R}$, then:

A *Crisp Random Crisp Variable* is a Crisp function, $X$,

$$X : \Omega \rightarrow \mathcal{R}$$

with a crisp probability structure defined on $\Omega$.

A *Crisp Randon Fuzzy Variable* is a fuzzy function, $\tilde{X}$,

$$\tilde{X} : \Omega \rightarrow \tilde{P}(\mathcal{R}).$$

with a crisp probability structure defined on $\Omega$.

A *Fuzzy Random Crisp Variable* is a crisp function, $X$,

$$X : \Omega \rightarrow \mathcal{R}$$

with a fuzzy probability structure defined on $\Omega$.

A *Fuzzy Random Fuzzy Variable* is a fuzzy function, $\tilde{X}$,

$$\tilde{X} : \Omega \rightarrow \tilde{P}(\mathcal{R})$$

with a fuzzy probability structure defined on $\Omega$.

For each of the four types of random variables one can ask: "What is the probability of an event," where the event can be either crisp or fuzzy. Furthermore, the question can be answered with either a crisp probability number or a fuzzy number expressed as a linguistic probability. Table 1 presents the 16 possible combinations of crispness and fuzziness pertaining to the probability structure over the space (random), the function into which the space is mapped (variable), the event, and the response. The rightmost column of Table 1 references the original work in each of these categories, some of which will be summarized in the next section. The many empty cells in the rightmost column emphasize that many of the combinations have not yet been explored. It is important to note that other definitions of random fuzzy variables (and their expectations) have been proposed

TABLE 1

*The 16 possible combinations of crispness and fuzziness pertaining to the probability structure (random), the function (variable), the event, and the response*

| Random | Variable | Event | Response | Original work |
|---|---|---|---|---|
| Crisp | Crisp | Crisp | Crisp | Classical probability theory |
| Crisp | Crisp | Crisp | Fuzzy | |
| Crisp | Crisp | Fuzzy | Crisp | Zadeh (1968) |
| Crisp | Crisp | Fuzzy | Fuzzy | Yager (1979; 1984*b*) |
| Crisp | Fuzzy | Crisp | Crisp | |
| Crisp | Fuzzy | Crisp | Fuzzy | Kwakernaak (1978; 1979) |
| Crisp | Fuzzy | Fuzzy | Crisp | |
| Crisp | Fuzzy | Fuzzy | Fuzzy | |
| Fuzzy | Crisp | Crisp | Crisp | |
| Fuzzy | Crisp | Crisp | Fuzzy | Zadeh (1975) |
| Fuzzy | Crisp | Fuzzy | Crisp | |
| Fuzzy | Crisp | Fuzzy | Fuzzy | Zadeh (1975) |
| Fuzzy | Fuzzy | Crisp | Crisp | |
| Fuzzy | Fuzzy | Crisp | Fuzzy | |
| Fuzzy | Fuzzy | Fuzzy | Crisp | |
| Fuzzy | Fuzzy | Fuzzy | Fuzzy | |

by Puri & Ralescu (1988), Nahmias (1978; 1979), Nguyen (1977), and Stein & Talati (1981).

Ideally, we would like to have one general model that accounts for human behaviour in all 16 cases. Such a model might treat all cells as special cases of the last row of Table 1, in which a fuzzy (linguistic) response is given to a query regarding the probability of a fuzzy event, given a fuzzy random fuzzy variable. However, from a psychological point of view it is not clear whether the same underlying judgment process operates in each of the 16 cells. It might be the case that individuals cope with combinations of randomness and liguistic inexactness in qualitatively different ways, depending on the exact combination and location of the linguistic inexactness.

The following section reviews four models developed in the fuzzy set literature. Each one pertains to a different row in the taxonomy of Table 1, and each was tested in the experiment to be described subsequently. The development of a unified model must await further research.

### 1.5. MODELS BASED ON FUZZY SET THEORY FOR REPRESENTING THE PROBABILITY OF A (FUZZY) EVENT GIVEN A (FUZZY) RANDOM (FUZZY) VARIABLE

One model, due to Zadeh (1968), pertains to the probability of a fuzzy event given a crisp random crisp variable. This is the only model tested in this work that assumes a numeric response. Yager's model (1979; 1984*b*) pertains to the same background information as Zadeh's, but assume a linguistic response. A model developed by Kwakernaak (1978; 1979) pertains to the probability of a crisp event given a crisp random fuzzy variable and assumes a linguistic response. Finally, another model developed by Zadeh (1975) pertains to the probability of a crisp event given a fuzzy random crisp variable, and assumes a linguistic response.

*1.5.1. The real-valued probability of a fuzzy event given a crisp random crisp variable (Zadeh, 1968)*

Consider the data in Table 2. What is the probability that a randomly selected person from this group will be *young*? This is a query that demands a numerical response about a fuzzy event given a crisp random crisp variable.

Let $(\mathcal{R}^n, \sigma, P)$ be a crisp probability space in which $\sigma$ is the $\sigma$-field of Borel sets in $\mathcal{R}^n$ and $P$ is a probability measure over $\mathcal{R}^n$. The probability of a fuzzy event $A$ (a fuzzy subset of $\mathcal{R}^n$ whose membership function is Borel measurable) is defined by Zadeh to be the Lebesgue–Stieltjes integral

$$P(A) = \int_{\mathcal{R}^n} \mu_A(x)\, \mathrm{d}P = E[\mu_A(x)],$$

where $\mu_A$ is the membership function of the fuzzy event $A$. If $A$ is a finite set, then $P(A)$ is simply the weighted sum of the membership values of the elements in $A$, with each value weighted by its respective probability. In more general terms, $P(A)$ is the expected value of the membership function of $A$. This is a straight forward generalization of the classical case where the probability of $A$ equals the expected value of the characteristic function of $A$.

*Example 1.* Let the concept *young* be represented by the following membership function:

$$\mu_{\text{YOUNG}}(u) = \begin{cases} 1 & \text{if} \quad u \le 25 \\ (1 + ((u-25)/5)^2)^{-1} & \text{if} \quad 25 < u < 100 \\ 0 & \text{if} \quad u \ge 100 \end{cases}$$

On the basis of the above function, the membership values of the numerical ages are presented in Table 2.

Using Zadeh's definition, the probability that a randomly selected person from

TABLE 2

*Probability distribution of numerical ages in a certain group and the membership values of the numerical ages in the fuzzy set YOUNG (see Example 1)*

| Probability | Age | $\mu_{\text{YOUNG}}$(age) |
|---|---|---|
| 0·300 | 50 | 0·038 |
| 0·100 | 43 | 0·072 |
| 0·200 | 37 | 0·148 |
| 0·200 | 30 | 0·500 |
| 0·050 | 28 | 0·735 |
| 0·025 | 20 | 1·000 |
| 0·025 | 19 | 1·000 |
| 0·100 | 17 | 1·000 |

Table 2 will be *young* is 0·335, calculated as:

$$(0{\cdot}3 \times 0{\cdot}038) + (0{\cdot}1 \times 0{\cdot}072) + (0{\cdot}2 \times 0{\cdot}148) + (0{\cdot}2 \times 0{\cdot}5) + (0{\cdot}75 \times 0{\cdot}735) + (0{\cdot}025 \times 1) + (0{\cdot}025 \times 1) + (0{\cdot}1 \times 1) = 0{\cdot}335,$$

where the first term in each product is the probability of a given age (from Table 2) and the second term is the age's membership value in the concept *young*. (For more details see Buoncristiani, 1980; 1983; Khalili, 1979; Klement & Schwghla, 1981; Ralescu & Ralescu, 1984; Smets, 1982*a*,*b*; Yager, 1982).

### *1.5.2. Linguistic models*

The next three models have two properties in commn. First, they all yield fuzzy probabilities (i.e., membership functions over [0, 1]), which can be interpreted as representing specific linguistic responses. Second, the models all proceed by (a) forming multiple crisp representations of the problem consistent with the overall fuzzy structure, (b) carrying out simplified calculations at each crisp representation, and (c) combining the results of these calculations to yield the resulting membership function over [0, 1]. For example, if the vagueness in a problem is due to a fuzzy event (e.g., *young*), then the model considers all possible crisp events (i.e., intervals of the real line) that agree more-or-less with the fuzzy concept. A probability is estimated for each of the crisp events, and a membership value is calculated for each probability. This procedure is consistent with theoretical developments in behavioral decision theory which postulate that subjects form and act upon simplified representations of problems (Slovic, Fishhoff, & Lichtenstein, 1977).

*The fuzzy probability of a fuzzy event given a crisp random crisp variable (Yager, 1979, 1984a,b; Klement 1982).* Yager noted that intuitively it appears unnatural for the probability of fuzzy events to be real numbers. It would be more natural if the probabilities were fuzzy subsets themselves. Hence, he assumed that for each possible numeric response $p$, subjects evaluate the truth value of the proposition: "the probability of $A$ is at least $p$," and "the probability of $A$ is at most $p$," and combine these evaluations through the min rule. The evaluations are accomplished by imagining or mentally simulating different possible crisp events ($A'_\alpha s$) associated with the linguistic term defining event $A$. Formally, let $A$ be a fuzzy subset of $\Omega$, and let $P$ be a crisp probability measure defined on $\Omega$, then using the extension principle, and following Zadeh's (1981) work on fuzzy cardinality, Yager defined the following three probabilities:

$$\mu_{\mathrm{FGP}(A)}(p) = \sup_{\alpha} \{\alpha \mid P(A_\alpha) \geq p\} \qquad p \in [0, 1]. \tag{1}$$

$\mu_{\mathrm{FGP}(A)}(p)$ should be interpreted as the truth value of the proposition: "the probability of $A$ is at least $p$."

$$\mu_{\mathrm{FLP}(A)}(p) = 1 - \mu_{\mathrm{FGP}(\bar{A})}(p) = \sup_{\alpha} \{\alpha \mid P(\bar{A}_\alpha) \geq 1 - p\}. \tag{2}$$

$\mu_{\mathrm{FGP}(\bar{A})}(p)$ should be interpreted as the truth value of the proposition: "the probability of not-$A$ is at least $p$," and $\mu_{\mathrm{FLP}(A)}(p)$ as the truth value of the

proposition: "the probability of $A$ is at most $p$."

$$\mathrm{FEP}(A) = \mathrm{FGP}(A) \cap \mathrm{FLP}(A), \tag{3}$$

where $\mu_{\mathrm{FEP}(A)}(p)$ is the truth value of the proposition: "the probability of $A$ is $p$."

For the properties of FEP($A$) see Yager (1984*b*).

*Example 2.* Let the concept *young* be defined as in Example 1. Then using the database in Table 2:

$$\begin{array}{lll}
A_\alpha = \{1, 2, 3, 4, 5, 6, 7, 8\} & P(A_\alpha) = 1 & 0 \le \alpha \le 0{\cdot}038 \\
A_\alpha = \{2, 3, 4, 5, 6, 7, 8\} & P(A_\alpha) = 0{\cdot}7 & 0{\cdot}038 < \alpha \le 0{\cdot}072 \\
A_\alpha = \{3, 4, 5, 6, 7, 8\} & P(A_\alpha) = 0{\cdot}6 & 0{\cdot}072 < \alpha \le 0{\cdot}148 \\
A_\alpha = \{4, 5, 6, 7, 8\} & P(A_\alpha) = 0{\cdot}4 & 0{\cdot}148 < \alpha \le 0{\cdot}5 \\
A_\alpha = \{5, 6, 7, 8\} & P(A_\alpha) = 0{\cdot}2 & 0{\cdot}5 < \alpha \le 0{\cdot}735 \\
A_\alpha = \{6, 7, 8\} & P(A_\alpha) = 0{\cdot}15 & 0{\cdot}735 < \alpha \le 1,
\end{array}$$

where $1, 2, \ldots, 8$ are the eight age groups in Table 2, and $A = \mathrm{YOUNG}$. Then FGP(YOUNG) is given by:

$$\mu_{\mathrm{FGP(YOUNG)}}(p) = \begin{cases} 1 & \text{if } 0 \le p \le 0{\cdot}15 \\ 0{\cdot}735 & \text{if } 0{\cdot}15 < p \le 0{\cdot}2 \\ 0{\cdot}5 & \text{if } 0{\cdot}2 < p \le 0{\cdot}4 \\ 0{\cdot}148 & \text{if } 0{\cdot}4 < p \le 0{\cdot}6 \\ 0{\cdot}072 & \text{if } 0{\cdot}6 < p \le 0{\cdot}7 \\ 0{\cdot}038 & \text{if } 0{\cdot}7 < p \le 1{\cdot}0 \end{cases}$$

FLP(YOUNG) is given by:

$$\mu_{\mathrm{FLP(YOUNG)}}(p) = \begin{cases} 0 & \text{if } 0 \le p < 0{\cdot}15 \\ 0{\cdot}265 & \text{if } 0{\cdot}15 \le p < 0{\cdot}2 \\ 0{\cdot}5 & \text{if } 0{\cdot}2 \le p < 0{\cdot}4 \\ 0{\cdot}852 & \text{if } 0{\cdot}4 \le p < 0{\cdot}6 \\ 0{\cdot}928 & \text{if } 0{\cdot}6 \le p < 0{\cdot}7 \\ 0{\cdot}962 & \text{if } 0{\cdot}7 \le p < 1{\cdot}0 \\ 1 & \text{if } p = 1{\cdot}0. \end{cases}$$

And FEP(YOUNG) is given by;

$$\mu_{\mathrm{FEP(YOUNG)}}(p) = \begin{cases} 0 & \text{if } 0 \le p < 0{\cdot}15 \\ 0{\cdot}265 & \text{if } 0{\cdot}15 \le p < 0{\cdot}2 \\ 0{\cdot}5 & \text{if } 0{\cdot}2 \le p < 0{\cdot}4 \\ 0{\cdot}148 & \text{if } 0{\cdot}4 \le p < 0{\cdot}6 \\ 0{\cdot}072 & \text{if } 0{\cdot}6 \le p < 0{\cdot}7 \\ 0{\cdot}038 & \text{if } 0{\cdot}7 \le p \le 1{\cdot}0. \end{cases}$$

*The fuzzy probability of a crisp event given a crisp random unimodal fuzzy variable (Kwakernaak, 1978, 1979).* Consider the information presented in Table 3 and the question: "What are the chances that a randomly selected person will be 43 to 48 years old?" This is a query about a crisp event given a crisp random fuzzy variable.

Note that Kwakernaak uses the term fuzzy random variable for what we call a crisp random fuzzy variable. Our usage emphasizes that the fuzzy component is the variable (function) rather than the probability structure.

Central to this model is the idea of an *original* of a fuzzy function. Consider again the data in Table 3 and imagine that these are six friends of yours whose exact ages you have forgotten. You can only assign each of them to a linguistic age category. Given this assignment you may consider many different numeric assignments, some of which are more compatible with the linguistic assignment than others. Each numeric assignment is treated as a potential original of the linguistic assignment.

Formally, let $(\Omega, \sigma, P)$ be a regular probability triple. Suppose that $U$ is a crisp random crisp variable defined on this triple. Assume now that we are perceiving this random variable through a noisy signal such that the best we can determine is the possibility that $U(\omega) = r$, where $\omega \in \Omega$, and $r \in \mathcal{R}$. Namely, we perceive $U$ as a fuzzy function $\tilde{X}$ mapping $\Omega$ to the set of all fuzzy subsets of the real line $(\tilde{P}(\mathcal{R}))$ given by

$$\omega \xrightarrow{\tilde{X}} X_\omega$$

where $\omega \in \Omega$ and $X_\omega \in \tilde{P}(\mathcal{R})$.

A crisp random fuzzy variable $\tilde{X}$ is called *unimodal* if for each $\omega \in \Omega$, the membership function $X_\omega$ is unimodal.

The random variable $U$, of which this crisp random fuzzy variable is a perception, is called an *original* of the crisp random fuzzy variable. Note that corresponding to a given crisp random fuzzy variable there exist many originals. The set of all possible originals of $\tilde{X}$ is a crisp set $U^*$ of all possible random variables defined on $(\Omega, \sigma, P)$.

The acceptability that $U \in U^*$ is the original of $\tilde{X}$ is given by

$$\inf_{\omega \in \Omega} \{X_\omega(U(\omega))\}.$$

Using Zadeh's extension principle, Kwakernaak defined the expectation of a crisp

TABLE 3
*Probability distribution of linguistic ages in a certain group*

| Probability | Age |
|---|---|
| 0·30 | Very old |
| 0·10 | Old |
| 0·20 | Approximately 45 |
| 0·20 | 45–40 |
| 0·05 | Young |
| 0·15 | Very young |

random unimodal fuzzy variable to be

$$(E\tilde{X})(x) = \sup_{U \in U^*} \{\inf_{\omega \in \Omega} \{X_\omega(U(\omega))\}\}, \qquad x \in \mathcal{R}.$$

Next, given a crisp event $A$, Kwakernaak defined a set of indicator functions of the fuzzy event $\tilde{X} \in A$. Each function is associated with one element of the sample space $\Omega$.

For a fixed $\omega \in \Omega$,

$$I_\omega^{\tilde{X} \in A}(\pi) = \begin{cases} r''_A(\omega) & \text{if } \pi = 0 \\ \min[r'_A(\omega), r''_A(\omega)] & \text{if } 0 < \pi < 1 \\ r'_A(\omega) & \text{if } \pi = 1. \end{cases}$$

where

$$r'_A(\omega) = \sup_{x \in A} X_\omega(x), \quad \text{and} \quad r''_A(\omega) = \sup_{x \in A'} X_\omega(x).$$

For a fixed $\omega \in \Omega$ and $\pi \in [0, 1]$, the number $I_\omega^{\tilde{X}}(\pi)$ indicates the acceptibility that a fraction $\pi$ of the point $\omega$ belongs to the event $A$.

Finally, the probability of $A$ (a Borel set in $\mathcal{R}$) is defined to be

$$P(\tilde{X} \in A) = EI^{X \in A}$$

where $I^{\tilde{X} \in A} = \{I_\omega^{\tilde{X} \in A} \mid \omega \in \Omega\}$, and associated with each $I_\omega^{\tilde{X} \in A}$ is the probability of $\omega$, hence $I^{\tilde{X} \in A}$ is a crisp random fuzzy variable.

This model assumes that given a crisp event $A$, subjects first evaluate the acceptability that the image of each member of the sample space under the fuzzy function belongs to the event $A$. The outcome of this evaluation is a function $I_\omega^{\tilde{X} \in A}: [0, 1] \to [0, 1]$. Associated with each function is the probability of $\omega$. Next, subjects mentally simulate all possible originals of $I^{X \in A}$ and compute the expected value under each possible original. The outcome is then combined according to the max–min rule.

*Example 3.* Consider again the information presented in Table 3 and the question: "What are the chances that a randomly selected person will be 43 to 48 years old?" Let the linguistic ages be represented by the following fuzzy subsets of $\mathcal{R}$:

$$\mu_{\text{OLD}}(u) = \begin{cases} 0 & \text{if } u \le 50 \\ (1 + ((u - 50)/5)^{-2})^{-1} & \text{if } 50 < u < 100 \\ 1 & \text{if } u \ge 100 \end{cases}$$

$$\mu_{\text{YOUNG}}(u) = \begin{cases} 1 & \text{if } u \le 25 \\ (1 + ((u - 25)/5)^{2})^{-1} & \text{if } 25 < u < 100 \\ 0 & \text{if } u \ge 100 \end{cases}$$

$$\mu_{\text{VERY OLD}}(u) = [\mu_{\text{OLD}}(u)]^2$$

$$\mu_{\text{VERY YOUNG}}(u) = [\mu_{\text{YOUNG}}(u)]^2$$

$$\mu_{\text{APPROXIMATELY}\,45}(u) = \begin{cases} 0 & \text{if} \quad u \leq 40 \\ (u-40)/(45-40) & \text{if } 40 < u \leq 45 \\ (50-u)/(50-45) & \text{if } 45 < u \leq 50 \\ 0 & \text{if } 50 < u \end{cases}$$

$$\mu_{45-40}(u) = \begin{cases} 1 & \text{if } 40 \leq u \leq 45 \\ 0 & \text{otherwise} \end{cases}$$

Using Kwakernaak's definition (and algorithms for the discrete case), the probability that a randomly selected person from the population described in Table 3 will be 43 to 48 years old is given by:

$$\mu_{P(X \in [43,48])}(p) = \begin{cases} 0{\cdot}6 & \text{if } 0 \leq p < 0{\cdot}2 \\ 1 & \text{if } 0{\cdot}2 \leq p \leq 0{\cdot}4 \\ 0{\cdot}07 & \text{if } 0{\cdot}4 \leq p \leq 0{\cdot}45 \\ 0{\cdot}0049 & \text{if } 0{\cdot}45 < p \leq 0{\cdot}6 \\ 0 & \text{if } p > 0{\cdot}6 \end{cases}$$

(For other developments in this category see Kruse, 1982; 1984; Miyakoshi & Shimbo, 1984.)

*The fuzzy probability of a crisp event given a fuzzy random crisp variable (Zadeh, 1975).* Consider the information presented in Table 4 and the question: "What are the chances that a randomly selected person will be 43 to 48 years old?" This is a query about a crisp event given a fuzzy random crisp variable.

The probability should be the "sum" of *doubtful, almost impossible,* and *unlikely.* If linguistic probabilities are represented as fuzzy numbers of the [0, 1] interval, it is necessary to have a formula for adding fuzzy numbers. However, the fuzzy numbers assigned to these linguistic terms are not independent of each other. For example, if we consider a probability of 0·7 for *likely* (in Table 4), then we cannot consider the value 0·4 for *doubtful.* Thus, the corresponding values are said to be *interactive,* and the usual formula for adding fuzzy numbers (using the extension principle) must be modified to accommodate interactive fuzzy numbers. Zadeh (1975), and Dubois & Prade (1981) have investigated several kinds of interaction, and provided practical methods for the computation with interactive fuzzy numbers.

TABLE 4
*Linguistic probability distribution of ages in a certain group*

| Probability | Age |
|---|---|
| Doubtful | 44 |
| Almost impossible | 45 |
| Unlikely | 47 |
| Likely | 50 |

Zadeh's (1975) model assumes that given a fuzzy probability structure, subjects mentally simulate all possible permissible originals of the fuzzy probability structure. For each permissible original they compute the probability of the crisp event. The outcomes are then combined according to the max–min rule.

*Example 4.* Consider again the information presented in Table 4 and the query: "What are the chances that a randomly selected person will be 43 to 48 years old?"

Let the linguistic probabilities be presented by the following triangular *LR* fuzzy numbers ($L = R$) (see Dubois & Prade, 1978*b*).

$$\text{Doubtful} = (0{\cdot}2, 0{\cdot}1, 0{\cdot}1)_{LR}$$

$$\text{Almost impossible} = (0{\cdot}1, 0{\cdot}05, 0{\cdot}05)_{LR}$$

$$\text{Unlikely} = (0{\cdot}3, 0{\cdot}3, 0{\cdot}3)_{LR}$$

$$\text{Likely} = (0{\cdot}4, 0{\cdot}2, 0{\cdot}2)_{LR}.$$

Then the interactive sum of *doubtful*, *almost impossible*, and *unlikely* is again an *LR* fuzzy number given by (Dubois & Prade, 1981):

$$P(X \in [43, 48]) = (0{\cdot}6, 0{\cdot}05, 0{\cdot}05)_{LR}.$$

Note that $P(X \in [43, 48])$ represents a greater probability than what is understood as the meaning of the concept *likely* ($0{\cdot}6 > 0{\cdot}4$), and is as precise as the term *almost impossible* (0·05).

### 1.6. PURPOSE OF THE EXPERIMENT

Giles (1983) has described the current character of research on fuzzy reasoning as follows:

> A prominent feature of most of the work in fuzzy reasoning is its ad hoc nature. . . . If fuzzy reasoning were simply a mathematical theory there would be no harm in adopting this approach; . . . However, fuzzy reasoning is essentially a practical subject. Its function is to assist the decision-maker in a real world situation, and for this purpose the practical meaning of the concepts involved is of vital importance (Giles, 1983, p. 263).

Fuzzy set theory would benefit from becoming a behavioural science, having its assumptions validated, and having its models verified by empirical findings (Kochen, 1975). In particular, there has been virtually no experimental work done with regard to probability inference in the presence of linguistic inexactness. The experiment to be described next empirically tested the models described in Section 1. Subjects were instructed to estimate the probabilities of certain events given databases similar to those presented in Examples 1 through 4. The subjects' linguistic or numeric responses were then compared to the predicted ones based on the tested models.

Shafer & Tversky (1985) compared a subjective probability model to a formal language. It has a vocabulary—a scale of degrees of probability. Attached to this vocabulary is a semantic structure—a scale of canonical examples that show how the vocabulary is to be interpreted, and psychological devices for making the interpretation effective. Elements of the vocabulary are combined according to a syntax—the theory's calculus. In the context of fuzzy probabilities the vocabulary is a set of functional representations of probabilities, namely their membership functions over

the unit interval. While previous work (e.g., Wallsten *et al.* 1986*a*) was more concerned with the vocabulary itself, this work is aimed at testing the "semantics" of the proposed subjective probability language.

Testing the "usefulness" of these models is the first step toward verifying the semantics of the linguistic probability language (Shafer & Tversky, 1985). A second step, which is outside the scope of the current research, is to validate the language syntax. The final step will be to incorporate this language into a formal decision analysis theory.

## 2. Method

### 2.1. SUBJECTS

Twenty native speakers of English were recruited by placing an advertisement in the students' newspaper, and by distributing this advert among students who participated in a first-year graduate level course in statistics. The advert offered the opportunity to earn cash ($50) for participation in a multi-session experiment on probability estimation conducted by the psychology department. From an initial pool of volunteers, 10 subjects with no probability/statistics background were assigned to the "naive" group (Group N), and another 10 subjects with moderate to advanced background in probability/statistics were assigned to the "sophisticated" group (Group S).

### 2.2. GENERAL PROCEDURE

Subjects were tested for five sessions of approximately 1 to $1\frac{1}{2}$ h each. Sessions 2 and 3 were one integral unit broken into two parts due to the lengthy nature of this unit. Similarly, sessions 4 and 5 were an integral unit. In what follows, sessions 2 and 3 will be referred to as Part 1, and sessions 4 and 5 as Part 2 of the experiment. Part 2 was a replication of Part 1. Session 1 was for practice, and Parts 1 and 2 were for data collection. The practice and four 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 all sessions, subjects worked through five types of tasks comprised of: (1) a probability estimation task (PE), (2) a linguistic probabilities scaling task (LPS), (3) a linguistic ages scaling task (LAS), (4) a linguistic probabilities similarity judgment task (SIM), and (5) a probability estimation—scaling task. The data collected in task (5) proved to be too sparse for proper analysis; hence this task will not be considered further. In addition, sessions 2 and 4 included a sixth type of task called a linguistic probabilities ranking task (RANK). Trials from all tasks were presented in all sessions in an inter-mixed random order. In what follows, tasks (2) through (6) will be referred to as the *auxiliary tasks,* referring to their status as an aid to the models' *verification* rather than directly related to the core issue of probability estimation in the presence of linguistic inexactness.

In session 1 (the practice session), the general nature of the study was described, and then each task was described in more detail. Subjects familiarized themselves with all tasks and modes of response.

Each task will now be described in more detail as it was presented in Session 1.

### 2.2.1. PROBABILITY ESTIMATION TASK (PE)

The instructions for this task read in part (for complete instructions see Zwick, 1987*b*):

> "... Every 4th of July, a group of civic clubs organizes a family event in Kenan Stadium. Imagine an especially great success 1 year. Some families brought their youngest children with them, some of whom were only several weeks old. Other families included the great and even great great grand parents, some of whom were 100 years old. So the audience in Kenan Stadium that day was extremely variable with respect to age, with some people who were just born, and others who were as old as 100 years. But all of them were there for the same reason—celebrating Independence Day.
>
> Now imagine that we select 100 people from this crowd, and we move them to the club house at the far end of the field. We will select people on the basis of their age only. For example, we might decide that we want to select 100 people such that 10 of them are 20 years old, 30 of them are 35 years old and the rest are 50 years old. Or we might decide that we want 100 people, two persons from each age group of 50 up to 100 (two persons who are 50 years old, two are 51 years old, two are 52 years old, and so on up to 99 years old). Or we might decide to choose 100 people such that there are more old people than young people in this group, and only few are very young.
>
> Now suppose that after we have made our selection of people and they are all in the club house, we randomly pick one of them to win a valuable prize. Certainly the chances of picking a person of a certain age will depend on the age structure of the group we brought to the club house. For example, if we decided to choose 100 people all of whom are 50 years old, then it is certain that a 50 year old person will be selected to win the prize. Or if we decided to choose 100 people such that all of them are either old or very old, but more are very old, then the chances that a very old person will win the lottery are higher than the chances that an old person will win it.
>
> On each trial we will select a different group of 100 people. On some trials we will tell you how we selected the ages for the group of 100 people, and then we will ask you to judge the chances that a randomly selected person from the group (the winner of the prize) will belong to a certain age category. For example we might ask you what are the chances that a randomly selected person from this group is '*middle age*'.
>
> Sometimes we will tell you exactly how this group is structured . on other trials we might tell you the probability structure of the age groups only in general terms. . . . After we have told you in some fashion about the age structure, we will ask you about the chances that a randomly selected person will be of a certain age or age category. We might ask about a precise numerical age such as 35 years old, or about an age category expressed by a linguistic term such as old, young, etc
>
> Sometimes we will ask you to respond with a probability number (i e , a number between 0 and 1), and on other trials to respond with a probability word, such as *slight chance, probable*, etc."

There were six different types of probability estimation trials, corresponding to the six rows with models in Table 1. Table 5 presents the structure of each type of trial (including the symbolic notation of each type, and the model each type is testing). "Probability structure" refers to the way the information regarding the age composition of the selected group was presented "Attribute" refers to the way the actual ages were presented. "Event" refers to the way the question was formulated, and finally "Response" refers to the instruction given to the subjects as to how they should respond. Due to errors in presenting type I.N1.1. data to the subjects, such that the trials were not compatible with Zadeh's model (1975) assumptions, this type was eliminated from further consideration Figures 1 to 5 present one example of the computer display from each of the remaining types

TABLE 5
*The structure of the problems included in this study*

| Probability structure (Random) | Attribute (Variable) | Event | Response | Symbolic notation (Type) | Model Tested |
|---|---|---|---|---|---|
| Numeric | Numeric | Numeric | Numeric | NNNN | Classical probability theory |
| Numeric | Numeric | Linguistic | Numeric | NNLN | Zadeh (1968) |
| Numeric | Numeric | Linguistic | Linguistic | NNLL | Yager (1984) |
| Numeric | Linguistic | Numeric | Linguistic | NLNL | Kwakernaak (1978) |
| Linguistic | Numeric | Numeric | Linguistic | LNNL | Zadeh (1975) |
| Linguistic | Numeric | Linguistic | Linguistic | LNLL | Zadeh (1975) |

For the numeric probability structures, a uniform distribution was used for simplicity. In each part, subjects responded to five problems from type NNNN, and 15 problems from each of the other types (NNLN, NNLL, NLNL, and LNNL), for a total of 65 different problems. All problems were presented in a random order intermixed with all other tasks. In a numeric response problem (types XXXN), subjects were instructed to type in their response (any number from 0 to 1) using the keyboard. When a linguistic response (types XXXL) was required, a list of probability words and a list of modifiers was presented at the bottom of the screen.

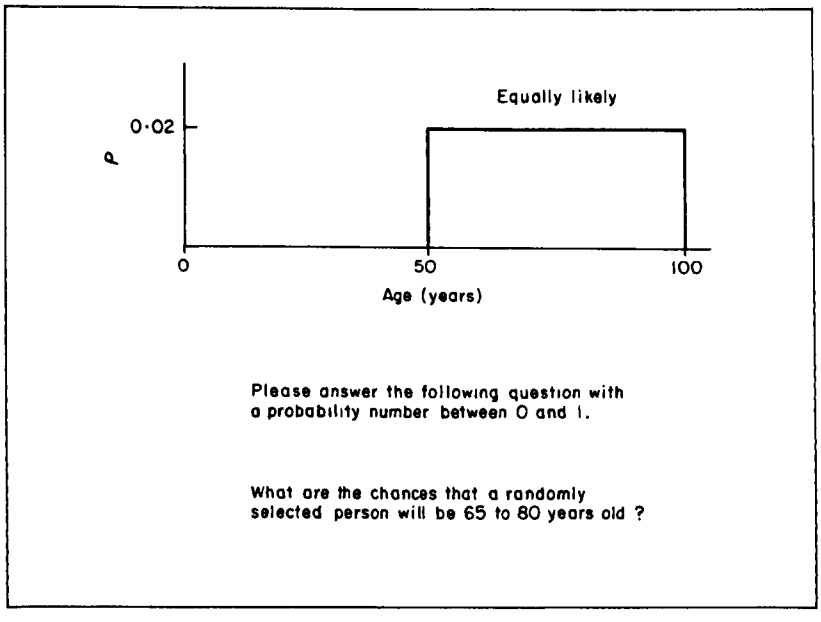



FIG. 1. Probability estimation trial, type NNNN.

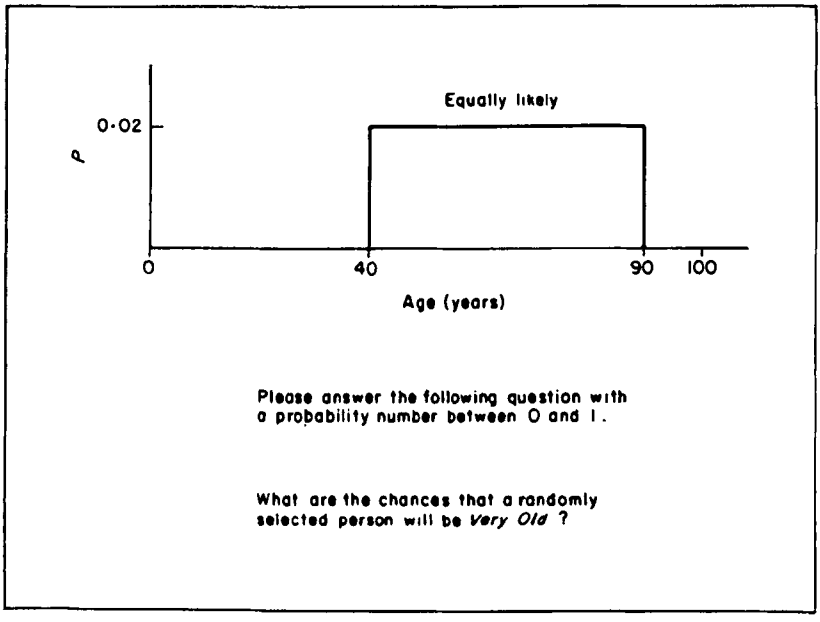



FIG. 2. Probability estimation trial, type NNLN.

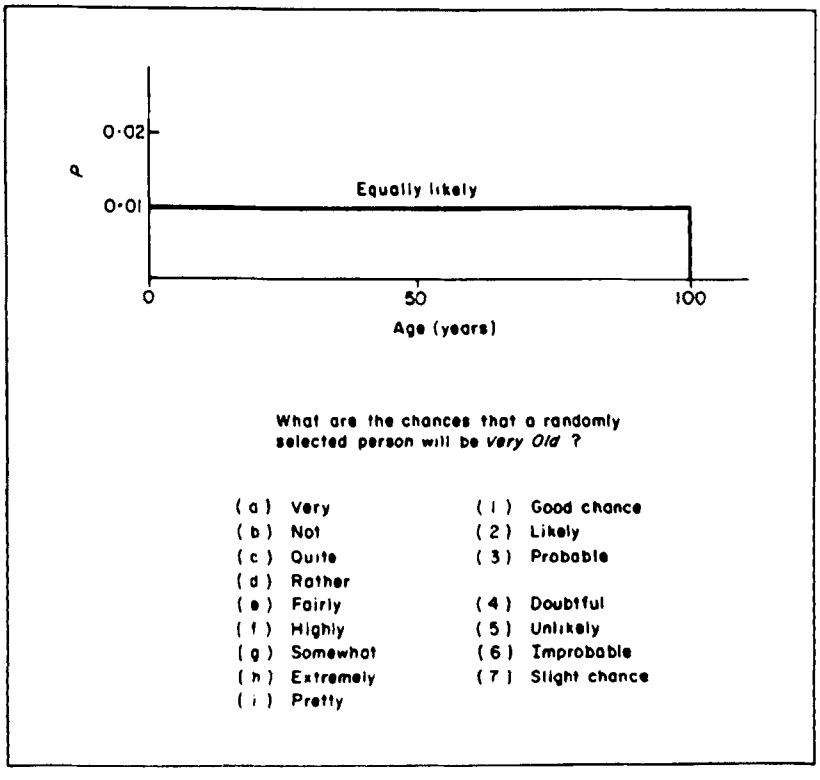


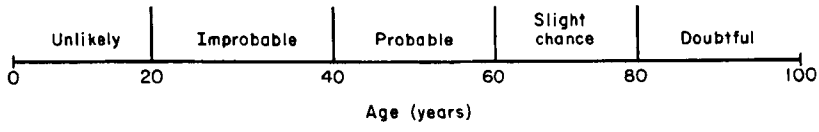

FIG. 3. Probability estimation trial, type NNLL.

| *Probability* | *Age* |
|---|---|
| 0·16 | Very Old |
| 0·28 | Old |
| 0·11 | Middle Age |
| 0·40 | Young |
| 0·05 | Very Young |

What are the chances that a randomly selected person will be *13 to 45 years old* ?

| | |
|---|---|
| ( a ) Very | ( 1 ) Good chance |
| ( b ) Not | ( 2 ) Likely |
| ( c ) Quite | ( 3 ) Probable |
| ( d ) Rather | |
| ( e ) Fairly | ( 4 ) Doubtful |
| ( f ) Highly | ( 5 ) Unlikely |
| ( g ) Somewhat | ( 6 ) Improbable |
| ( h ) Extremely | ( 7 ) Slight chance |
| ( i ) Pretty | |

FIG. 4. Probability estimation trial, type NLNL.




What are the chances that a randomly selected person will be *20 to 60 years old* ?

| | |
|---|---|
| ( a ) Very | ( 1 ) Good chance |
| ( b ) Not | ( 2 ) Likely |
| ( c ) Quite | ( 3 ) Probable |
| ( d ) Rather | |
| ( e ) Fairly | ( 4 ) Doubtful |
| ( f ) Highly | ( 5 ) Unlikely |
| ( g ) Somewhat | ( 6 ) Improbable |
| ( h ) Extremely | ( 7 ) Slight chance |
| ( i ) Pretty | |

FIG. 5. Probability estimation trial, type LNNL.

Subjects could choose one of seven primary terms, or they could combine one primary term with one or two of nine modifiers. The primary terms were: *good chance, likely, probable, doubtful, unlikely, improbable,* and *slight chance.* The modifiers were: *very, not, quite, rather, fairly, highly, somewhat, extremely,* and *pretty.* These terms were chosen on the basis of a pilot study, in which subjects generated their own linguistic responses. The terms that were used most frequently by subjects in the pilot study were included in the current study. The primary terms were presented in a column at the bottom right of the screen. These terms were ordered from most to least likely based on each subject's response in the linguistic probabilities ranking task (details of this task will be presented later in this section).

2.2.2. LINGUISTIC PROBABILITIES SCALING TASK (LPS)

The objective of this task was to establish the subject's membership function for various linguistic probability phrases. Recently, Wallsten, *et al.* (1986*a*) developed a method for empirically establishing the memberhsip functions of fuzzy concepts, based on conjoint measurement and utilizing a graded pair-comparison technique. Rapoport *et al.* (1987) further established that the methods of direct magnitude estimation and graded pair-comparison yield similar membership functions. In this study we adopted the direct magnitude estimation technique, which is much shorter. Instructions for this task said in part (the probability number and phrase refer to those that appeared on the screen in the practice session:

> "... At the top of the screen is a question:
> 'How well is 0·5 described by *probable*?'
> The probability phrase should be thought of in the age context we have been using, namely that it represents the chances that a randomly selected person will be a particular age or a particular age category. You are to indicate how well the probability phrase describes the numerical probability in this context. If you think that the probability number (0·5) is very well described by the phrase, move the arrow all the way to the right.... If you think that the probability number is not at all well described by the phrase, move the arrow all the way to the left.... The relative location of the arrow on the line should correspond to how well (right) or how poorly (left) the phrase described the numerical probability."

The phrases each subject used were classified as expressing high or low probability, according to the following rules:

(1) A priori classification of the primary terms. *Good chance, likely,* and *probable* were classified as expressing high probability. *Doubtful, unlikely, improbable,* and *slight chance* were classified as expressing low probability.
(2) Any combination of a "primary term" [see (1)] with a modifier, except the modifier *not,* was classified as belonging to the primary term category.
(3) Any combination that included the modifier *not,* was classified as belonging to the category that the "primary term" does not belong to. (This procedure produced few misclassifications, all of which involved the combination of *not very* or *not highly* with a primary term. Fortunately, these combinations were seldom used).

"High" probability terms were presented once (in Part 1) with each of the following high "core" probabilities: 0·45, 0·55, 0·65, 0·75, 0·85, and 0·95. "Low" probability terms were presented once with each of the following low "core"

probabilities: 0·05, 0·15, 0·25, 0·35, 0·45, and 0·55. The probabilities that actually appeared on the screen were randomly chosen from the following interval ["core" −0·03, "core" +0·03]. We adopted this procedure because: (1) Presenting the same 12 numeric probabilities again and again might cause the subjects to treat the probablity distributions as discrete rather than continuous ones, as they should be treated; and (2) Assuming consistent responses between parts, this procedure facilitates subsequent curve-fitting (see Results section), by providing more data points along the probability axis. We also assumed that since the underlying scaled concepts are continuous, this technique would not impair the ability to test response consistency between parts. The total number of trials from this task varied depending on the number of probability phrases generated by each subject.

#### 2.2.3. LINGUISTIC AGES SCALING TASK (LAS)

The objective of this task was to establish the subject's membership function for various linguistic age phrases. Again, we adopted the direct magnitude estimation technique.

Instructions for this task said in part (the linguistic and numeric ages refer to those that appeared on the screen during the practice session:

> "... At the top of the screen is a question:
>
> 'How well is 45 described by old?
>
> You are to indicate how well the age phrase described the numerical age in the context of the whole population at Kenan Stadium. If you think the number (45) is very well described by the phrase (old), move the arrow all the way to the right.... If you think that the number is not at all well described by the phrase, move the arrow all the way to the left.... The relative location of the arrow on the line should correspond to how well (right) or how poorly (left) the phrase described the numerical age."

In Part 1 each linguistic age was presented once with each of 10 different numeric ages. The following numeric ages were used:

*Very Young:* 2, 5, 6, 8, 11, 12, 14, 16, 20, 26

*Young:* 7, 13, 16, 19, 24, 29, 31, 36, 39, 45

*Middle Age:* 23, 28, 33, 36, 44, 47, 50, 53, 62, 69

*Old:* 37, 46, 53, 61, 68, 72, 75, 82, 84, 93

*Very Old:* 43, 57, 62, 66, 74, 78, 85, 92, 98, 100

Thus, there was a total of 50 trials.

#### 2.2.4. LINGUISTIC PROBABILITIES SIMILARITY-JUDGMENT TASK (SIM)

The objective of this task was to choose the best similarity index between membership functions within a subject. This index was used to determine whether subjects selected the linguistic responses that were the "closest" to the ones that were predicted by the various models under investigation. Zwick, Carlstein & Budescu (1987) have noted that if the objective is to accurately model the behavior of a specific individual, then the ideal strategy is to determine the best similarity measure for that individual. This study follows that recommendation. Subjects were

asked to judge the similarity between probability phrases that were classified as belonging to the same category (high or low category).

Instructions for this task read in part:

> "At the center of the screen you see two non-numerical probability phrases. You have to judge how similar, or synonymous, you consider the two phrases to be with respect to describing the probability in the age context we have been using. If you think that the two phrases are absolutely similar, move the arrow all the way to the right. If you think that the two phrases are not at all similar, move the arrow all the way to the left. The relative location of the arrow on the line should correspond to how similar (right) or how dissimilar (left) the two phrases are.

In Part 1 subjects judged the similarity among eight high probability, and among eight low probability phrases. Each pair of words was presented once for a total of 56 trials. For each subject, each category contained the first eight probability phrases that were chosen by him or her.

#### 2.2.5. LINGUISTIC PROBABILITIES RANKING TASK (RANK)

Whenever a linguistic response was required in a probability estimation trial, a list of "primary terms" appeared at the lower right portion of the screen. It was assumed that ordering these terms from most to least likely would facilitate the search for the appropriate response. The objective of this task was to guarantee that this ordering would agree with each subject's opinion.

The instruction for this task said in part:

> "At the bottom of the screen you see a list of probability words. You must choose the word that expresses the highest probability. Note that corresponding to each word is a number. Type in the number that corresponds to the word that in your opinion expresses the highest chance. . . . As you can see this word has disappeared from the list and instead is appearing now at the top of a new list at the top of the screen.
>
> Next you have to choose again the word that expresses the highest probability among the remaining words in the list at the bottom of the screen. . . . Note again that the chosen word was eliminated from the bottom list and instead is now appearing at the second place in the top list. You have to repeat this task until no word is left in the bottom list.
>
> Note that the top list should agree with your opinion about the way to rank order these probability words from most to least likely."

Each part started with this task.

## 3. Results

Recall that in this study subjects were presented with both primary and auxiliary tasks. The auxiliary category included the two scaling tasks (linguistic ages and probabilities), and the linguistic probabilities similarity and ranking tasks. The primary category included the probability estimation task. The structure of the results section follows the category distinction. First the results of the auxiliary tasks will be reported. These results include subjects' reliability, practical estimation of membership functions, and choosing the best similarity index—all at the individual level. Next the results of testing the models will be organized and reported by response mode. Numerical models will be reported first (classical probability theory,

and Zadeh, 1968, [NNLN]), followed by the linguistic models (Yager, 1979, 1984*b*, [NNLL]; Kwakernaak, 1978, [NLNL]; and Zadeh, 1975, [LNNL]).

### 3.1. AUXILIARY TASKS

#### *3.1.1. Linguistic probabilities scaling task* (LPS)

*Reliability.* A prerequisite for any model testing is that subjects' responses be consistent beyond the variability expected due to the vagueness inherent in the meaning of the scaled concepts. Because Part 2 was a partial replication of Part 1, it is possible to assess subjects' reliability in the direct scaling tasks in which subjects were instructed to respond by locating the arrow on a directed line.

Linear correlations and the estimated slopes and intercepts of the linear structural relation between responses in Parts 1 and 2 were used to assess reliability. A consistent subject should demonstrate a high correlation and a linear structural relation with an intercept of 0, and a slope equal to 1. Based on the assumption that close probability numbers have close membership values in the same concept, repeated trials were included in this analysis only if the actual probabilities presented on the screen were not farther than 0·03 apart.

The linear correlations were high for all subjects (ranging from 0·84 to 0·53), but the intercept and slope of the linear structural relationships differed significantly from (0, 1 respectively) for subjects 6 and 11. Because the ability to measure the vague meanings of the probability terms is a prerequisite for any attempt to test the linguistic models, subjects 6 and 11 were removed from further linguistic model testing analysis. Considering the demonstrated high reliability of the other subjects in this task, all subsequent analyses in this task (LPS) were done by combining the two parts.

*Practical estimation of membership function (linguistic probabilities).* We have adopted what can be called an implicit analytical definition, or a parametric approach to modeling membership functions. In this approach, the general shape of the function (a family) is assumed to be known, and the particular member of the family is estimated by using the experimental data. Two types of membership functions are common throughout fuzzy set literature—monotonic and single-peaked (unimodal, nonmonotonic) functions. Hersh, Caramazza & Brownell (1979), adopted the linear family, and estimated the parameters by using regression techniques. Zysno (1981) adopted an exponential family for the monotonic functions, and a combination of two exponential for the single-peaked functions. The slopes and the inflection points of these functions were estimated by regression analysis.

Based on the shapes of membership functions obtained in previous research on scaling linguistic ages and probabilities (Zysno, 1981; Wallsten, *et al.*, 1986*a*; Rapoport *et al.*, 1987), we concluded that the cubic polynomial family can adequately represent the membership functions of these concepts. Cubic polynomials were fitted to the subjects' responses, using a least squares technique. Each function was restricted to be non-negative then rescaled and normalized to attain the

value of 1 on at least one point in the appropriate interval. (See Zwick (1987*b*) for details about the rescaling procedure.)

For the purpose of model testing it is necessary that the fitted cubic functions be good representations of a subject's responses. Note that the usual application of regression analysis is to predict the dependent variable given the independent variables. In the traditional case, a flat regression line, parallel to the $x$-axis, is not informative and the regular indices of $R^2$ and $F$ values would indicate that the regression model is not significant. In our context, a flat membership function is appropriate if it accurately reflects subjects' perception of the concept (e.g., 12% of the subjects in Wallsten *et al.* (1986*a*) Experiment 1 judged *possible* to have a flat membership function). For this reason, we report and rely only on the Root Mean Sqaure Error (RMSE) index as a goodness of fit measure.

As expected, the RMSE varied considerably between words within a subject, and between subjects within a word, The average RMSE across subjects ranged from 0·099 for Subject 1 to 0·233 for subject 5. By using the RMSE data we have identified those words that are poorly fitted by the cubic function, and to guard against the possibility that these poorly fitted phrases distorted the models' testing results, we tested the models with and without these words. There was virtually no influence on the results. Thus, we report only the results of the analysis that used the entire vocabulary of each subject.

### *3.1.2. Linguistic ages scaling task* (LAS)

*Reliability.* All subjects were highly reliable in this task as is evident by the high linear correlations between Parts 1 and 2 (range from 0·97 to 0·61) and by the inability to reject the assumption that the intercept and the slope of the linear structural relationships between responses in Parts 1 and 2 are simultaneously equal to (0, 1). Considering these results, all subsequent analyses were done over the two parts combined.

*Practical estimation of membership function* (*linguistic ages*). The same procedure was used as with the linguistic probabilities. As with linguistic probabiities, the RMSE varied considerably between words within a subject, and between subjects within a word. The average RMSE across subjects ranged from 0·088 for subject 17 to 0·248 for subject 12. The small number of points within each distribution (5) does not allow the reliable detection of outliers. When the RMSE index for linguistic probabilities is compared to that for linguistic ages (within a subject), it turns out that in all cases the RMSE value associated with an outlier in the LPS task is much higher than the maximum RMSE associated with a linguistic age. Based on the above, we did not eliminate any of the linguistic ages.

### *3.1.3. Similarly task* (SIM)

*Reliability.* Most subjects were not consistent in this task as evaluated by the linear correlations and by the slopes and intercepts of the linear structural relationships between responses in Parts 1 and 2. Recall that in this task, subjects judged the similarity between probability words separately within the high and low probability

groups. Most subjects judged the words within each group to be very similar to each other, and consequently used only the right-most part of the response continuum, close to the label *absolutely similar*. This restriction of range resulted in low reliability measures. Subjects whose responses did not deviate significantly from the consistency line judged the members of one or two pairs of words to be quite dissimilar—judgments that increased response variability and consequently increased the chances of no significant deviation from the consistency line.

*Determining the best similarity index.* Recall that the models discussed in the Introduction (except Zadeh's, 1968) generate membership functions over the unit interval to represent the appropriate probabilities. Generally, one would not expect the resulting fuzzy set to exactly correspond to any of the fuzzy sets assigned to linguistic probabilities in the LPS task. One approach is to find the probability phrase (presented by its cubic polynomial, estimated by the LPS data) that is most "similar" to the fuzzy set resulting from the models' computations. Such a term would then be called a *linguistic approximation*. This is an analogy to statistics, where empirical distribution functions are often approximated by well-known standard distribution functions.

The purpose of the similarity task was to determine the best similarity index between fuzzy sets. Considering the unreliability in the similarity judgments between parts, we compared the discriminative power of all indices in each part separately. For each subject the index that performed well in both parts was used in testing the models. (See Zwick *et al.* (1987) for a description of the indices tested, and for theoretical background.)

Similarly indices were compared within parts in the following way: For a particular subject and a particular similarity measure (within a part), the correlation between the "true" similarity rating (SIM task) and the similarity measures over all pairs of words was calculated. For each subject and part, all indices were ranked from best to worst (from the highest to the lowest correlation). Within subjects, the index with the highest average rank was chosen. We preferred the highest average rank, rather than the highest average correlation, to avoid choosing an index that performs extremely well in one part, but quite poorly in the other. Using the best average rank procedure guarantees that the index is robust to a moderate fluctuation in the similarity judgments. This is important considering that all other analyses (fitting membership functions, and model testing) were done over the two parts combined. Generally, the best indices predict subjects' judgments relatively well.

### 3.2. PRIMARY TASKS: MODEL TESTING

#### *3.2.1. Numeric response models*

*Type NNNN (classical probability theory).* This probability estimation task was included in the experiment as an additional indicator of the statistical sophistication of the subjects. The problems are classical probability questions similar to those dealt with in any introductory probability course. Since the two groups (N and S) were formed to differ in probability sophistication, we assumed that they would differ in the performance of this task.

Table 6 presents the number of times each subject's response were precisely correct in this type of a problem and the mean absolute difference (MAD) between the observed and the correct response. (There were five problems in each part for a total of 10 NNNN problems.) As can be seen, this task failed to discriminate between the two groups. The average number of correct responses in both groups is 5·4, and the mean MAD for groups N and S is 0·046 and 0·041, respectively. This finding and the fact that four subjects from the "sophisticated" group (subjects 15, 16, 17, and 20) responded incorrectly more often than correctly, put in doubt the assumption that this task measures probability sophistication. Considering these findings, we do not consider this task as an indication of statistical sophistication.

*Zadeh's model (1968)* (*Type* NNLN). An *a priori* question is whether this task was meaningful from the subjects' point of view. Subjects might be unable to answer these questions intelligently, and since they were required to respond, they did so randomly. To test this possibility a reliability check, similar to the one performed on the scaling data was performed.

Based on the correlations and the intercepts and slopes of the structural relationships between responses to replicated problems in Parts 1 and 2, subjects 1,

TABLE 6

*Number of correct responses (out of 10) and the mean absolute difference (MAD) between observed and correct responses in type NNNN (task PE)*

| Group | Subject | Number of correct responses | MAD |
|---|---|---|---|
| N | 1 | 3 | 0·084 |
| | 2 | 1 | 0·095 |
| | 3 | 8 | 0·012 |
| | 4 | 2 | 0·119 |
| | 5 | 0 | 0·099 |
| | 6 | 8 | 0·006 |
| | 7 | 8 | 0·012 |
| | 8 | 8 | 0·012 |
| | 9 | 10 | 0·000 |
| | 10 | 6 | 0·016 |
| Mean | | 5·4 | 0·046 |
| S | 11 | 9 | 0·006 |
| | 12 | 9 | 0·006 |
| | 13 | 7 | 0·046 |
| | 14 | 6 | 0·016 |
| | 15 | 0 | 0·110 |
| | 16 | 2 | 0·109 |
| | 17 | 4 | 0·077 |
| | 18 | 8 | 0·012 |
| | 19 | 8 | 0·012 |
| | 20 | 1 | 0·020 |
| Mean | | 5·4 | 0·041 |

2, 4, 5, 10, 15 and 17 were identified as unreliable in this task, and were not used to test Zadeh's model.

Table 7 presents the mean absolute deviations (MAD) and the intercepts and slopes of the regression lines between observed and predicted responses. The model is unsuccessful in predicting the responses of 10 out of 13 reliable subjects. In most cases, the model over-estimated subjects' responses.

### *3.2.2. Linguistic-response models*

*Consistency.* A prerequisite for any model testing is that subjects' responses be consistent across replications. This issue of consistency is complicated by the nature of the linguistic response. Since different phrases might convey the same meaning, consistency should be defined by the type rather than the token of the response. Because subjects almost never replicated their exact linguistic responses across parts, it is necessary to determine whether the two responses to a problem are similar in meaning. The average distance between replicated responses by subject and problem type was computed by applying the best similarity measure to the

TABLE 7

*Average mean absolute deviations (MAD) and slopes and intercepts of the regression lines between observed and predicted responses in type NNLN (task PE)*

| Group | Subject | Intercept | Slope | MAD |
|---|---|---|---|---|
| N | 1(UR) | — | — | — |
| | 2(UR) | — | — | — |
| | 3 | 0·010 | 0·645* | 0·135† |
| | 4(UR) | — | — | — |
| | 5(UR) | — | — | — |
| | 6 | −0·019 | 0·466* | 0·342† |
| | 7 | −0·004 | 0·742* | 0·096† |
| | 8 | 0·047 | 0·555* | 0·128† |
| | 9 | −0·038 | 0·942 | 0·071 |
| | 10(UR) | — | — | — |
| S | 11 | −0·043 | 0·912 | 0·090 |
| | 12 | 0·028 | 0·619* | 0·105† |
| | 13 | 0·056 | 0·675* | 0·070† |
| | 14 | 0·086 | 0·447* | 0·157† |
| | 15(UR) | — | — | — |
| | 16 | −0·056 | 1·030 | 0·105 |
| | 17(UR) | — | — | — |
| | 18 | 0·061 | 0·363* | 0·192† |
| | 19 | 0·004 | 0·752* | 0·087† |
| | 20 | 0·028 | 0·574* | 0·176† |

* Indicates that the assumption that the slope = 1 or that the intercept = 0 should be rejected ($p < 0{\cdot}05$).

† $p < 0{\cdot}05$ (Wilcoxon test).

UR Unreliable subject in this task.

membership function representations of the linguistic responses. Using the distances between all pairs of words in each subject's vocabulary, one can find the (approximate) sampling distribution of the mean distance between 15 randomly selected pairs (corresponding to the 15 problems in each type). This is equivalent to the assumption that subjects are responding randomly, hence each possible pair of responses is equally likely. The probability of finding a mean distance that is less than or equal to the observed mean distance under the random choice assumption can be found. Most subjects passed this weak test of consistency ($p < 0{\cdot}05$). In addition to Subjects 6 and 11 who were found to be inconsistent in the LPS task, Subjects 4 and 12 demonstrated a lack of consistency across types and were removed from further analysis regarding these models. In addition, the following subjects demonstrated a lack of consistency in one type or another and were removed from testing the corresponding model ($p > 0{\cdot}05$): subjects 1, 5, 15 and 17 from testing Yager's model (NNLL), subjects 2 and 8 from testing Kwakernaak's model (NLNL), and subject 1 from testing Zadeh's model (LNNL).

*Model testing technique.* As was explained in Section 2, the models (Yager's, Kwakernaak's, and Zadeh's) were tested by comparing them to an alternative simple baseline model. Recall that in these problems (NNLL, NLNL, and LNNL) subjects were instructed to respond by choosing one of seven primary linguistic terms, alone or combined with one or two of nine modifiers. Given these possibilities, each model predicts subjects' responses based on the subjects' membership functions for the lingusitic terms. The alternative model (*unrestricted baseline model*) uses a random process to predict a response from each subject's own vocabulary.

Since many of the words used by the subjects might be considered synonyms we have used a cluster analysis technique to reduce each subject's own vocabulary into fewer equivalence classes. This reduction was done by considering the pairwise distances (using the best similarity index) between all words in each subject's vocabulary. The distances between predicted and observed response clusters were computed for each problem within a subject. In order to test the model, we first obtained a sampling distribution for a null model (the unrestricted baseline model) by considering the discrete sampling distribution of the distances between a randomly selected cluster and the observed response. This is simply the set of all possible distances from the observed response to the set of all clusters in the vocabulary (call this random variable $X_i$). The unrestricted baseline model assumes that responses are independent within a problem (across parts) and between problems (within and across parts). Hence, under the unrestricted baseline model assumptions we have a sequence of 30 independent random variables, $\{X_i\}_{i=1}^{30}$, the mean and standard deviation of which are known. According to the Liapunov version of the central limit theorem (Rao, 1965, p. 107)

$$\gamma_n = \frac{\sum_{i=1}^{n} (X_i - \mu_i)}{C_n}$$

tends to the standard normal. Where

$$E(X_i) = \mu_i, \quad \text{and} \quad C_n = \left(\sum_{i=1}^{n} \sigma_i^2\right)^{1/2}.$$

Based on this approximate sampling distribution we can compute the probability of the standardized observed mean distance (or a smaller value) under the unrestricted baseline model assumptions. A small probability value indicates that the model tested is outperforming the unrestricted baseline model, while a sizeable probability value indicates that the model tested does not improve prediction beyond the performance of a random unrestricted baseline model. All models should pass this initial test to deserve further consideration.

To further investigate the predictive power of the tested models, the number of clusters from which the baseline model is allowed to randomly choose a predicted response is successively restricted. The restriction mechanism is analogous to considering successively more powerful null hypotheses since the predictive power of the unrestricted baseline model depends on the number of clusters in the vocabulary. Few clusters increase the random baseline hit rate, making it harder for the tested model to distinguish itself. We sequentially carried out the same analysis reported above eliminating at each state the cluster that is the farthest away from the observed one on the previous stage.

For each model the analysis is reported by the probability of the standarized mean distance (or a smaller value) under the unrestricted baseline model, and by the percentage of clusters that are eliminated before the model tested ceased to outperform the restricted baseline model ($p > 0{\cdot}05$). Note that low probabilities and high percentage support the tested model versus the unrestricted and the restricted baseline models. (For a full development of the model testing techniques see Zwick (1988).

*Yager's model.* Table 8 presents the results of the analysis described in the previous section for each of the models. For each subject Table 8 presents the number of words used in task PE, and the number of clusters found at the 99·9% level (using the average linkage method of cluster analysis). We consider Yager's model (type NNLL in task PE) first. The unrestricted baseline model column (UBM) presents the probability of the standard observed mean distance, or a smaller value, between the observed linguistic response and the predicted cluster under the unrestricted baseline model assumptions. Recall that a small probability value indicates that Yager's model predicts a subject's responses better than the unrestricted baseline model. Note that this is the case for 11 out of 12 reliable subjects. In most of these cases the probability is very low ($p < 0{\cdot}001$). The restricted baseline model columns (RBM) presents the percentage of clusters that are eliminated before the model ceased to outperform the restricted baseline model. For three subjects (2, 10 & 16) a small restriction (under 31%) is required, indicating that the superiority of Yager's model over the unrestricted baseline model is a weak one. For another six subjects (3, 7, 13, 14, 18 & 19) a moderate restriction is required (between 44% & 67%), indicating a good performance by Yager's model. Finally, for 2 subjects (8 & 9) a large restriction is required (71·4% and 83·3%, respectively) indicating very good predictive ability by Yager's model with regard to these subjects.

*Kwakernaak's model.* The next major columns of Table 8 presents the results of the analysis of Kwakernaak's model (type NLNL). Kwakernaak's model predicts subjects' responses significantly better than the unrestricted baseline model for 10

TABLE 8

*Testing Yager's, Kwakernaak's and Zadeh's models versus the unrestricted (UBM) and the restricted baseline models (RBM)*

| | | | | Yager's model | | Kwakernaak's model | | Zadeh's model | |
|---|---|---|---|---|---|---|---|---|---|
| Group | Subject | Number of words | Number of clusters ($R^2 = 0{\cdot}999$) | UBM $p$ | RBM Eliminated(%) | UMB $p$ | RBM Eliminated(%) | UMB $p$ | RBM Eliminated(%) |
| N | 1 | 33 | 27 | — | — | 0·00 | 11·1 | — | — |
| | 2 | 43 | 33 | 0·00 | 30·3 | — | — | 0·09* | 00·0 |
| | 3 | 19 | 15 | 0·00 | 66·6 | 0·00 | 86·7 | 0·00 | 73·3 |
| | 4 | 33 | 9 | — | — | — | — | — | — |
| | 5 | 43 | 40 | — | — | 0·25* | 00·0 | 0·03 | 22·5 |
| | 6 | 26 | — | — | — | — | — | — | — |
| | 7 | 50 | 14 | 0·00 | 57·1 | 0·00 | 21·4 | 0·00 | 28·6 |
| | 8 | 34 | 7 | 0·00 | 71·4 | — | — | 0·76* | 00·0 |
| | 9 | 15 | 12 | 0·00 | 83·3 | 0·55* | 00·0 | 0·00 | 83·3 |
| | 10 | 32 | 12 | 0·01 | 16·6 | 0·07* | 00·0 | 0·21* | 00·0 |
| S | 11 | 44 | — | — | — | — | — | — | — |
| | 12 | 29 | 15 | — | — | — | — | — | — |
| | 13 | 29 | 20 | 0·00 | 45·0 | 0·00 | 40·0 | 0·99* | 00·0 |
| | 14 | 46 | 38 | 0·00 | 44·7 | 0·02 | 5·3 | 0·11* | 00·0 |
| | 15 | 35 | 18 | — | — | 0·00 | 16·6 | 0·08* | 00·0 |
| | 16 | 28 | 4 | 0·02 | 25·0 | 0·57* | 00·0 | 0·00 | 50·0 |
| | 17 | 48 | 40 | — | — | 0·00 | 35·0 | 0·03 | 22·5 |
| | 18 | 42 | 21 | 0·00 | 47·6 | 0·00 | 33·3 | 0·00 | 28·6 |
| | 19 | 33 | 31 | 0·00 | 48·6 | 0·00 | 48·4 | 0·00 | 58·1 |
| | 20 | 43 | 36 | 0·27* | 00·0 | 0·00 | 77·8 | 0·00 | 88·9 |

—Unreliable subject in this type, based on previous analysis.
* $p \geq 0{\cdot}05$.

out of 14 reliable subjects. For six subjects (1, 7, 14, 15, 17 & 18) a small restriction (less than or equal to 35%) on the vocabulary size is sufficient to eliminate the significant difference between the performance of Kwakernaak's model and the restricted baseline model. For two subjects (13 & 19) a moderate restriction is required (between 40% and 49%), and finally for two subjects (3 and 20), a sizable restriction is required (86·7% & 77·8%, respectively).

*Zadeh's model (1975).* The final major columns of Table 8 presents the results of the analysis of Zadeh's model. For nine out of 15 reliable subjects, Zadeh's model outperforms the unrestricted baseline model. For four subjects (5, 7, 17 & 18) a relatively small restriction (<34%) on the vocabulary size is sufficient to eliminate the significant difference between the predictive power of Zadeh's and the restricted baseline models. For two subjects (16 & 19) a moderate restriction is required (between 44% & 59%), and for another three subjects (3, 9 & 20) a large restriction is required (73·3%, 83·3% & 88·9% respectively).

## 4. Discussion

This study experimentally tested four models proposed in the fuzzy set literature to represent the probability of a fuzzy or a non-fuzzy event, in a fuzzy or a non-fuzzy environment (as well as the classical probability model in a totally crisp situation). The structure of the Discussion section is as follows. First, we discuss the success rate of each model, starting with the numeric-response models (classical probability theory, and Zadeh's, 1968) and ending with the linguistic-response models (Yager, 1979, 1984*b*; Kwakernaak, 1978; Zadeh, 1975). Next, we compare the success rate of the models within and between subjects. Finally, the implications of these comparisons with respect to the underlying psychological processes in estimating probabilities under linguistic inexactness are discussed.

### 4.1. NUMERIC RESPONSES

In two types of problems subjects were instructed to respond numerically. The first type (NNNN) was intended to provide a measure for further discriminating among subjects on the basis of statistical sophistication (the main discrimination was by group membership). The second type (NNLN) tested the descriptive validity of Zadeh's model (for estimating the probability of a fuzzy event given a crisp random crisp variable).

#### *4.1.1. Classical probability theory (type* NNNN)

Classical probability theory failed to describe subjects' responses in this task. On average, subjects responded incorrectly more often than correctly, although substantial individual differences were found (Table 6). The MAD scores indicated that most subjects, although not always correct, were not too inaccurate. On average, the error was within 0·04 of the correct response. Accuracy in this task (measured by the number of correct responses and the MAD index) was group independent, although the two groups were formed *a priori* to differ in terms of statistical sophistication. There were only three possible correct responses to the

NNNN problems: 0·2 to problems 1, 2, and 3; 0·3 to problem 4; and 0·4 to problem 5. All the correct responses are "simple" numbers, in that only one digit is required after the decimal point. Hence the assumption that errors were due to a numerical approximation process should be rejected. We believe that most subjects failed to recognize the simplicity of these problems, although subjects in group *S* had seen similar ones many times in probability courses. These problems (NNNN) were presented to the subjects along with the other types in an intermixed and random order. It is very likely that the subjects failed to recognize the simplicity of these problems due to the linguistic vagueness introduced in the other types. Subjects responded to these problems using the same frame of mind that was needed to cope with the vagueness that existed in other problems, and instead of solving the problems using an exact numerical method they resorted to other techniques that yielded only approximate responses such as considering relative areas on the density graphs. This *Einstellung* effect is very well documented in psychological literature. For example Luchins' (1942) classical experiments showed that subjects, given a series of problems with the same rule for solution, solved test problems by a "blind" application of that rule, even when a much simpler solution was possible.

The findings that our subjects misperceived the nature of the problems in this type should alert us to the possibility that this may also be the case with the other types of problems.

*4.1.2. Zadeh's model (1968)* (*Type* NNLN)

Although most subjects were consistent in this task, the fact that seven out of 20 subjects were not consistent emphasize the difficulty of responding with an exact probability number in the presence of a vague linguistic event. However, the moderate degree of "concordance" among the reliable subjects (0·67) suggests that the task was meaningful at least for these subjects. Nevertheless, Zadeh's model fails to describe the responses of 10 out of 13 reliable subjects.

It can be argued that Zadeh's model fails to predict subjects' responses in this task because the fuzzy-set representations of the linguistic ages are inadequate. This argument can be rejected on two related grounds: (1) The same representations were used with the other models, and they were more successful; and (2) subjects were highly consistent in the linguistic age scaling task, and a good fit was found between the cubic functions and the data.

We believe that the failure of Zadeh's model is due to its underlying assumption that subjects behave as if they integrate probabilities and membership values on the entire range over which the membership function is not zero. The fact that in most cases Zadeh's model predicted higher than observed probability estimates suggests that subjects disregard elements of the fuzzy event if their membership values are below some predetermined threshold.

This possibility raises the issue of the extent to which the subjects actually used graded category membership in this task. Lakoff (1973) suggested that some speakers seem to turn relative judgment of category membership into absolute judgments by assigning the member in question to the category in which it has the highest degree of membership. Kochen (1975) proposed that some people might be "thresholders" who assign an item to a category if its membership value is beyond a certain level, while others might be "estimators" and think in terms of genuine

gradations of membership. He found that approximately half of his sample treated graded categories in an essentially unfuzzy way. Pipino, Van Gigch & Tom (1981) found evidence for the existence of both types of people in the same cognitive task. And finally, Zwick *et al.* (1987) have found that distance measures between fuzzy sets with consistently good descriptive performance all share the property of concentrating their attention on a single value, rather than on the entire function.

All of the above considerations suggest several alternative models to Zadeh's.

(1) *A threshold model.* According to this model, whenever a person is instructed to respond numerically on the basis of a linguistic concept, he or she simplifies the task by acting upon certain precise values consistent with the vague concept. If $\tilde{A}$ is the fuzzy-set representation of the linguistic concept, then the precise values belong to an $\alpha$-level set of $\tilde{A}$ for a specific $\alpha$. $P(\tilde{A})$ is then equal to $P(A_\alpha)$ for some $\alpha \in [0, 1]$. The $\alpha$-level can be perceived as the threshold of belonging to $\tilde{A}$. Assuming that subjects adopt the same $\alpha$-level in all problems, the $\alpha$ parameter can be estimated and the fit of the model to the data determined.

(2) *Graded membership models.* This is a family of models that assumes that subjects do think in terms of genuine gradations of membership. However, this family accepts the possibility that subjects consider fuzzy sets defined over only one part of their support—the most significant part. Namely, $P(\tilde{A}) = P(\tilde{A}_\alpha)$ for some $\alpha \in [0, 1]$. This family of models also allows the integration of probabilities and membership values to be not necessarily multiplicative in nature. With regard to the last point, Yager (1982) has extended Shafer's theory of evidence so that belief structures may involve fuzzy sets. He then obtained under the condition of Bayesian belief structure a family of possible definitions for the real-valued probability of a fuzzy event given a crisp random crisp variable. For example Yager has shown that if $\Omega = \{x_1, x_2, \ldots, x_n\}$ is a finite sample space on which probability (Bayesian) structure, $p(x_i) = p_i$, is defined and if $\tilde{A}$ is a fuzzy subset of $\Omega$, then

$$P(\tilde{A}) = \sum_{i=1}^{n} T(p_i, \mu_A(x_i))$$

for any $t$-norm. The significance of this result is that there exists a family of possible definitions for the probability of a fuzzy event. For example, if we consider multiplication as our $t$-norm function, then $P(\tilde{A}) = \sum_{i=1}^{n} p_i \cdot \mu_A(x_i)$, which is the definition suggested by Zadeh (1968). However if we consider the min operator then we get

$$P(\tilde{A}) = \sum_{i=1}^{n} \min(p_i, \mu_A(x_i)).$$

The threshold and the graded membership models with different $t$-norm operations will be experimentally tested in future research.

To summarize, Zadeh's model (1968) failed to describe the behavior of subjects in this task (NNLN). We believe that this failure is related to Zadeh's model assumptions that subjects consider the entire range on which the membership

function is not zero, and that the integration of probabilities and membership values is multiplicative in nature. Better assumptions might be that either (1) subjects treat vague concepts in an essentially unfuzzy way, or (2) that subjects consider only a restricted range on which the fuzzy set is defined, although on this range they think in terms of graded membership. Furthermore, within this range the integration of probabilities and membership values may take forms other than multiplication (Zwick, Budescu & Wallsten, 1988).

### 4.2. LINGUISTIC RESPONSES

Before evaluating the separate models, it is necessary to discuss two issues that are common to all linguistic models, those of consistency and vocabulary size.

#### *4.2.1. Consistency*

Most subjects were quite consistent in responding linguistically to the problems in this experiment. Only two subjects (4 & 12) were unreliable across all three types of problems that required linguistic responses (NNLL, NLNL, and LNNL). Other subjects were unreliable in only one of the three types.

Several authors have suggested that a greater degree of response consistency over trials will occur if subjects are allowed to give imprecise verbal reports about a vague concept than if they are forced to give precise numeric responses (Kochen, 1975; Zimmer, 1983). Our data do not support this claim. Recall that tasks NNLN and NNLL differed only along the response dimension, hence a direct consistency comparison is possible. For the same problem, subjects were instructed to respond numerically in type NNLN, and linguistically in type NNLL. There was a high degree of correlation between consistency in both tasks. Five out of the seven subjects who were found to be inconsistent in task NNLN were found to be inconsistent in task NNLL as well. Two subjects were unreliable in only the numeric task and three subjects were unreliable in only the linguistic task. Overall, neither the linguistic nor the numerical mode exhibited greater consistency. The reader should note that different consistency tests were performed with the numeric and the linguistic data. Hence these results should be treated with caution.

#### *4.2.2. Vocabulary*

A substantial vocabulary for uncertainty was demonstrated. The number of phrases generated by subjects in the PE task ranged from 15 (subject 9) to 50 (subject 7), with an average of 35·25 phrases. This result can be compared to an average of 13 phrases (both self-produced and from the list) generated by Budescu, Weinberg & Wallsten's (1988) subjects, and is in marked contrast to Zimmer's (1983) results, in which subjects' active lexicons for uncertainty seemed to contain, on average, between five and six expressions each. It is important to note that these experiments are different in many ways, any of which might be responsible for the conflicting findings.

Zimmer (1983) elicited from his subjects the verbal labels (in German) that they were most familiar with and used most frequently, while in our study subjects composed responses from an *a priori* closed set of primary and modifier terms. Familarity and frequency of use were not emphasized. It is possible that in the present study subjects tried to show that they are shrewd problem solvers, and since

it was emphasized that there are no right or wrong answers, subjects expressed their sophistication by producing a rich vocabulary. In light of the conflicting findings, further research is needed to determine the number of linguistic terms that is sufficient to span the probability interval, although individual differences should be recognized.

#### 4.2.3. *Yager's model (type* NNLL)

The quality of a lingusitic model can be evaluated both in terms of the number of subjects for whom the model outperformed the unrestricted baseline model, and in terms of the proportion of clusters that must be eliminated before the model ceases to outperform the restricted baseline model for a given subject.

The results strongly support Yager's model, both across and within subjects. Yager's model outperformed the unrestricted baseline model for 11 out of 12 reliable subjects (91%) (Table 8). The power of the model was further demonstrated by the findings that in eight out of 11 successes almost half or more of the vocabulary's clusters had to be eliminated before the model ceased to outperform the restricted baseline model. The results support the underlying assumptions of the model: that people focus on different crisp events associated with the fuzzy event, and integrate the results of the computations in each of these levels by the max–min rule. The success of the model is even more significant considering the large number of computations needed to derive the predictions. Recall that in contrast with the other models, Yager's model produces a subnormal membership function. To interpret this function as a linguistic probability, the function should be normalized. This process was not necessary in deriving the predicted responses of the other models. The fact that Yager's model was found, nevertheless, to be relatively more successful than the other models, (see next sections) emphasizes its robustness.

#### 4.2.4. *Kwakernaak's model (type* NLNL)

This model outperformed the unrestricted baseline model for 10 out of 14 reliable subjects (71·4%) (Table 8). However, this success is rather weak given that in six out of the 10 cases (subjects 1, 7, 14, 15, 17 & 18) a restriction of less than or equal to 35% on the vocabulary size was enough to eliminate the superiority of Kwakernaak's model over the restricted baseline model.

It is possible that in at least two ways the subjects did not perceive the data in the manner assumed by Kwakernaak's model. If so, this could bias the results of the model evaluation. First, recall that the lingistic vagueness in this case was introduced in the values of the age attribute. The fact that the probabilities assigned to the various linguistic ages sum to one (although these categories are not mutually exclusive) might seem artificial. Kwakernaak (1978) explained this probability structure as if it were generated by the members of the sample space itself, in which each of them classifies himself or herself into one and only one of the linguistic (age) categories. Such an interpretation may work in some cases, however, in our experiment this interpretation was not given in the instructions. It is possible that the fact that the probabilities add to 1 encouraged subjects to treat the linguistic ages as exclusive categories, rather than as elastic fuzzy concepts. Second, a possible "original" to the fuzzy function can be such that, for example, the numerical age

that is assigned to the age category *very old* is less than the numerical age that is simultaneously assigned to the age category *old*. This is explained by Kwakernaak as being possible if one person classifies him/her self as *old*, and the second person classifies him/her self as *very old*, but nevertheless the second person is younger than the first one. This makes sense if we accept the premise that the linguistic assignment is self generated. However, in this experiment the linguistic age assignment was presented to the subjects with no further elaboration on the assignment process. It is very likely that subjects perceived all persons in one age category (say *old*) to be younger than all the persons in a second age category (say *very old*). Such an interpretation once again encourages subjects to treat the age categories as mutually exclusive, and eliminates many possible "originals" that Kwakernaak's model considers. These two points should be dealt with an any future research by explicitly presenting the linguistic age structure as self generated.

Finally, Kwakernaak's model assumes a sophisticated underlying process which is highly demanding. With this in mind it is interesting to note the different performance of the model with regard to the two groups. Kwakernaak's model outperformed the unrestricted baseline model for seven out of eight of group S's reliable subjects. In contrast, the model outperformed the unrestricted baseline model for only three out of six reliable subjects in group N. The differential performance of the model, although not significant, may suggest that this model describes well the behavior of experts who are trained in probability. This hypothesis requires further research with a larger sample.

*4.2.5. Zadeh's model (1975) (type* LNNL)

The model outperformed the unrestricted baseline model for nine of 15 reliable subjects (60%) (Table 8). In five out of the nine successes half or more of the vocabulary's clusters had to be eliminated before the model ceased to outperform the restricted baseline model. In another four cases, a small restriction (under 30%) was sufficient. The relatively weak performance of the model across subjects is particularly disappointing, considering that in this type of problem the same component of vagueness was used in both the problem formulation and in the response side. That is, subjects were asked to "add" two linguistic probabilities, and to express the sum as a third one. Hence the required manipulation should be easier than dealing with different kinds of imprecision as was the case in types NNLL and NLNL. Also, more subjects were consistent in this task than in any of the other linguistic tasks. This shows that the task was relatively easy (as measured by consistency) from the subjects' point of view. As with Kwakernaak's model, it is possible that the weak performance is due to problems in data presentation and scaling technique rather than with the model itself.

The model predicts that subjects will respond with the interactive sum of the two fuzzy numbers representing the linguistic terms. Furthermore the model assumes that the linguistic probability assignment list associated with the sample space is $\beta$-interactive (Zadeh, 1975), in the sense that an additional constraint is imposed on the *joint membership function* of the terms (i.e., $p_1 + \cdots + p_n = 1$, in which $p_i$ is a numerical probability associated with the linguistic $\tilde{P}_i$). This interaction is taken into account in the sum operation. This assumption ignores the possibility that, in reality, the interaction between the members of the linguistic probability list might influence the *individual* membership function as well as the joint one.

It has been shown that phrase meanings vary over contexts within an individual (Hersh *et al.*, 1979; Cohen, 1986; Wallsten, Fillenbaum & Cox, 1986*b*). In these studies the following factors were shown to affect the individual membership functions of specific phrases: (1) the nature of the communication task, namely, whether one receives the phrase in communication from another person or selects the phrase in order to communicate to someone else; (2) the available vocabulary (little effect on the meanings of core phrases); (3) event desirability; and (4) base rate effects. It is very likely that an additional context effect is present in this study, namely, that of list composition. Wallsten *et al.* (1986*a*) did not find such an effect in their study (experiment 1), however, the comparison was between groups, and no external restriction was imposed on the lists. In the present study, subjects knew that the linguistic probability assignment list described the probability distribution over mutually exclusive and collectively exhaustive categories, hence, the meanings of individual terms might depend on the other members of the list. Imagine, for example, that you are told that it is *likely* to rain tomorrow and *improbable* not to rain tomorrow. On a different occasion, you are told that it is *likely* to rain tomorrow, but there is a *slight chance* that it will not rain tomorrow. If you understand *improbable* to express a different level of probability than *slight chance*, then it is possible that your perception of the vague meaning of *likely* will differ in these cases. This hypothesis can be experimentally tested. It predicts, for example, that the sum of the same two linguistic probability terms depends on the other members of the assignment list. Unfortunately, we cannot test this prediction in this study, because the sum of two different linguistic terms was required in each problem. This context effect can explain the relatively weak performance of Zadeh's model across subjects.

Based on theoretical grounds Stein (1985) concluded that a joint membership function over the entire linguistic probability distribution must be specified rather than using the marginals as was done in this study. He further introduced a class of joint membership function, called fuzzy beta, to solve this problem. In this work we are interested in the experimental validity of the mathematical models, hence Stein's formulation is useful only if the joint membership function is derived experimentally.

### 4.3. COMPARISON OF MODELS

Since each of the models tested in this study pertains to a different information presentation mode and to a different required response (Table 5), a direct comparison of models is impossible. Nevertheless, it is possible to compare the performance of the models within and across subjects, each in its own setting.

Several trends can be observed: (1) There is a high correlation between accuracy in type NNNN, expressed by the MAD scores, and between reliability in type NNLN. The average MAD score in type NNNN for subjects who were unreliable in type NNLN is 0·086 compared to 0·021 for the reliable subjects. This difference is significant based on the normal approximation to the Wilcoxon rank sum test (Conover, 1980) ($S = 111{\cdot}5$, $z = 2{\cdot}999$, $p < 0{\cdot}003$). The same relationship does not exist with the consistency in the other three problem types (NNLL, NLNL & LNNL). These findings may suggest that statistical sophistication is related to the

manipulation of linguistic information if and only if a numeric response is required. (2) There is no consistent pattern in the performance of the linguistic models (XXXL) within a subject. However, all three models were relatively succesful in predicting the responses of subjects 3, 7, 18 and 19. (3) If more than one of the models was successful within a subject, than the level of success seems to be comparable across models. See for example subjects 3, 9 and 20 for a powerful predictive ability across models, and subjects 7, 13, 18 and 19 for a weaker predictive ability across models. (4) Comparing the overall success rate of the models, across reliable subjects, each in its own setting, reveals that Yager's model is the most successful (91·7%), Kwakernaak's model is next (71·4%) and Zadeh's model (1975) is the least successful (60%). However, a different picture emerges when comparing the strength of the models within subjects. On average about half of the vocabulary clusters had to be eliminated before Yager's (1979; 1984*b*) and Zadeh's (1975) models ceased to outperform the restricted baseline model (48·72% and 50·6% respectively), compared to an average of 37·56% with regard to Kwakernaak's model. These findings indicate that Kwakernaak's model outperformed the unrestricted baseline model in more cases than did Zadeh's model (1975), however among these cases Zadeh's model (1975) was more powerful. (5) Zadeh's model (1968) clearly failed to capture the psychological processes of numerically estimating the probability of a fuzzy event given a crisp random variable.

It is important to note that the above discussion with regard to the performance of the models ignores the possibility that the experimental setting was in some cases at fault, rather than the models themselves. We have discussed these possibilities with regard to Kwakernaak's (1978) and Zadeh's (1975) models in previous sections. Any final conclusion should await further research in which this possible pitfall is eliminated.

Models may have more to offer than the success of their predictions on a particular body of data. One cares about their ability to suggest extensions that add to predictive power and to provide insights into the underlying psychological processes and into behaviour outside the laboratory. The remainder of this section discusses these points.

At the lowest level, the study is conducive to minor changes that might affect the quality of the predictions. Some of these were mentioned in this paper and others have been examined but not considered here (Zwick, 1987*b*). An example is an alternative scaling procedure, or using experts as subjects.

The relative success of the linguistic models and the failure of the numeric one (Zadeh, 1968) supports the common underlying feature of the linguistic models. These models assume that subjects are unable to cope with the overall structure of the problem and instead reformulate it to reduce its complexity by considering multiple-crisp representations of the problem. The solution at each crisp level of representation is easy, since it corresponds to a simple probability problem. The final response is the combined solutions from the surrogate crisp levels of representations.

Finally, one cares about the extent to which the models form useful inputs for other theoretical structures such as to decision analysis. We are planning to extend the current research in this direction.

This research was supported by Contract MDA 903-83-K-0347 from the US 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.

## References

ADAMO, J. M. (1980). Fuzzy decision trees. *Fuzzy Sets and Systems*, **4**, 207–219.

BECKER, S. W. & BROWNSON, F. O. (1964). What price ambiguity? Or the role of ambiguity in decision making. *Journal of Political Economy*, **72**, 62–73.

BEHN, R. D. & VAUPEL, J. W. (1982). *Quick Analysis for Busy Decision Makers*. New York: Basic Books, Inc.

BELLMAN, R. E. & ZADEH, L. A. (1970). Decision-making in a fuzzy environment. *Management Science*, **17**, B141–B164.

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

BLACK, M. (1937). Vagueness. *Philosophy of Science*, **4**, 427–455.

BRYANT, G. D. & NORMAN, G. R. (1980). Expressions of probability : Words and numbers (letter). *New England Journal of Medicine*, **302**, 411.

BUDESCU, D. V. & WALLSTEN, T. S. (1987). Subjective estimation of precise and vague uncertainties. In G. WRIGHT & P. AYTON, Eds. *Judgmental Forecasting*. pp. 63–82. New York: John Wiley & Sons.

BUDESCU, D. V., WEINBERG, S. & WALLSTEN, T. S. (1988). Decisions Based on Numerically and Verbally Expressed Uncertainties. *Journal of Experimental Psychology: Human Perception and Performance*, **14**, 281–294.

Buoncristiani, J. F. (1980). *Probability on Fuzzy Sets*. Unpublished PhD dissertation. Boston University.

Buoncristiani, J. F. (1983). Probability on fuzzy sets and *E*-fuzzy sets. *Journal of Mathematical Analysis and Applications*, **96**, 24–41.

CHANG, S. & ZADEH, L. A. (1972). On fuzzy mapping and control. *IEEE Transactions, SMC 2*, 30–34.

COHEN, B. (1986). *Effects of Independent Outcome Desirability on the Meanings of Probability Phrases*. Unpublished M.A. Thesis, University of North Carolina at Chapel Hill.

CONOVER, W. J. (1980). *Practical Nonparametric Statistics*, second edition, New York: John Wiley & Sons.

DE FINETTI, B. (1977). Probabilities of probabilities: A real problem or a misunderstanding? In A. AYKAC & C. BRUMAT, Eds. *New Directions in the Application of Bayesian Methods*. Amsterdam: Elsevier.

DUBOIS, D. & PRADE, H. (1978*a*). *Fuzzy Algebra, Analysis, Logics*. Technical Report TR-EE 78/13. Purdue University, Lafayette, Indiana.

DUBOIS, D. & PRADE, H. (1978*b*). Operations on fuzzy numbers. *Systems Science*, **9**, 613–626.

DUBOIS, D. & PRADE, H. (1981). Additions of interactive fuzzy numbers. *IEEE Transactions on Automatic Control*, **26**, 926–936.

DUBOIS, D. & PRADE, H. (1982*a*). Towards fuzzy differential calculus. Part 1: Integration of fuzzy mappings. *Fuzzy Sets and Systems*, **8**, 1–17.

DUBOIS, D. & PRADE, H. (1982*b*). Towards fuzzy differential calculus. Part 2: Integration on fuzzy intervals. *Fuzzy Sets and Systems*, **8**, 104–116.

DUBOIS, D. & PRADE, H. (1982*c*). Toward fuzzy differential calculus. Part 3: Differentiation. *Fuzzy Sets and Systems*, **8**, 225–233.

DUTTA, A. (1985). Reasoning with imprecise knowledge in expert systems. *Information Sciences*, **37**, 3–24.

ELSBERG, D. (1961). Risk, ambiguity, and the savage axioms. *Quarterly Journal of Economics*, **75**, 643–669.

FOX, J., BARBER, D. C. & BARDHAN, K. D. (1980). Alternative to Bayes? A quantitative comparison with rule-based diagnostic inference. *Method of Information in Medicine*, **19**, 210–215.

GAINES, B. R. (1978). Fuzzy and probability uncertainty logics. *Information and Control*, **38**, 154–169.

GARDENFORS, P. & SAHLIN, N. E. (1982). Unreliable probabilities, risk taking, and decision making. *Synthese*, **53**, 361–386.

GILES, R. (1983). The practical interpretation of fuzzy concepts. *Human Systems Management*, **3**, 263–264.

GOGUEN, J. A. (1968–69). The logic of inexact concepts. *Synthese*, **19**, 325–373.

HERSH, H. M., CARMAZZA, A. & BROWNWELL, 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: Elsevier.

KAHNEMAN, D. & TVERSKY, A. (1982). Variants of uncertainty. *Cognition*, **11**, 143–157.

KENT, S. (1964). Words of estimated probability. *Studies in Intelligence*, **8**, 49–65.

KHALILI, S. (1979). Independent fuzzy events. *Journal of Mathematical Analysis and Applications*, **67**, 412–420.

KLEMENT, E. P. (1982). Some remarks on a paper by R. R. YAGER. *Information Sciences*, **27**, 211–220.

KLEMENT, E. P. & SCHWYHLA, W. (1981). Fuzzy probability measures. *Fuzzy Sets and Systems*, **5**, 21–30.

KOCHEN, M. (1975). Applications of fuzzy sets in psychology. In L. A. ZADEH, K. S. FU, K. TANAKA & M. SHIMURA, Eds. *Fuzzy Sets and Their Applications to Cognitive and Decision Processes*. pp. 395–408. London: Academic Press.

KOCHEN, M. (1979). Enhancement of coping through blurring. *Fuzzy Sets and Systems*, **2**, 37–52.

KRUSE, R. (1982). The strong law of large numbers of fuzzy random variables. *Information Sciences*, **28**, 233–241.

KRUSE, R. (1984). Statistical estimation with linguistic data. *Information Sciences*, **33**, 197–207.

KWAKERNAAK, H. (1978). Fuzzy random variables, I: Definitions. *Information Sciences*, **15**, 1–29.

KWAKERNAAK, H. (1979). Fuzzy random variables, II: Algorithms and examples for the discrete case. *Information Sciences*, **17**, 253–278.

LAKOFF, G. (1973). Hedges: a study in meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Logic*, **2**, 458–508.

LUCHINS, A. S. (1942). Mechanization in problem solving—The effect of Einstellung. *Psychological Monographs*, **54**, 1–27.

MIYAKOSHI, M. & SHIMBO, M. (1984). A strong law of large numbers for fuzzy random variables. *Fuzzy Sets and Systems*, **12**, 133–142.

NAGY, T. J. & HOFFMAN, L. J. (1981). *Exploratory evaluation of the accuracy of linguistic versus numeric risk assessment of computer security* (Technical Report GWU-IIST-81-07). Computer Security Research Group. The George Washington University.

NAHMIAS, S. (1978). Fuzzy variables. *Fuzzy Sets and Systems*, **1**, 97–110.

NAHMIAS, S. (1979). Fuzzy variables in a random environment. In M. M. GUPTA, R. K. RAGADE & R. R. YAGER, Eds. *Advances in Fuzzy Set Theory and Applications*. pp. 165–180. Amsterdam: Elsevier.

NAKAO, M. A. & AXELROD, S. (1983). Numbers are better than words. *American Journal of Medicine*, **74**, 1061–1065.

NATIONAL RESEARCH COUNCIL GOVERNING BOARD COMMITTEE ON THE ASSESSMENT OF RISK (1981). *The Handling of Risk Assessments in FRC Reports*. Washington, DC: US National Research Council.

NEGOITA, C. V. & RALESCU, D. A. (1975). *Applications of Fuzzy Sets to Systems Analysis*. New York: John Wiley & Sons.

NGUYEN, H. T. (1977). On fuzziness and linguistic probabilities. *Journal of Mathematical Analysis and Applications*, **61**, 658–671.

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

PIPINO, L. L., VAN GIGCH, J. P. & TOM, G. (1981). Experiments in the representation and manipulation of labels of fuzzy sets. *Behavioral Science*, **26**, 216–228.

PURI, M. L. & RALESCU, D. A. (1989). Fuzzy random variables. *Journal of Mathematical Analysis and Applications*. (In press).

RALESCU, A. L. & RALESCU, D. A. (1984). Probability and fuzziness. *Information Science*, **34**, 85–92.

RAO, C. R. (1965). *Linear Statistical Inference and its Applications*. New York: John Wiley & Sons.

RAPOPORT, A., WALLSTEN, T. S. & COX, J. A. (1987). Direct and indirect scaling of membership functions of probability phrases. *Mathematical Modelling*, **9**, 397–417.

SAVAGE, L. J. (1954). *The Foundations of Statistics*. New York: Wiley.

SHAFER, G. & TVERSKY, A. (1985). Languages and designs for probability judgment. *Cognitive Science*, **9**, 309–339.

SLOVIC, P., FISCHHOFF, B. & LICHTENSTEIN, S. (1977). Behavioral decision theory. *Annual Review of Psychology*, **28**, 1–39.

SMETS, P. (1982*a*). Subjective probability and fuzzy measures. In M. M. GUPTA & E. SANCHEZ, Eds. *Fuzzy Information and Decision Processes*. pp. 87–91. Amsterdam: Elsevier.

SMETS, P. (1982*b*). Probability of a fuzzy event: an axiomatic approach. *Fuzzy Sets and Systems*, **7**, 153–164.

STEIN, W. E. (1985). Fuzzy probability vectors. *Fuzzy Sets and Systems*, **15**, 263–267.

STEIN, W. E. & TALATI, K. (1981). Convex fuzzy variables. *Fuzzy Sets and Systems*, **6**, 271–283.

SZOLOVITS, P. & PAUKER, S. G. (1978). Categorial and probabilisitc reasoning in medical diagnosis. *Artificial Intelligence*, **11**, 115–144.

WALLSTEN, T. S., BUDESCU, D. V., RAPOPORT, A., ZWICK, R. & FORSYTH, B. (1986*a*). Measuring the vague meanings of probability terms. *Journal of Experimental Psychological: General*, **115**, 348–365.

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

WYDEN, P. (1979). *Bay of Pigs*. New York: Simon & Schuster.

YAGER, R. R. (1979). A note on probabilities of fuzzy events. *Information Sciences*, **18**, 113–129.

YAGER, R. R. (1982). Generalized probabilities of fuzzy events from belief structures. *Information Sciences*, **28**, 45–62.

YAGER, R. R. (1984*a*). Probabilities from fuzzy observations. *Information Sciences*, **32**, 1–31.

YAGER, R. R. (1984*b*). A representation of the probability of a fuzzy subset. *Fuzzy Sets and Systems*, **13**, 273–284.

YATES, J. F. & ZUKOWSKI, L. G. (1976). Characterization of ambiguity in decision making. *Behavioral Science*, **21**, 19–25.

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

ZADEH, L. A. (1968). Probability measures of fuzzy events. *Journal of Mathematical Analysis and Applications*, **23**, 421–427.

ZADEH, L. A. (1975). The concept of linguistic variable and its application to approximate reasoning. Parts 1, 2 and 3. *Information Science*, **8**, 199–249; **8**, 301–357; **9**, 43–96.

ZADEH, L. A. (1981). Fuzzy probabilities and their role in decision analysis. *Proceedings of the Fourth MI/ONR Workshop on Command, Control and Communications*. pp. 159–179. MIT.

ZADEH, L. A. (1984). Fuzzy probabilities. *Information Processing and Management*, **20**, 363–372.

ZADEH, L. A. (1986). Is probability theory sufficient for dealing with uncertainty in AI: A negative view. In L. N. KANAL & J. F. LEMMER, Eds. *Uncertainty in Artificial Intelligence*. North Holland: Elsevier.

ZIMMER, A. C. (1983). Verbal versus numerical processing of subjective probabilities. In R. W. SCHOLZ, Ed. *Decision Making Under Uncertainty*. pp. 159–182. North Holland: Elsevier.

ZIMMERMANN, H.-J. (1987). *Fuzzy Sets, Decision Making, and Expert Systems*. Boston: Kluwer Academic.

ZWICK, R. (1987*a*). A note on random sets and the Thurstonian scaling methods. *Fuzzy Sets and Systems*, **21**, 351–356.

ZWICK, R. (1987*b*). Combining stochastic uncertainty and linguistic inexactness: Theory and experimental evaluation. Unpublished doctoral dissertation. University of North Carolina at Chapel Hill.

ZWICK, R. (1988). The Evaluation of Verbal Models. *International Journal of Man–Machine Studies*, **29**, 149–157.

ZWICK, R., BUDESCU, D. V. & WALLSTEN, T. S. (1988). An empirical study of the integration of linguistic probabilities. In T. Zétényi, Ed. *Fuzzy Sets in Psychology*. pp. 91–125. North Holland: Elsevier.

ZWICK, R., CARLSTEIN, E. & BUDESCU, D. V. (1987). Measures of similarity among fuzzy concepts: A comparative analysis. *International Journal of Approximate Reasoning*, **1**, 221–242.

ZYSNO, P. (1981). Modeling membership functions. In B. B. RIEGER, Ed. *Empirical Semantics*. pp. 350–375. Bochum, FRG: Rockmeyer.