PROCESSING LINGUISTIC PROBABILITIES: GENERAL PRINCIPLES AND EMPIRICAL EVIDENCE

David V. Budescu and Thomas S. Wallsten

I. Overview

How do people use and understand linguistic expressions of probability? Is information processing, choice behavior, or decision quality more optimal in any well-defined sense when subjective uncertainty is expressed linguistically or numerically? Do people communicate with each other better in one modality or another? These and related questions are of considerable practical and theoretical importance. Practical issues arise because weighty decisions often depend on forecasts and opinions communicated from one person or set of individuals to another. Examples of decisions that depended on the communication of expert judgment include the Bay of Pigs invasion (Wyden, 1979, pp. 89–90), safety assessments regarding components of the space shuttle (Marshall, 1986, 1988) or of nuclear power plants (Vesely & Rasmusen, 1984), or simply one's own selection of investments or medical treatments. The standard wisdom (e.g., Behn & Vaupel, 1982; Moore, 1977; von Winterfeldt & Edwards, 1986) has been that numerical communication is better than linguistic, and therefore, especially in important contexts, it is to be preferred. But a good deal of evidence suggests that (1) this advice is not uniformly correct, and (2) it is inconsistent with strongly held preferences. A theoretical understanding of the preceding questions is an important step toward the development of means for improving communication, judgment, and decision making under uncertainty.

The theoretical issues concern how individuals interpret imprecise linguistic terms, what factors affect their interpretations, and how they combine those terms with other information (itself vague or imprecise to some degree) for the purpose of taking action. The action may be a discrete choice, an evaluation of options, or the communication of one's own opinion to other people. In this chapter, we review the relevant literature in order to develop a theory of how linguistic information about imprecise continuous quantities is processed in the service of decision making, judgment, and communication. We restrict ourselves almost entirely to research on qualitative expressions of subjective uncertainty, where we have done most of our work on the topic. However, without much effort, the theoretical ideas can be extended to and tested in other domains. The theory that we will present has evolved over the years as our research has progressed, and pieces of it in one form or another can be found in our and our colleagues' publications. We take advantage of this chapter to present our current view, which has evolved inductively, to substantiate it where the data allow, and to suggest where additional research is needed.

The relevant literature is a rich one; indeed, it dates back at least half a century to Simpson's (1944) paper on relative frequency terms. The literature has moved in many directions, but to keep this chapter focused, we do not review research on frequency, quantity, or amount terms and their use in questionnaires and surveys (see Bass, Cascio, & O'Connor, 1974; Bradburn & Miles, 1979; Hammerton, 1976; Mosier, 1941; Newstead, 1988; and Pepper, 1981, for some key representative examples). Nor do we discuss the empirical literature comparing verbal and numerical representations of other (i.e., nonprobability) attributes of options in decision situations (e.g., Schkade & Kleinmuntz, 1994; Stone & Schkade, 1991; Svenson & Karlsson, 1986).

In the next section, we set the stage by presenting a useful three-way taxonomy of sources of vagueness. This typology leads naturally to a pair of background assumptions that underlie our review. The third section then summarizes the research on meanings of qualitative probability expressions, and the fourth section compares judgments and decisions made on the basis of vague and precise (generally linguistic and numerical) probabilities. Each section consists of a review of the empirical regularities that have emerged, followed by a theoretical statement in the form of general principles that can explain the regularities, and, when available, a summary of research supporting the principles. The final section brings the background assumptions and the empirical principles together into a unified theoretical statement and provides some concluding remarks.

II. Setting the Stage

A. Stimulus and Task Constraints

Any theory of how people process information about uncertain events must be constrained by stimulus and task considerations. The former refers to the type of event in question, the nature of the uncertainty about it, and the manner in which that uncertainty is represented. The latter concerns the purpose for which the information is being processed. Indeed, specifying those constraints goes a long way toward developing the structure of the theory. Therefore we must consider stimulus and task characteristics before laying out our theoretical principles.

Zwick and Wallsten (1989) proposed and Wallsten (1990) somewhat modified a useful three-way taxonomy according to the nature of the event, the uncertainty about that event, and the manner in which the uncertainty is represented. Each of the three ways is a continuum that ranges from absolute precision at one end to absolute vagueness at the other. Considering events first, the distinction between those that are precise and those that are vague is straightforward. For example, the event of *noontime temperature exceeding 75°F* is precise, whereas the event of *a warm noontime* is vague. More generally, an event is precise if it is defined such that any outcome in the universe of discourse either is or is not an exemplar of it (for any day, the noontime temperature either does or does not exceed 75°F). All other events are vague to some degree. In set-theory terms, an event is precise if all outcomes in the universal set unequivocally have membership of either 0 or 1 in the subset that is defined by that event. If membership of 0 or 1 cannot be definitively assigned to each outcome, then the event is vague to some extent. For example, the event of a warm noontime is somewhat but not absolutely vague, because some temperatures, say below 40°F and above 110°F, are definitely outside the concept *warm*. It is vague to some degree because individuals will consider some temperatures between those two extremes more worthy of the designation *warm* than others. (Murphy & Brown, 1983, and Murphy, Lichtenstein, Fischhoff, & Winkler, 1980, provide additional illustrations of the distinction between vague and precise events.)

The notion that uncertainty varies from vague to precise is more subtle and less easily described. Generally, uncertainty is precise if it depends on external, quantified random variation; and it is vague if it depends on internal sources related to lack of knowledge or to judgments about the nature of the database. But regardless of the source of the uncertainty, we say that it is precise if the judge can place the event in question within a total likelihood ordering of a sufficiently rich set of events. Specifically, consider the set of canonical events $\{E_0, E_1, \ldots, E_i, \ldots, E_{N-1}, E_N\}$, defined as the radial sectors of a perfectly balanced spinner with relative areas of i/N for any positive integer N. E_0 is the null event, E_N is the universal event, and the events are ordered such that

$$E_0 < E_1 < \ldots < E_i < \ldots < E_{N-1} < E_N, \tag{1}$$

where < means "is less likely than."[1] The uncertainty of event A can be considered absolutely precise if for any N there exists two successive events, E_i and E_{i+1}, such that the judge agrees that

$$E_i < A < E_{i+1}. \tag{2}$$

For virtually any event A in a real situation, there will exist a sufficiently large N that an individual will not feel comfortable judging Eq. 2 to hold for any E_i. In that sense, subjective uncertainty is rarely, if ever, absolutely precise; but in a practical sense, people often will judge Eq. 2 to hold for reasonably large values of N and we would deem the event uncertainty to be precise. Thus, defining A as the event of a coin landing heads up, one might, after a sufficient number of test flips, be willing to state for $N = 99$ that

$$E_{49} < A < E_{50}.$$

The subjective uncertainty in that case would be very precise.

More often, the uncertainty is relatively vague. For example, defining the event A as noontime temperature exceeding 75°F, a weather forecaster might be unwilling to endorse Eq. 2 with any E_i for N as small as 10, but would be willing to assert for $N = 5$ that

$$E_3 < A < E_4. \tag{3}$$

Our goal here is not to develop an index of vagueness, but rather to develop the important intuition that the degree to which a judge can place events in a connected, transitive likelihood ordering varies as a function of the nature and source of the supporting information. Note that this conception of subjective uncertainty differs markedly from a common viewpoint that vague uncertainties (generally misleadingly called ambiguous

[1] Canonical events of this sort, with the assumption that $P(E_i) = i/N$, are included in some axiom systems leading to subjective probability representations (e.g., DeGroot, 1970, chap. 6). We are concerned here with qualitative likelihood judgments, not with numerical representations, but want to convey the idea of precise events that would be judged equally spaced in likelihood.

probabilities) can be represented by second-order probability distributions, that is, precise probabilities over all possible probability distributions (see, e.g., the review by Camerer & Weber, 1992). As Gärdenfors and Sahlin (1982) have suggested, there are psychological and epistemological differences between unspecified, or unspecifiable, probabilities and those that are given in the form of a distribution. The former are vague to some degree and the latter are precise. The distinction matters for some purposes and not for others, but to represent both types of uncertainty by second-order probability distributions is to miss the problems for which it is important.

Finally, representations of uncertainty vary from precise to vague. Probability theory provides a language for precise representation. Imprecise portrayals include numerical intervals, quantified numbers (e.g., *approximately .6*), and linguistic probabilities of the sort we are considering in this chapter. An advantage of numerical intervals is that they signal both location and degree of imprecision. Verbal expressions do not have this benefit, but they compensate by conveying greater nuances of meaning. Their rich semantic structure allows one to convey not only approximate location and degree of imprecision, but also relative weights over levels of uncertainty within an implied range, and perhaps also other aspects of the communicator's knowledge or opinions beyond degrees of uncertainty.

The three continua along which stimulus vagueness can vary—event type, uncertainty type, and representation type—are distinct but not fully independent (Budescu & Wallsten, 1987). That is, a representation can be no more precise than the underlying uncertainty, which in turn can be no more precise that the event in question. The converse, however, does not hold: The uncertainty of an event can be vague to any degree, as can the representation of that uncertainty. Thus it would be perfectly natural for a television weather forecaster to (1) consider the precise event that the temperature at noon will exceed 75°F, (2) judge its uncertainty as vaguely as indicated by Eq. (3), and (3) say, "There is a good chance that the noontime reading will be above 75°F." On the other hand, it would be very strange for him or her to say, "There is a 70% chance that the temperature at noon will be warm," because a precise probability of a vague event is meaningless.

Tasks are not so easily taxonomized. They may involve making choices, rating or ranking alternatives, forming judgments for later use or for communication to other people, or any of a myriad of other possibilities. We should be mindful, however, that the manner in which the task is carried out depends on the nature of the information, and, correspondingly, the way in which the information is processed depends on the purpose to which it is being put.

B. Background Assumptions

Our goal is to explain how humans process vague, especially linguistic, information about uncertainty and how they combine it with and trade it off against information about other stimulus dimensions. The theory consists of two background assumptions and five principles. Some of the principles are no more than restatements of robust empirical regularities in theoretical terms, whereas others have further testable consequences. They are presented at the conclusion of the relevant review sections, along with supporting data where they are available. Here, we focus on the two background assumptions. Both are falsifiable, but they strike us as reasonable and we are treating them without evidence as true. These assumptions have, however, testable corollaries and we will summarize the relevant results.

1. Background Assumption B1

Except in very special cases all representations are vague to some degree in the minds of the originators and in the minds of the receivers. This assumption builds the logical constraints previously discussed into our theory of how humans process information about uncertain events. It implies that to the extent possible people consider the event definition and the data base both when deciding how to represent their own uncertain judgment and when interpreting a representation that they have received from someone else.

2. Background Assumption B2

People use the full representation whenever feasible, but they narrow it, possibly to a single point, if the task requires them to do so. This assumption expresses the idea that the task determines whether and how an individual resolves representational vagueness. Thus, for example, people treat separate judgments in their vague form when receiving and combining them, but they restrict attention to a narrow range of uncertainty or to a single point value when making specific decisions. Put crudely, one can have imprecise opinions, but one cannot take imprecise actions.

III. Meanings of Qualitative Probability Expressions

As already mentioned, probability phrases have rich semantic structures and it is likely that we use them to communicate more information than simply an approximate location on a [0,1] scale (Moxey & Sanford, 1993; Teigen & Brun, 1993). Nevertheless, most behavior studies of probability terms have focused on their numerical referents; and on this topic, there is a voluminous literature. The representations investigated vary from a single number (e.g., Beyth-Marom, 1982), through a range of numerical values (e.g., Hamm, 1991), to functions representing acceptability (Mosteller & Youtz, 1990; Reagan, Mosteller, & Youtz, 1989) or membership (Wallsten, Budescu, Rapoport, Zwick, & Forsyth, 1986) over the [0,1] interval. Among the elicitation methods used are words-to-numbers translation (e.g., Lichtenstein & Newman, 1967), numbers-to-words conversion (e.g., Reagan et al., 1989), comparison of the appropriateness of various numbers to describe a word and of various words to describe a given numerical value (Wallsten, Budescu, Rapoport, Zwick, & Forsyth, 1986), and elicitation of numerical and verbal responses to identical events (Budescu, Weinberg, & Wallsten, 1988). We review the empirical literature in four subsections, each focusing on distinct issues related to communication with these terms: intra-individual vagueness of probability terms, interindividual variance in the understanding and use of probability phrases, intra-individual sensitivity to context, and preferences for various modes of communicating degrees of uncertainty.

A. Intra-individual Vagueness of Probability Terms

One important aspect of our work in this domain was to establish that a probability term within a context can be modeled by means of membership functions over numerical values.[2] We think of a probability phrase as a vague or fuzzy concept whose members are the numerical probabilities in the [0,1] interval. A membership function assigns to each numerical probability a real number, which we refer to as its *membership* in the concept defined by the phrase (in that context). In principle, these membership values are ratio scaled and range from 0, for probabilities that are absolutely not included in the concept, to a maximum value, arbitrarily fixed at 1, for probabilities that are ideal or perfect exemplars of the concept being considered; intermediate values represent intermediate degrees of membership. There are no special constraints on the shape, or any other property, of these functions. In particular, a membership function is not a

[2] It should be clear that phrases are always considered within a context of some sort, even if that context is no more than the single phrase in isolation or within a list, as has sometimes been the case in research. A simple list or individual phrase may be as close to a "null" context as one can get, or it may be an invitation for subjects to provide their own idiosyncratic frame of reference, as Moxey and Sanford (1993) suggest, but it is nevertheless a context. Thus, when we speak of representing the meaning of a probability phrase by a membership function, we are always referring to its meaning in a particular context regardless of whether or not we explicitly say so. Unfortunately, we have not been sufficiently careful to stress this point in the past. Indeed, as will become apparent, an important question is whether membership functions of phrases change systematically across contexts.

density function. It need not be continuous and the area under it need not integrate to 1. For example, precise (crisp) terms are characterized by a membership function with only two values (1 for all probabilities that are ideal exemplars of the concept, and 0 for all other values).

The concept of a membership function provides a useful generalization of simpler, perhaps more intuitive representations of linguistic uncertainties, such as a best probability and a range of probabilities. Presumably, the best point representation of a term characterized by a membership function is some central measure of that function. Of the various measures possible, a natural choice is the probability with the highest membership (or if the value is not unique, a summary, such as the mean of all of the probabilities with maximal membership). Another measure, analogous to the mean, was proposed by Yager (1981) as

$$\mathrm{W}_e = \frac{\int_0^1 \mu_e(p)\, p\, dp}{\int_0^1 \mu_e(p)\, dp}, \tag{4}$$

where $\mu_e(p)$ represents the membership value of p in expression e. W_e is simply an average of the probabilities in the [0,1] interval, weighted by their normalized (forced to sum to unity) membership values. A discrete version of Eq. (4) is straightforward. Other location measures have been proposed in the fuzzy sets literature (see Bortolan & Degani, 1985, for a review), and an interesting empirical question is which, if any, of these measures is the best.

The range of acceptable probabilities for a term within a particular context, usually called the support of the membership function, consists of all of the values with positive membership. A convenient measure of spread, which we have used occasionally, is analogous to the variance of a density. Making use of W_e in Eq. (4), this measure is

$$\mathrm{V}_e^2 = \frac{\int_0^1 \mu_e(p)\, (p - \mathrm{W}_e)^2\, dp}{\int_0^1 \mu_e(p)\, dp}. \tag{5}$$

Finally, membership functions implicitly define subsets of probabilities that are members of the concept implied by the phrase to specified degrees. This definition is accomplished by restricting subset membership to probabilities with membership values above some threshold, v, where $0 < v < 1$. Subsets monotonically decrease (become narrower) as v increases. The notion of thresholds provides a convenient quantitative way to theorize about probabilities that are "sufficiently well described by a phrase," and we will use it subsequently. Figure 1 presents hypothetical examples of some membership functions of verbal expressions.

Wallsten, Budescu, Rapoport, Zwick, and Forsyth (1986) and Rapoport, Wallsten, and Cox (1987) have shown that membership functions of the sort shown in Fig. 1 can be empirically derived and validated at the individual subject level. For each phrase used in these studies, the subjects were asked to compare various pairs of probabilities (represented by spinners) and to indicate which member of the pair better represents the meaning of the given term and how much better it does so. The graded pair-comparison judgments were reliable over replications. Conjoint measurement techniques showed that the functions satisfied the ordinal properties of a difference or a ratio representation (see also Norwich & Turksen, 1984), and goodness of fit measures indicated that both metrics scaled the judgments equally well. Rapoport et al. (1987) showed that the scales derived from the pair-comparison judgments can be approximated well by simpler and quicker direct rating techniques, which we have applied successfully in

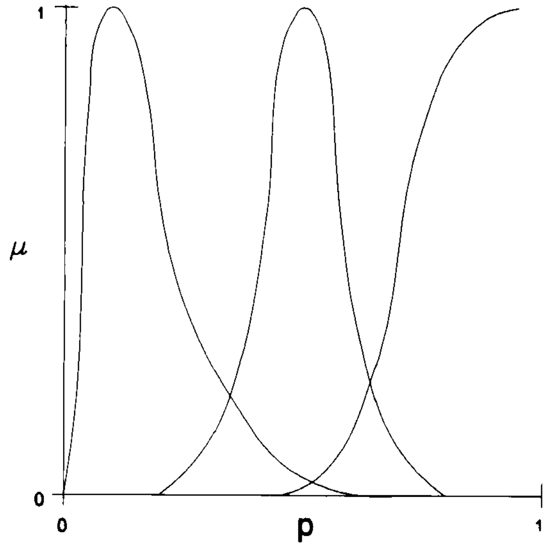

Fig. 1. Generic membership functions for three phrases.

several subsequent papers (e.g., Budescu & Wallsten, 1990; Fillenbaum, Wallsten, Cohen, & Cox, 1991; Jaffe-Katz, Budescu, & Wallsten, 1989; Tsao & Wallsten, 1994; Wallsten, Budescu, & Zwick, 1993).

Most derived membership functions are single-peaked and a sizeable minority are monotonic, decreasing from probabilities close to 0 for low terms and increasing to probabilities close to 1 for high terms. (In many cases, these monotonic functions may actually be single-peaked, but just appear to be monotonic because we failed to include probabilities sufficiently close to the end points. See also Reagan et al.'s, 1989, footnote on this topic.) Single-peaked functions may be considerably skewed in one direction or another. Most functions cover a relatively large range of values, indicating that the terms are vague to individuals. Calculations of the spread measure, V_e^2 in Eq. (2), confirms this impression (e.g., Fillenbaum et al., 1991; Tsao & Wallsten, 1994). Only a small minority of the functions can be classified as relatively "crisp" (i.e., having a narrow support with uniformly high membership). Relatively crisp functions generally represent terms such as *toss-up* or *even odds,* which tend to convey specific values.

It is important to realize that because membership functions are derived from temporally stable judgments at the level of individual subjects, they indicate that phrases are *intra-individually vague.* Reagan et al. (1989) present similarly appearing functions, but they are relative frequency distributions of acceptable or best probability judgments aggregated over individuals. There are two problems in interpreting the Reagan et al. functions. First, as Rubin (1979) pointed out in a different domain of vague terms, such response distributions are just as easily interpreted in terms of error variance as in terms of vagueness or fuzziness. Second, regardless of whether the functions represent noise or vagueness, they relate to inter- not intra-individual differences in translating phrases to numbers.

Returning to single-subject data, other procedures, which require fewer judgments than needed to establish membership functions and therefore are simpler for respondents, also indicate that terms are vague to individuals. The simplest procedure is just to ask for a range—a lower and an upper probability—that the phrase represents. Wallsten, Budescu, Rapoport, Zwick, and Forsyth (1986) posed that question prior to eliciting membership functions. The results are summarized in Fig. 2, reproduced from their article. Median within-subject ranges were substantial, varying from approximately .1 for the terms *almost impossible* and *almost certain* to more than .5 for *possible.* Hamm (1991) also asked subjects to provide lower and upper bounds for 19 terms. The median range (taken across 65 subjects) for 14 of these terms was greater than .10, and it was 0 only for the anchor terms *absolutely impossible, toss-up,* and *absolutely certain.*

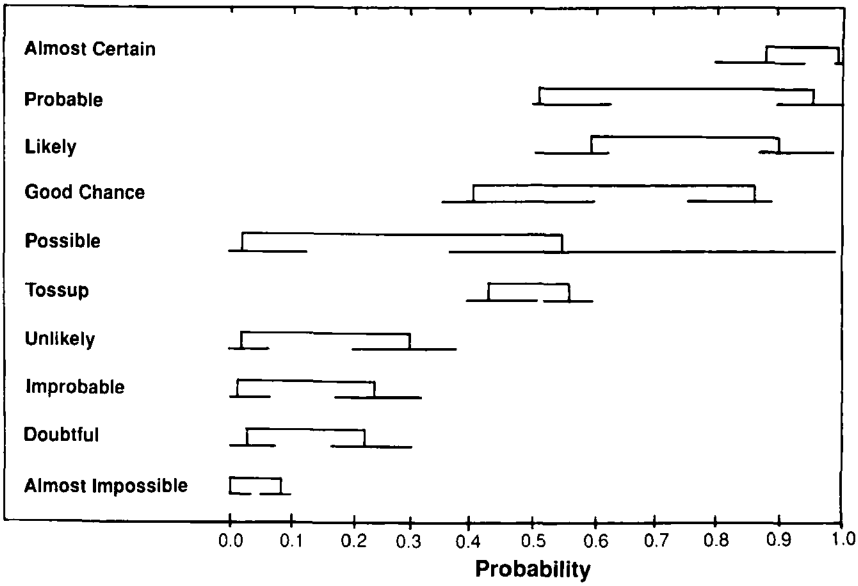

Fig. 2. First, second, and third quartiles over subjects of the upper and lower probability limits for each phrase in Experiment 1 of Wallsten, Budescu, Rapoport, Zwick, and Forsyth (1986).

If probabilities belong in varying degrees to the concept defined by a phrase, one might expect subjects to give differing probability ranges for a particular phrase, depending on the degree of acceptability they deem necessary to include the probability in its concept. In membership function terms, subjects respond to the task of giving a lower and an upper probability by setting an acceptability level, v, and reporting the lower and upper probabilities, p_* and p^*, respectively, for which

$$\mu\,(p_*) = \mu\,(p^*) = v, \tag{6}$$

assuming single-peaked functions. In the case of truly monotonic functions, one of the limits would be either 0 or 1. Presumably, v would be sensitive to instructions, payoffs, or other factors, and would be constant over all phrases within a particular context. To our knowledge, this interesting prediction has not been checked. However, it would explain why when Weber and Hilton (1990) and Wallsten, Fillenbaum, and Cox (1986, Experiment 1) gave their subjects the opportunity to use a range of values, rather than a single number, only a few subjects chose to do so. Perhaps, because they were simply allowed but not required to give a range, they eased their task as much as possible by setting $\upsilon = 1$.

Other data relevant to intra-individual vagueness come from studies in which subjects are asked for point numerical translations on more than one occasion. As already indicated, the problem with such data from our perspective is that one cannot determine the degree to which nonperfect replicability reflects error variance rather than vagueness. Following our membership function theme, if people set $\upsilon = 1$ for purposes of providing point estimates, then any differences in replication are the result of random error. In fact, standard words-to-numbers translations within a fixed (or in the absence of any) context appear to yield relatively narrow (although clearly greater than 0) ranges of values within each individual (e.g., Beyth-Marom, 1982; Clarke, Ruffin, Hill, & Beamen, 1992). Similar results, but with other summary measures, have been reported by Bryant and Norman (1980), Budescu and Wallsten (1985), Johnson (1973), Mullet and Rivet (1991), and Sutherland et al. (1991).

Yet another was to assess the intra-individual vagueness of a phrase is to determine the range of probabilities for when an individual uses it. In the first stage of the experiment of Budescu et al. (1988), 20 subjects judged probabilities of 11 spinners six times. The procedure provided subjects with the opportunity to invoke the same word to describe different displays, and provided us with a convenient way to quantify the considerable within-subject variability of probabilities which can be described by any given term. On the average this variance was larger than in the case of numerical judgments, and in approximately one-half of the cases, the within-subject variance exceeded the between-subject component.

We have assumed (assumption B1) that all representations are vague to some degree. A somewhat counterintuitive corollary of this statement is that the scale meanings of numbers should also show imprecision. This prediction has been sustained by at least two studies, which have established that the meanings of numbers are both imprecise and subject to context effects. We concentrate here only on the imprecision and consider context effects later. Mullet and Rivet (1991) had children and adolescents rate the meanings of various sentences (in French) that described the chances of certain children passing or failing in school the next year. The sentences included both numerical (e.g., *one-in-four chance*) and verbal (e.g., *doubtful*) phrases. Ratings were made on a continuous (unnumbered) response scale. Between- and within-subject variances in the ratings of each sentence were not substantially different in the verbal and numerical cases. Shapiro and Wallsten (1994) had subjects rate verbal and numerical probabilities both in isolation and in the context of forecasts. On an unnumbered response scale, they provided a best, a highest, and a lowest rating in each case. The measured difference between the highest and lowest ratings is an index of vagueness. Numbers were less vague than words, but not completely precise. Finally, Budescu et al. (1988) also found in the first stage of their study that both numerical and verbal expressions were applied to describe more than a single graphical display, although numerical expressions were used less than verbal expressions.

Despite the evidence just reviewed that people treat numbers as vague, we assume they would believe numbers to be relatively more precise if they knew the values represented extensive relative frequency data rather than other individuals' judgments. But, to our knowledge, this fact has not been established.

To summarize, a vast array of data at the individual subject level must be interpreted not as error variance, but as indicating that phrase (and number) meaning extends over a range of probabilities. Moreover, the data strongly suggest that a given expression represents a set of probabilities to varying degrees, a point that we formalize below as Principle P1. After considering this principle, we turn to the questions of how these vague meanings differ over people and contexts.

B. Principle P1

Membership functions can be meaningfully scaled. Membership functions, $\mu_3(p)$, over the probability interval [0,1] can be used meaningfully to scale interpretations of numerical or verbal probability expressions or of other subjective representations of uncertainty. The operative term in this statement is *meaningful.* As reviewed earlier, numerous procedures have been devised to establish such scales; our assumption is that they capture subjective meanings and interpretations in a manner that allows prediction of independent behavior.

The first test of meaningfulness in this sense was reported by Wallsten, Budescu, Rapoport, Zwick and Forsyth (1986). In that study, we scaled membership functions for individual phrases by having subjects judge the degree to which a given phrase better represented one spinner probability than another. We then used those scaled values to predict successfully judgments in which pairs of phrases were shown with single probabilities and subjects had to indicate how much more descriptive one of the phrases was than the other.

This could be viewed as a weak test, as the two types of judgments are fundamentally similar, but membership functions have passed stronger hurdles as well. One such test was that of Jaffe-Katz et al. (1989), who examined the semantic congruity and semantic distance effects (e.g., Holyoak, 1978) as applied to verbal (V) and numerical (N) expressions of uncertainty. One goal of that work was to show that the two modes of expression represent the same underlying construct of subjective uncertainty and therefore operate identically and (given the right model) interchangeably in such comparisons. Subjects performed speeded paired comparisons of terms under instructions to choose the larger (or the smaller) term. We used both single mode (VV, NN) and mixed (NV) pairs. The time required for a decision exhibited identical qualitative patterns in all three conditions. Consistent with the symbolic distance effect obtained in other domains, the closer were two terms, the longer it took to compare them; and consistent with semantic congruity effects observed elsewhere, the larger (smaller) of two large (small) terms was identified faster than the larger (smaller) of two small (large) terms. The NN comparisons were made more rapidly than the NV and VV, which did not differ significantly from each other. Important to this discussion is that membership functions were obtained and successfully used in an expanded reference point model that performed 70% better than did the basic version (Holyoak, 1978). Thus, it is fair to conclude that the membership functions predicted well the complex reaction time patterns obtained in speeded magnitude comparisons. In another study, Wallsten, Budescu, and Erev (1988) used membership functions to predict the stochastic properties of repeated choices between lotteries for constant amounts of money based on the outcomes of events described verbally or numerically. As this study provides the underpinning for a subsequent principle, we defer discussion of it here.

Finally, relative membership values were successfully predicted in a bidding study by Budescu and Wallsten (1990, Experiment 2). In that experiment, decision makers bid for lotteries with chance events that had been described verbally and numerically by independent forecasters. The decision makers also provided judgments from which their membership functions for the forecaster's phrases could be inferred. In 23% of the cases, the decision makers' membership values were in fact 1.0 at the forecasters' estimated probability. In another 27% of the cases, the membership was below 1.0 but still higher than the memberships of all of the competing phrases of that probability. Thus, on approximately one half of the occasions, we can conclude on the basis of membership values that the decision makers considered the same particular phrase to be better than any other used by the forecaster for the intended probability. This is a remarkable result, considering the tremendous range of vocabulary the forecasters used.

Given all of the preceding results, we feel justified in concluding that when membership values are properly scaled by means of pair comparison or by suitably constrained magnitude estimation procedures, they provide meaningful representations of an individual's understanding of a phrase within the context that it is being used. Therefore, these membership values can be used in quantitative model testing.

C. Interindividual Variance in the Understanding and Use of Probability Phrases

A widely accepted generalization is that people differentially understand probability phrases (e.g., Clark, 1990). Consequently, different individuals use diverse expressions to describe identical situations and understand the same phrases differently when hearing or reading them. The data strongly support both of these statements and show that people have surprisingly rich and individualized lexicons of uncertainty.

For example, consider the range of expressions people use in identical situations. In the Budescu et al. (1988) experiment, 20 subjects spontaneously generated 111 distinct phrases to describe 11 different graphically displayed probabilities. Similarly, 40 forecasters in a study by Tsao and Wallsten (1994, Experiment 5) freely selected 148 unique phrases to describe 11 probabilities of drawing balls of specific colors from urns. Zwick and Wallsten (1989) report an experiment in which 20 individuals used an average of over 35 distinct expressions to represent the uncertainty of 45 real-world events. Many subjects in a revision of opinion experiment (Rapoport, Wallsten, Erev, & Cohen, 1990) used over 30 phrases despite a request to limit the number to 15. Wallsten, Budescu, and Zwick (1993) asked 21 subjects to create a vocabulary for expressing degrees of confidence in the truth of almanac-type statements. Each respondent was required to include the anchor terms *certain, toss-up,* and *impossible* and to select 8 additional phrases from a list of 64 such that he or she considered the whole [0,1] probability interval to be covered. Overall, 60 distinct phrases were selected, of which only 8 were chosen by more than 5 subjects and 20 were each selected by only a single individual. Finally, in a study by Erev and Cohen (1990), four experts each freely generated between 10 and 17 distinct terms in making probability judgments about 27 basketball events. Thus, without question, people have different working vocabularies for expressing degrees of confidence or uncertainty and create different lexicons for themselves when given the opportunity to do so.

To investigate the flip side of the problem of how people understand phrases when they receive them, it is necessary to compare individuals' responses to distinct expressions. Numerous studies of phrase-to-number conversion have reported vary large degrees of between-subject variability in the assessments of the same terms in a fixed context or in the absence of a specified context. The list of studies and replications is too long to be reproduced here. Johnson and Huber (1977) using Army personnel, Budescu and Wallsten (1985) using psychology graduate students and faculty, and Mullet and Rivet (1991) using children between the ages of 9 and 15 all found that variability between subjects far exceeded that within. Other studies finding considerable interpersonal variability in interpreting probability phrases among either lay people or experts within their professional domains include Beyth-Marom (1982), Brackner (1985), Bryant and Norman (1980), Chesley (1985), Clarke et al. (1992), Farkas and Makai-Csasar (1988), Hamm (1991), Kong, Barnett, Mosteller, and Youtz (1986), Lichtenstein and Newman (1967), Merz, Druzdzel, and Mazur (1991), Murphy et al. (1980), Nakao and Axelrod (1983), and Sutherland et al. (1991). Two exceptions of interest include a study by Brun and Teigen (1988, Study 2) that demonstrated greater consensus among physicians than among parents of young children in assigning numerical meanings to linguistic probabilities. The other is by Timmermans (1994), who noted that experienced and resident internists and surgeons interpreted probability terms similarly when applied to describe symptoms.

Of course, the degree of interindividual variance is not identical for all terms. Consensus regarding the meaning of phrases tends to be greatest near the ends of the continuum (e.g., *almost certain* or *practically impossible*) or in the vicinity of 0.5 (e.g., *even odds*), and to be least between these anchor points. An interesting result, especially from a linguistic perspective, is that negation (e.g., Reyna, 1981) and symmetric reversals (e.g., Clarke et al., 1992; Reagan et al., 1989; Lichtenstein & Newman, 1967) do not necessarily lead to complementary numerical estimates.

The phenomenon is not an artifact of the differential use of a numerical scale. Budescu and Wallsten (1985) elicited rankings of various terms and found interindividual rank reversals (see also Moore & Thomas, 1975). Using somewhat different methodology, Reagan et al. (1989) asked 115 subjects to provide words-to-numbers translations, words-to-numbers (range) acceptability functions, and numbers-to-words acceptability functions using 18 probability phrases and 19 numerical values (from .05 to .95 in steps of .05), and analyzed the joint results. They found large variance across subjects in all the tasks. They were also able to demonstrate high levels of consistency (mean correlations above .85 for all pairs of tasks compared) in the distribution of responses across individuals.

All of the studies just reviewed demonstrate extreme variation over individuals when people are required to translate phrases into numerical equivalents or to indicate which numbers are acceptable translations or expressions. Analysis of individual membership funtions carries this result a step further. We have shown (Wallsten, Budescu, Rapoport, Zwick, and Forsyth, 1986; Rapoport et al., 1987; Budescu & Wallsten, 1990) that the location and spread of functions representing any given term vary considerably across subjects. Even more impressive is the fact that the shape of these functions is not universal. For example, Budescu and Wallsten (1990) report that 25% of the functions describing *unlikely* are monotonically decreasing and 67% are single-peaked; of the functions describing *very good chance,* 44% are single-peaked and the same proportion are monotonically increasing. Thus, not only do phrases differ over individuals in their central meaning, but they also differ in the extent and nature of their vague referents. To the degree that membership functions carry implications for the semantics of terms, such semantics vary considerably over individuals.

D. Intra-individual Sensitivity to Context

One obvious solution to potential communications problems raised by intraindividual vagueness and interindividual variability in understanding probability phrases is to standardize the language. That is, develop a verbal scale by identifying a reasonably small subset (7 to 13 members) of frequently used terms, impose a ranking, and associate a range of probabilities with each of the terms on the list. It is a little known fact that the National Weather Service (NWS) has done just that with respect to probability of precipitation (POP) forecasts (National Weather Service, 1984, Chapter C-11). In issuing POP forecasts, NWS meteorologists can translate .10 and .20 only to the term *slight chance;* .30, .40, and .50 only to *chance,* and .60 and .70 only to *likely.* Other terms are not allowed in POP forecasts.

The general argument for a standardized probability language is best articulated by Mosteller and Youtz (1990) (comments by other researchers for and against their proposals plus their rejoinder immediately follow their article, and we refer readers to the entire interesting discussion). Too many scales have been proposed to mention all of them here, but representative examples include Beyth-Marom (1982), who grouped 19 terms into seven categorical ranges; Hamm (1991), who suggested a single best term for each of the 19 intervals of width 0.05 in the .05 to .95 range; and Kadane (1990), who proposed 11 terms to label 11 intervals. And, of course, there is the infamous scale used by NASA engineers that was linked (Marshall, 1986, 1988) to the space shuttle accident.

In a similar spirit, artificial intelligence researchers who need to incorporate uncertainty in expert systems have suggested that a selected subset of terms be represented by a family of partially overlapping membership functions. For example, Bonissone, Gans, and Decker (1987), in an expert system named RUM, suggested using nine terms from Beyth-Marom's (1982) list and describing each one by a trapezoidal function; Degani and Bortolan (1988) proposed triangular membership functions for 13 terms; and López de Mántaras, Meseguer, Sanz, Sierra, and Verdaguer (1988) described a medical diagnostic expert system (MILORD) in which uncertainty is captured by nine terms represented by trapezoidal functions.

A standardized scale is feasible if: (1) people can suspend or suppress the meanings they normally associate with particular terms, and (2) meanings and representations of selected terms are invariant over all contexts in which they are applied. The former condition has not been studied extensively. In fact, the only study that has addressed this issue was reported by Wallsten, Fillenbaum, and Cox (1986, Experiment 1). They showed that weather forecasters were subject to base-rate effects in a domain outside of their expertise with the very phrases that had been endowed with standardized meaning in the context of POP forecasts. The effects were substantial. For example, *chance* was interpreted on average as indicating a probability of .39 when referring to the likelihood that an ankle twisted in a soccer game was sprained rather than broken, but as .18 when referring to the likelihood that severe life-threatening effects would accompany a flu shot. This result strongly suggests that meanings cannot be legislated.

The second condition is the primary subject of this section. As it turns out, most empirical results obtained to date show that this condition is systematically violated in interesting ways. Just as with terms of frequency (e.g., Pepper, 1981) and quantity (e.g., Newstead, 1988), probability phrases tend to change their meanings according to the context in which they are used.

The first result of note is simply that context matters. In Beyth-Marom's (1982) study, a group of political forecasters translated 14 common probability terms to numbers on two occasions. The phrases first were presented in isolation and then embedded in paragraphs from reports published by their organization. Interestingly, the second administration lead to greater interindividual variance for most expressions. This pattern of increased variability in specific contexts was replicated in two studies reported by Brun and Teigen (1988). Mapes (1979) reported a study in which physicians assigned different distributions of numerical values to probability (and frequency) terms when used to describe likelihood (and frequency) of side effects in response to different medications. Beyth-Marom (1982) speculated that there were three reasons why interindividual variance may have increased with the context she supplied, (1) because the context included other vague terms (e.g., *hostile activities* or *severe illness*) that also required interpretation so that the subjects perceived the nominally identical frames differently, (2) because individuals imposed their own judgments on the contexts rather than just interpreting the phrases, or (3) because perceived outcome values influenced the interpretations of the phrases. Whatever the cause, and it may have been all three, the results certainly demonstrate that the interpersonal variability in translating phrases to numbers commonly observed in laboratory settings does not result from the *absence* of a specified context as has been occasionally suggested (e.g., Moxey & Sanford, 1993; Parduci, 1968).

Context, of course, is not a well-defined unidimensional concept. To reach more specific generalizations, we review next a few situational variables that have been shown to affect systematically probability phrase meanings. We first consider perceived *base rates.* Following Pepper's (1981; see also Pepper & Prytulak, 1974) lead with frequency expressions, Wallsten, Fillenbaum, and Cox (1986) have shown that the numerical interpretation of a term used to describe the chances of a given event occurring depends on whether the context implies a high or a low base rate. For example, the numerical translation of *probable* when referring to the likelihood of snow in the North Carolina mountains is higher when the statement specifies the month of December than when it specifies October. In an elaborate study (Wallsten, Fillenbaum, & Cox, 1986, Experiment 2), the average effect size due to base rate exceeded .11 over the nine probability terms used, but was considerably stronger for the high (e.g., *likely*) and neutral (e.g., *possible*) terms than for the low (e.g., *improbable*) terms. Detailed analyses at the group level suggested that the meaning assigned to any given term may be a weighted combination of its meaning in isolation and the perceived scenario base rate.

Recently, Shapiro and Wallsten (1994) replicated the effect with a different set of scenarios covering a wider range of base rates and also with numerical as well as verbal expressions. Because they had the same subjects both judge base rates and interpret numbers and phrases, they were able to analyze data at the individual level. It appeared on this basis that the averaging may have been an artifact of analyzing group rather than individual data. Shapiro and Wallsten suggested that relatively few people rely on base rates when interpreting low expressions because base rates generally are used only in the narrow range below some "neutral" level and therefore are only associated with those probabilities. The range is narrow because neutral points are never greater than .5, and often are much less. That is, when a phrase is used to describe the chances of an event within a context in which only two alternative events are possible, neutral is naturally taken as .5. If there are four possible outcomes, neutral is .25.

A few studies have looked at phrase meaning directly as a function of the *number of possible alternatives, n.* Tsao and Wallsten (1994, Experiments 1–4) encoded subjects' membership functions for low, neutral, and high phrases when *n* equaled 2 or 4. The functions tended to shift left when $n = 4$ relative to when $n = 2$ for probabilities from 0 to roughly .60. As a consequence, the function for the neutral term *even chance* tended to peak at .50 and .25 for $n = 2$ and $n = 4$, respectively; functions for high terms, which increased monotonically in the 0 to .60 range, described lower probabilities better when $n = 4$ than when $n = 2$; and functions for low terms, which tended to decrease monotonically in that range, did the opposite. These are sensible results on the assumption that low, neutral, and high are assessed relative to $1/n$.

Teigen (1988a) obtained parallel results when he examined how the number and relative likelihoods of alternatives affect people's understanding and use of (Norwegian) probability phrases in real-world contexts. He showed that the number of alternatives did not affect the fraction of subjects who indicated that high rather than low terms (e.g., *great chances* vs. *small chances; not improbable* vs. *improbable*) more appropriately described the likelihood of specific outcomes. Teigen interpreted this result as showing that people's tendencies to overestimate probabilities increase as the number of alternatives increases. We prefer the interpretation that meanings of high phrases depend on a perceived neutral point, $1/n$. As n increases (and $1/n$ decreases), high phrases are increasingly appropriate for lower probabilities. If this is so, then their continued use as n increases is not surprising.

Teigen (1988b) did find, however, that the source of the uncertainty dramatically affected phrase selection. The pattern of results suggested subtle semantic effects not understood simply in terms of implied probabilities. Recently, Gonzales and Frenck-Mestre (1993) reported a series of experiments in which various groups of subjects read short scenarios about probabilistic events. The various versions of these vignettes differed with respect to the type and nature of information provided (base rate, trend, global and local weight, etc.), but, in most cases, the event's (objective) probability remained fixed. They showed that the verbal responses (in French) were more sensitive than the numerical ones to variability in base rates and local weight. Both papers emphasize the importance of semantic concerns beyond those captured by simple numerical translations, but their results may depend to some degree on idiosyncratic aspects of the scenarios considered (only a fraction of the results were significant) and/or on the special nature of the response mechanism used (only a small number of phrases were used and subjects were asked to rate their appropriateness by using scales with two to seven categories).

Weber and Hilton (1990) investigated base-rate effects but also *outcome severity* in the context of medical scenarios. In three experiments (using some of the scenarios used by Wallsten, Fillenbaum, & Cox, 1986), subjects judged the meanings of probability phrases associated with each scenario, the case's severity, and its prevalence (base rate). Weber and Hilton replicated the base-rate effect, but also found that on average the phrases were interpreted to imply greater probabilities when associated with events of more severe consequence. The authors correctly pointed out that event severity and prevalence are negatively correlated in medical contexts (serious illnesses are generally less common than mild ones) and therefore their separate effects on meanings of probability terms are hard to disentangle. This correlation can explain the results reported by Merz et al. (1991), who found that the numerical values associated with some expressions decreased when associated with more severe outcomes. Severity effects are not universal. Sutherland et al. (1991) report that cancer patients responded similarly to verbal descriptions of different ("death" vs. "illness") side effects of blood transfusion.

Closely related to the notion of outcome salience is that of *outcome valence,* whether the outcome is positively or negatively valued. Mullet and Rivet (1991) compared scale values assigned to 24 French expressions used in predictions of children's chances of passing or failing a test. On average, the positive context induced higher estimates. A similar pattern was found in two experiments described by Cohen and Wallsten (1992), in which subjects compared pairs of lotteries whose (positive or negative) payoffs depended on the outcomes of binary probability spinners. For most subjects and most words in both studies the inferred probability of an expression was higher when associated with a positive than a negative outcome.

Yet another context effect relates to the *characteristics of the available uncertainty vocabulary* (such as its length, composition, and structure) when judging particular expressions. Hamm (1991) found less interindividual variance and better discrimination among the meanings of 19 phrases when they were presented in ordered rather than random lists. Clarke et al. (1992) found similar results. Fillenbaum et al. (1991) elicited membership functions for a list of core expressions (*likely, probable, possible, unlikely,* and *improbable*) and for the same words when embedded in longer lists including modified expressions (such as *very likely, quite probable,* etc.) or anchor terms (such as *toss-up* and *almost certain*). The additional terms did not affect the shape, location (W), or scatter (V^2) of the core phrases, but they did decrease the degree to which these words were judged to be most appropriate for certain events.

Finally, a particularly important factor in determining phrase meaning concerns one's *role in a dialogue or exchange.* The recipient of a communication may understand a probability expression differently than intended by the originator because the two individuals differentially interpret the base rates, valences, or severities or outcomes, or because meaning depends more generally on the direction of communication. We are not aware of any research that has compared recipients' and originators' perceptions of context, but a number of studies have looked at the overall effects of communication direction.

In a study of dyadic decision making (Budescu & Wallsten, 1990), one member of the dyad (the forecaster, F) saw a probability spinner and communicated the chances of the target event to another subject (the decision maker, DM) who used this information to bid for a gamble. The DM had no direct access to the spinner and the F was not informed of the amounts to be won or lost. Both subjects provided numerical translations for all of the words used. A clear pattern emerged: The recipients of the verbal forecasts generally assigned values closer to 0.5 than originally intended by the communicators. Fillenbaum et al. (1991) obtained analogous results with elicited membership functions. Functions were located closer to 0.5 (in terms of their W values) and were broader (in terms of V^2) for phrases selected by others than for those selected by the subjects themselves for communication. Thus, both the Budescu and Wallsten and the Fillenbaum et al. studies suggested that people interpret phrases more broadly and more centrally when receiving them than when selecting them.

To summarize, context effects on the interpretation of probability terms are pervasive. Individual differences in assigning numerical values to expressions are greater when the expressions are used in discourse than when they appear alone. More specifically, perceived base rate, number of alternatives, valence and severity of outcomes, and direction of communication all have been shown to have large and systematic effects on phrase meaning. To a lesser degree, meanings may also depend on the available vocabulary and on the order in which phrases are seen. It is interesting to note that studies have been carried out in many languages and in many cultures with no appreciable differences in the magnitudes or directions of results. Nevertheless, in view of the cultural differences in probabilistic thinking documented in other domains (Wright et al., 1978; Yates, Zhu, Ronis, Wang, Shinotsuka and Toda (1989), it would be of interest to search carefully for any differences that might exist.

E. Principle P2

The location, spread, and shape of membership functions vary over individuals and depend on context and communication direction. This principle implies that the individual differences summarized in section III(B) and the systematic effects on phrase meaning described in section III(C) can be represented by membership function location, spread, and shape. Individual differences in membership functions are already well documented and were discussed in section III(C). But the representation of context and communication effects on such terms has not been directly tested.

A particularly interesting implication of principle P2, should it be correct, is that membership functions may provide the quantitative handle necessary for theorizing about the semantic features of probability expressions. Moxey and Stanford (1993), in their very selective review, suggest that the use of membership functions or other quantitative measures is inconsistent with understanding the semantic functions of such phrases, because, among other reasons, phrase meaning is affected by context. That is a very pessimistic view. An alternative one is that each expression in an individual's lexicon can be represented by a basic membership function, which is operated upon by the particular context, communication direction, and communication intention. Parsimonious theory development along such lines requires that a particular context, direction, or intention affect all basic membership functions in a similar way. Research testing this idea remains to be done.

F. Communication Mode Preferences

Do people prefer verbal or numerical communications, and to what degree do any such preferences depend on the properties of language that we have previously documented? A corollary of background assumption B1, that all representations are vague, is that people generally prefer to communicate their opinions verbally because this mode conveys the underlying vagueness. The general folklore is consistent with this expectation and so are the empirical results. Erev and Cohen (1990) had basketball experts give their opinions of the likelihood of future basketball events. Three of the four experts spontaneously used verbal rather than numerical probabilities for almost every prediction. Subsequently, 21 students who had used these judgments were asked to provide their own judgments for future decision makers. Of these, 14 (67%) spontaneously did so verbally. In a medical context, Brun and Teigen (1988) found that 50 of 66 physicians (76%) preferred conveying their likelihood judgments in a verbal fashion. The suggestion that these results occurred simply because it is easier for people to communicate verbally rather than numerically is not a criticism. Rather, it is a natural explanation for this pattern if one assumes that it is easier to communicate in the modality that provides the most accurate representation of the underlying opinion.

This interpretation and related findings are supported in a survey of 442 respondents (undergraduate, graduate nursing, and MBA students) by Wallsten, Budescu, Zwick, and Kemp (1993). Overall, 77% thought that most people prefer communicating verbally rather than numerically and 65% indicated that mode as their initial personal preference. An analysis of reasons given for the preferences suggested that verbal communications generally are preferred unless the underlying opinions are based on solid evidence and are therefore relatively precise or unless the importance of the occasion warrants the effort to attempt greater precision.

Assumption B1 provides less guidance about what to expect about people's preferences for receiving judgments of uncertainty, but the data are clear: most people prefer receiving such judgments numerically. In the Erev and Cohen (1990) study, 27 (75%) of the decision makers always chose to look at numerical rather than verbal information and only 4 (11%) always did the reverse. Brun and Teigen (1988) found that 38 of 64 patients (60%) preferred receiving numerical rather than verbal judgments. In the Wallsten, Budescu, Zwick, and Kemp (1993) survey, 70% of the respondents expressed an initial personal preference for receiving information numerically. The same reasons were given as for communicating to others. Those who preferred verbal judgments did so because they are more natural, more personal, and easier; and those who preferred numerical judgments did so because they are more precise.

The fact that most people have initial preferences for communicating to others verbally and for receiving from others numerically means that there must be some people who have both preferences simultaneously. The Wallsten et al. survey found this pattern for 35% of their respondents, and Erev and Cohen found it for 47% of the decision makers. The reverse preference pattern, to communicate numerically to others and to receive verbally from others, virtually never appeared.

In a series of papers, Teigen and Brun have shown that linguistic probabilities have connotations that are not captured by their numerical counterparts, which provides another reason why their use often is preferred. Among these connotations are "affective intensity" (Brun & Teigen, 1988) and "directionality" (Teigen, 1988b; Teigen & Brun, 1993). Another indication of people's sensitivity to the vagueness of probability terms is the fact that many experts insist on communicating their opinions verbally. The evidence is primarily anecdotal but quite convincing (see Behn & Vaupel, 1982, and Wallsten, 1990, for some interesting examples). Beyth-Marom (1982) speculates that this preference is related to the perception that (because of their vagueness) linguistic forecasts cannot be verified and evaluated as can their numerical counterparts (but see the recent method proposed by Wallsten, Budescu, and Zwick, 1993).

G. Principle P3

Communication mode choices are sensitive to the degrees of vagueness inherent in the events being described, the source of the uncertainty, and the nature of the communication task. This principle is really a corollary of background assumption B2, but it also summarizes the main results of the previous section. It captures the idea that most people prefer the communication mode that they can use most easily and accurately for the task at hand. Thus, they communicate their opinions to others in a verbal rather than numerical form because it is the easiest way to convey different locations and nuances of imprecision. On the other hand, even when recognizing that opinions are not precise, people prefer to receive information numerically because they know that individuals differ greatly in their phrase selection. Therefore, people believe they can more easily and accurately obtain a fix on the rough uncertainty location another individual intends to communicate when he or she does so numerically rather than verbally.

One implication of this principle was developed and tested by Erev, Wallsten, and Neal (1991). They suggested that in many contexts vague communications promote the well-being of a society to a greater extent than do precise ones, and that people will select their communication modes accordingly. Vague language is beneficial because it is more likely to result in heterogeneous rather than homogeneous actions, which in turn increase the chances that at least some individuals will be successful and the society as a whole will survive. Their experiment involved groups (societies) of seven subjects each. On each trial, one subject was the F, three were DMs, and the remaining three were passive participants. Subjects' roles rotated over trials. The F alone saw a probability spinner partitioned into three sectors and communicated to the group the probability of its landing on each one. The F had the option of selecting numerical or verbal terms to communicate these assessments. The DMs then individually and privately chose one of the three events to bet upon. Two different groups were run with different payoff structures. In the first condition, the group's total outcome was maximized if all subjects agreed in their choices; but in the second group, the total outcome was maximized by a pattern of heterogeneous selections. As predicted, the proportion of cases in which Fs used verbal terms increased systematically across trials in the second group but remained stable in the first. The clear implication is that subjects realized the condition in which vagueness was beneficial and precision harmful, and chose to use linguistic terms in that case.

IV. Judgment and Decision with Verbal and Numerical Probabilities

The picture that emerges from the previous section is that linguistic terms are imprecise and subject to multiple interpretations by different individuals and under different conditions. Therefore, they are problematic in communication and should be avoided for important decisions. We (Budescu & Wallsten, 1985) have referred to the possible "illusion of communication" induced by the use of these terms, and decision analysts routinely have recommend avoiding them (e.g., Behn & Vaupel, 1982; Moore, 1977; von Winterfeldt & Edwards, 1986). In this section, we review empirical results comparing judgmental accuracy and decision quality given probabilistic information in verbal and numerical form. Most of these studies involved within-subject comparisons of judgments and decisions performed with the two modes of information. The next subsection focuses on studies of probabilistic judgments and inferences, and the following subsection is concerned with experiments in which subjects made choices, gave bids, or in some way took action on the basis of verbal or numerical information.

A. Accuracy of Judgments

According to background assumption B1, all representations are vague to some degree. On this basis, contrary to intuition, verbal judgments are not necessarily less accurate than their numerical counterparts. Indeed, Zimmer (1983, 1984) argued persuasively that they should be more accurate. Beginning with the premise that the mathematics of uncertainty (probability theory) developed only in the seventeenth century, whereas the language of uncertainty is ubiquitous and much older, Zimmer (1983) wrote, "It seems unlikely that the mathematically appropriate procedures with numerical estimates of uncertainty have become automatized since then. It is more likely that people handle uncertainty by customary verbal expressions and the implicit and explicit rules of conversation connected with them" (p. 161). He continued the argument in another article by writing,

> It seems plausible to assume that the usual way humans process information for predictions is similar to putting forward arguments and not to computing parameters. Therefore, if one forces people to give numerical estimates, one forces them to operate in a "mode" which requires "more mental effort" and is therefore more prone to interference with biasing tendencies (Zimmer, 1984, p. 123)

To evaluate these claims, Zimmer (1983, 1984) ran experiments (in German) in which he tested for biases with verbal responses that typically are found with numerical ones (overconfidence in judgment and conservatism in revision of opinion). The data appear to substantiate his prediction of greater accuracy in the verbal mode, but more thorough tests are required. As Clark (1990) pointed out, both studies employed very coarse response scales (median number of categories approximately five), and lacked appropriate direct comparisons with numerical responses. Recently, we tested Zimmer's thesis by allowing subjects a much richer vocabulary (in English) and comparing the results with closely matched numerical controls.

In one study, Zimmer (1983, 1984) had subjects use verbal responses to estimate posterior probabilities in a simple two-alternative Bayesian revision of opinion paradigm. His subjects saw items, bricks or balls, as they were sampled from one of two bins. They knew that sampling was a priori equally likely to be from either bin and they knew the compositions of both bins. After each item was sampled, the subject provided an updated verbal judgment of "the chance of getting a brick or a ball the next time from the bin" (Zimmer, 1983, p. 177). It appears from the description that subjects were limited to approximately five phrases and that for purposes of comparing the responses to the optimal Bayesian values, each was converted to the central probability with the maximum empirically determined membership value for that phrase. The resulting judgments were more accurate (i.e., closer to the predicted Bayesian response) than generally found in studies using numerical responses (e.g., Edwards, 1968).

Rapoport et al. (1990) replicated the study with certain crucial differences. For each problem, subjects saw two urns containing specified proportions of red and white balls. Prior sampling probabilities were always equal, but the urn compositions changed from problem to problem. A within-subject design was used in which each subject provided verbal and numerical judgments in different sessions and a payoff scheme was used to motivate careful responding. Subjects constructed their own vocabularies, and individual membership functions were established for the 14 to 16 phrases each subject used most frequently. Verbal terms were quantified in two ways, by the probability with the highest membership value and by the location measure, W. The results indicated near equivalence in the quality of verbal and numerical judgments. That is, verbal judgments were more variable than numerical, as Budescu and Wallsten (1990) had also found, and consequently were less accurate when measured by mean absolute deviation from optimal. However, on average, verbal judgments were more accurate than numerical ones when the verbal responses were quantified by peak values. To complicate matters, the two modes yielded equal degrees of conservatism when the phrases were quantified by W.

In a related Bayesian revision opinion problem, Hamm (1991) also found very little difference between verbal and numerical modes. In contrast, Timmermans (1994) claimed in a medical context using physician subjects that probability judgments were more accurate (closer to the Bayesian posterior probabilities) in response to the numerical information than the verbal information. Some of Timmermans' conclusions are particularly hard to evaluate, as it is not clear that (1) the "objective" Bayesian values calculated and the estimates provided by the physicians pertain to the same events, or (2) the numerical meanings of the terms, which were elicited in the absence of any context, can be generalized to the diagnostic problems.

Thus, although the data do not support Zimmer's strong claim that reasoning is better in the revision of opinion context when people can respond verbally rather than numerically, neither do they show the opposite to be true. Rather, it appears that reasoning is about equivalent in these tasks given the two modes of responding. The same seems to be true in the case of judgment, in which the usual finding with numerical responses is that people are overconfident. Zimmer (1983) reported results of a study in which 90 soldiers answered 150 political science questions and assessed their chances of being correct by means of verbal expressions. Unfortunately, the description of the experiment is too sparse to properly evaluate it, but the calibration curve that is presented (based on the median value of the words) is closer to the identity line than generally occurs with numerical responses (e.g., Lichtenstein, Fischhoff, & Phillips, 1982). The implication is that subjects are less overconfident and more accurate when their judgments are verbal than when they are numerical. We recently replicated and extended this study and arrived at somewhat different conclusions.

Wallsten, Budescu, and Zwick (1993) asked 21 subjects to report their confidence in the truth of 600 factual items (in the domains of geography, demography, and history) consisting of 300 true statements and their semantically identical false complements. Subjects selected their own vocabularies (in addition to three specified anchor terms). They provided verbal and numerical judgments in separate sessions and encoded individual membership functions for each linguistic term. Verbal and numerical judgments correlated very highly within individual subjects and behaved very similarly in most aspects analyzed. The only differences we observed between the two modes were that the central numerical response (0.5) was used more frequently than its verbal counterpart (*toss-up*) and that the level of overconfidence (excluding the middle category) was higher for the verbal terms. A possible interpretation of these differences is that judges may use the central category more frequently in the numerical than in the verbal mode to represent imprecise judgments, because that category is equally defensible regardless of the outcome. In contrast, the verbal mode allows more honest representations of vaguely formed opinions without resort to the central category. If this is true, then overconfidence, as usually measured, may actually be greater than observed when subjects respond numerically.

Two points must be added in summarizing this evidence and its relation to background assumption B1. First, recent developments by Erev, Wallsten, and Budescu (1994) suggest that the underconfidence common in opinion revision studies and overconfidence common in judgment studies may both arise from one set of processes, leading to data that researchers analyze differently in the two paradigms. Erev et al. (1994) noted that when objective probabilities are independently defined, as in Bayesian revision of opinion studies, researchers analyze mean subjective estimates as a function of objective values and generally find underconfidence. In contrast, when objective probabilities are not independently definable, they analyze percentage correct as a function of the subjective estimates and generally find overconfidence. Erev et al. (1994) applied both types of analyses to three data sets and found that judgments appeared underconfident when analyzed one way and overconfident when analyzed the other way. They generated simple models that assumed well-calibrated underlying true judgment plus an error function and showed that such models yield the full range of observed results under standard methods of analysis. Erev et al. were concerned only with situations of numerical responses, but there is every reason to believe that their conclusions apply to verbal responding as well: that is, verbal probabilities imply underconfidence in one case and overconfidence in the other at least in part because they are an errorful representation of underlying judgment. Furthermore, because the degrees of under- and overconfidence are equivalent in the verbal and numerical cases, we can assume that so is the extent of error. Erev et al. (1994) concluded that questions of true under- and overconfidence cannot be properly addressed without a substantive and an error theory relating responses to judgment. The same applies, of course, to verbal responses. Moreover, this approach suggests that the same theories should apply in both cases.

The second point to make in concluding this summary is that although no systematic evidence has accrued thus far indicating that one mode of responding is more accurate than another, many avenues remain to be explored. For example, Zimmer (1983) suggested that the type of information subjects think about in making forecasts or predictions depends on the mode in which they must respond. He claims that people focus more strongly on qualitative data when they can respond verbally and on quantitative data when they can respond numerically. This is an intriguing idea that is consistent with the "compatibility effect" (Tversky, Sattath, & Slovic, 1988) and deserves follow-up.

Nevertheless, the available data provide no indication that the one mode of representation systematically leads to more accurate judgments than does the other. This conclusion is not implied by assumption B1, but it is eminently consistent with it, and no further summary principle is required.

B. Choice and Decision Quality

Although Knight (1921) and Keynes (1921) distinguished between various types of uncertainty, it was Ellsberg's (1961) famous paradox that made decision theorists pay closer attention to the effects of vagueness (or ambiguity, as it is often, but incorrectly, described) on preference (see Budescu & Wallsten, 1987, and Camerer & Weber, 1992, for reviews). Ellsberg's results, and subsequent empirical studies (e.g., Curley & Yates, 1985; Einhorn &

Hogarth, 1985) seem to indicate that, everything else being equal, most people prefer precise probabilities over vague (ambiguous) representations of uncertainty. This pattern of "avoidance of ambiguity" is consistent with findings, described previously, indicating that people prefer receiving numerical rather than verbal information. As phrases are relatively more vague than numbers, one would expect that choices and overt decisions (as operationalized by bids, attractiveness ratings, or rankings of risky options) reflect this preference. On the other hand, Zimmer's argument (1983, 1984) regarding the superiority of the verbal mode makes just the opposite predictions. Next, we describe a series of studies in which these two conflicting predictions were tested empirically.

Budescu et al. (1988) reported two experiments comparing the quality of decisions based on verbal and numerical probabilities. In the first stage of the study, we determined the single best numerical and verbal representations for 11 distinct spinners for each subject. Next, we used these representations to describe lotteries (some involving gains and others losses), which the subjects evaluated either by bidding for the lotteries (Experiment 1) or rating their attractiveness (Experiment 2). Although, on the average, subjects won more (by 1.2.%) and lost less (by 4.7%) with numerical probabilities, the bids, attractiveness ratings and decision times (after eliminating the possibility of calculations) were almost identical under the three presentation modes. Because the results may reflect the high level of intra-individual consistency in the use of words (further reinforced by the initial judgment stage), Budescu and Wallsten (1990) replicated the study with dyads. The F saw probability spinners and communicated the probabilities of the target events in either numerical or verbal form to the DM who had to bid for gambles. The DM could not see the spinner and the F was not informed of the amounts to be won or lost for each specific gamble. Yet, the mean bids, and the expected gains, were identical under the two modes of communication.

Erev and Cohen (1990) used a more realistic version of this dyadic paradigm. Four expert sportscasters provided numerical and verbal forecasts for 27 basketball events (e.g., one player will score more points than another) in a randomly chosen game from a group of games identified by the experimenters. Then, 36 students were asked to rate the attractiveness of gambles whose outcome probabilities were given by the experts' forecasts. The actual events were disguised, forcing the subjects to rely on the expert judgments and not on their own knowledge. The gambles were presented in sets consisting of eight events (with verbal or numerical probabilities and monetary outcomes) that the subjects ranked from the most to the least attractive under a payoff scheme that motivated careful responding. Subjects' ratings and expected payoffs were well above chance level but did not vary as a function of the mode of communication.

In a recent experiment (Gonzáles-Vallejo, Erev, and Wallsten, 1994), Fs made predictions about video games in which the event probabilities were controlled. Fs observed events on the computer screen and provided simultaneous verbal and numerical estimates of the event probabilities. Each DM subsequently was presented with sets of six monetary gambles, whose probabilities were the numerical or verbal forecasts of a particular F, and had to rank the gambles from the most to the least attractive under a payoff structure similar to that used by Erev and Cohen (1990). Once again, the DMs' expected profits were practically identical under the two modes of forecasting.

The single exception to this line of results is Experiment 5 of Tsao and Wallsten (1994). Fs provided numerical and verbal estimates of event chances when the number of possible outcomes was $n = 2$ or $n = 4$. On the basis of these estimates, DMs subsequently estimated (under a payoff scheme designed to promote accuracy) how many times out of 100 the event would occur. Estimates based on verbal and numerical judgments were equally accurate when $n = 2$, but not when $n = 4$. In the latter case, quality suffered in the verbal mode. This task differed from the other decision tasks in that it required the estimate of a sample statistic, rather than a choice or a bid for a gamble. Nevertheless, this suggests that more careful work is needed in contexts in which $n > 2$.

To summarize, with the single exception just noted, all of the experiments described here agree that, on the average, decision quality is unaffected by the mode in which the probability information is provided. To say the least, the result is puzzling in light of the material discussed in section III. In part, it may be a result of the fact that judgmental accuracy is roughly equivalent given verbal and numerical information, as described in section IV(A). But that explanation alone will not do, as it still does not explain the approximately equal decision quality given the wide interpersonal variability in interpreting phrases and the considerably greater imprecision of verbal versus numerical expressions. However, the next principle provides the link that in conjunction with the findings of section IV(A) brings together the two otherwise contradictory sets of results. We propose one more principle that suggests a subtle but important way in which the two modes, verbal and numerical, do differentially affect trial by trial decision behavior while generally leaving average results unaffected.

C. Principle P4

The question we must address is How are imprecise assessments of uncertainty resolved for the purpose of taking action? (Assumption B2 asserts that such resolution takes place, but does not specify the mechanism.) A natural answer would be to assume that when faced with an expression covering a wide range of probabilities—but representing some values within that range better than others—one restrict attention to the best-described values. An extreme version of this assumption would be that for purposes of making a decision, one treats a phrase as equivalent to that probability (or to the mean of those probabilities) for which the membership function, $\mu_e(p)$ is maximal. If this were true, however, then decision variability would be equal under verbal and numerical information conditions, and it generally is not. For example, Budescu and Wallsten (1990) found, on a within-subject basis, that the variance of bids based on verbal probabilities was larger than the variance of numerical bids by a factor of 2.44. Thus, a weaker assumption is needed. We propose, instead, that **when combining, comparing, or trading-off information about uncertainty with information about other dimensions, such as outcome values, the uncertainty representation, $\mu_e(p)$, is converted from a vague interval to a point value by restricting attention only to values of p with membership above a threshold v, that is, for which $\mu_e(p) \geq v$.** A specific point value p^* is then selected probabilistically according to a weighting function proportional to the $\mu_e(p) \geq v$. Expressed formally, p^* is selected for expression e according to the density $f_e(p^*)$ defined by

$$f_e(p^*) = \frac{\mu_e(p^*)}{\int_x^y \mu_e(p)\,dp}, \tag{7}$$

where

$$x = \left[\begin{array}{l} 0 \text{ if } \mu_e\,(p) \text{ monotonically decreases} \\ \text{Min}\,(p|\mu_e(p) = v) \text{ otherwise,} \end{array}\right.$$

and

$$y = \left[\begin{array}{l} 1 \text{ if } \mu_e\,(p) \text{ monotonically increases} \\ \text{Max}\,(p|\mu_e(p) = v) \text{ otherwise,} \end{array}\right.$$

Equation (7) says that membership values above the threshold v are converted to (proportional) choice probabilities.

Good but limited support for this principle comes from a study by Wallsten et al. (1988), in which subjects chose between two gambles, (a, p, 0) and (a, q, 0). That is, one gamble offered outcome a with probability p, and outcome 0 with probability 1-p; and the other offered the same outcomes with probabilities q and 1-q, respectively. The value, p, was easily estimated as the relative area of a visible spinner; whereas the value q was conveyed by means of a probability phrase. For each of the 10 probability phrase–outcome value combinations (five phrases by two outcomes), subjects made nine choices at each of six levels of p. Subsequent to the choice phase of the study, each subject provided membership functions for each phrase used. Finally, a single threshold parameter v was sought for each subject that provided for the best prediction of his or her tendency to choose the gamble with the visible spinner by converting membership values greater than or equal to the threshold to sampling weights. Although the model did not fit perfectly, the mean deviation between predicted and observed choice probabilities was very small, .01, for 9 of the 10 subjects.

We have replicated these results in unpublished work,[3] but have not pursued the obvious question of what factors control the placement of v. Various hypotheses are reasonable. One might argue, for example, that as decision importance increases, so too does motivation to consider the full range of possible meanings conveyed by an expression. Consequently, v should decrease with decision importance. Thus, as we argued earlier, when subjects are allowed but not required to give a probability range in situations of vague uncertainties, and the response is of no real consequence, they simplify their task to the greatest extent possible by setting $v = 1$. Presumably, they would set v lower for more important decisions. An alternative line of reasoning, however, suggests that v increases with decision importance, because, as illustrated earlier, people typically prefer precise information for important decisions; that is, as the stakes increase, so too does the necessity to make a clear decision. These competing conjectures remain to be tested. Along another dimension, assuming that the greater v, is the easier and more rapidly can the decision be made, one might expect that v increases with time pressure. Again, we have no data.

Principle P4 provides a qualitative explanation of Budescu and Wallsten's (1990) dyadic decision results, which were discussed previously, but has not been further tested. Clearly, additional evaluation is required. Nevertheless, assuming the principle's validity, we are in a position to understand why average decision quality is unaffected by probability mode. Note first that when a phrase is converted to a probability value, p^*, for decision purposes, its expected value, $E_e(p^*)$, is equivalent to W_e defined for $\mu_e(p)$ restricted

[3] Wallsten and Erev ran an additional nine subjects in essentially the same design, but with more observations per point, and obtained equivalent results.

to values greater than or equal to v. This fact can be seen by using Eq. (7) to obtain $E_e(p^*)$ and comparing the result to Eq. (1). Moreover, for symmetric single-peaked functions, W_e equals the peak probability, that is, the value p^{**} such that $\mu_e(p^{**}) = 1$. However, regardless of the shape of the membership function, as v increases $E_e(p^*)$ approaches p^{**} (or the mean of the p**, should the value not be a single point). Therefore, we can claim that in general, when making decisions, people interpret phrases as equivalent to probabilities in the neighborhood ranging from their central value, W, to their peak value(s), p^{**}.

Two more steps are needed to complete the explanation of why average decision quality is unaffected by probability mode. Recall, first, that people treat numbers received from others as vague at least to some degree. Therefore, we can assume that just as occurs with phrases, DMs interpret numbers somewhat more broadly and centrally than the F intended. Second, recall from section IV(A) that verbal and numerical information is processed with roughly equivalent accuracy. On that basis, we can deduce that for a given DM, there is no systematic difference between W_e or p^{**} of the phrase selected by a F and the interpreted meaning of the numerical probability communicated by that F. Given this entire train of argument, it is not surprising that average decision quality was unaffected by probability mode in the experiments reviewed in section IV(B), while, simultaneously, decision variance was somewhat greater when communication was verbal than when it was numerical.

D. Principle P5

Decision patterns differ in more than just variance given verbal and numerical information, although the additional difference is sufficiently subtle that it eluded us for some time. We express this result as the following principle: **When combining, comparing, or trading-off information across dimensions, the relative weight accorded to a dimension is positively related to its precision.** In other words, the narrower $\mu_e(p)$ is, the more weight is accorded to p^*.

The finding leading to this principle was first evident in the data of a pilot study run by González-Vallejo et al. (1994). The effect was replicated in a second, much more substantial experiment, which we now consider. In the first phase, Fs observed uncertain events in a video game and expressed their judgments of the probabilities both verbally and numerically. These events were then used as the basis for gambles of the form (a, p, 0). DMs saw the gambles in sets of six and rank-ordered them from most to least preferred under a payoff scheme. DMs never saw the actual probabilities but only the Fs' numerical estimates for some sets and verbal estimates for others. The sets all were constructed such that the outcomes and probabilities were negatively correlated, and that the gambles' expected values agreed in rank order with the outcomes for one half of the sets, and with the probabilities for the other half. When the probabilities were expressed verbally, subjects' rankings correlated positively with the payoffs. In contrast, when the probabilities were expressed numerically, the subjects' rankings were positively related to the probabilities. Consequently, when ranking gamble sets in which outcomes were positively correlated with expected value, subjects made more money given the verbal probabilities than the numerical probabilities; and when ranking gamble sets with the opposite structure, the reverse occurred. This result is consistent with Tversky et al.'s (1988) contingent weighting model, suggesting that the relative weight given to the probability and the outcome dimensions depended on the information format. Thus, we see that *on the average* subjects can do equally well under either mode, but not given a particular stimulus structure.

Principle P5 also motivated, in part, a study by Gonzáles-Vallejo and Wallsten (1992) on preference reversals. Six subjects, acting as Fs, provided verbal and numerical probability judgments of events in video displays. These events served as the basis for gambles (a, *p*, 0) shown to 60 DMs, who saw only the outcomes, a, and the verbal or numerical estimates of *p*. They, in turn, bid twice for individual gambles and chose twice from pairs of gambles, once in each task given the numerical expressions and once given the verbal expressions. The regular pattern of reversals between choices and bids (e.g., Grether & Plott, 1979) was found in both modes, but its magnitude was much smaller in the verbal case. Looked at differently, bids were very similar given the verbal and numerical representations, but the choices were not. Considerably greater risk aversion was shown in the numerical case than in the verbal case.

This entire pattern of results is expected, given the conjunction of P5 and the oft-repeated generalization that "probabilities loom larger in choice than in bidding" (Tversky et al., 1988). That is, when bidding for gambles, people tend to focus relatively more strongly on the outcomes than on the probabilities. Therefore, an additional decrease in the weight accorded the probability dimension resulting from vagueness in the verbal condition has little effect on the bids. In contrast, this weight decrease strongly affects choice, where probabilities are in primary focus.

V. Recapitulation and Conclusions

A. Assumptions and Principles

For ease of reference, it is useful to repeat here the two background assumptions and the five principles. We will then consider the epistemological status of each and their joint implications.

B1. Except in very special cases, all representations are vague to some degree in the minds of the originators and in the minds of the receivers.
B2. People use the full representation whenever feasible, but they narrow it, possibly to a single point, if the task requires them to do so.
P1. Membership functions can be meaningfully scaled.
P2. The location, spread, and shape of membership functions vary over individuals and depend on context and communication direction.
P3. Communication mode choices are sensitive to the degrees of vagueness inherent in the events being described, the source of the uncertainty, and the nature of the communication task.
P4. When combining, comparing, or trading-off information about uncertainty with information about other dimensions, such as outcome values, the uncertainty representation, $\mu_e(p)$, is converted from a vague interval to a point value by restricting attention to values of p with membership above a threshold v, that is, for which $\mu_e(p) \geq v$. A specific point value p^* is then selected probabilistically according to a weighting function proportional to the $\mu_e(p) \geq v$.
P5. When combining, comparing, or trading off information across dimensions, the relative weight accorded to a dimension is positively related to its precision.

Our story is straightforward. Rarely is one's opinion or judgment precise (B1), although one acts upon precise values when action is called for (B2). The nature and extent of one's vague representation depends on various individual, situational, and contextual factors (P2), and is measureable in a meaningful fashion (P1). One converts vague opinion to a point value for purposes of action by a probabilistic process that is sensitive to the form of the opinion and to the task (P4). However, despite this conversion, the degree of attention one accords to a dimension depends on its underlying vagueness (P5). Somewhat outside this stream that travels from opinion to action is the issue of communication mode preferences, which are systematically affected by degrees of vagueness (P3). Finally, also important but not formulated as a principle because it does not have theoretical content, is the fact that judgment is approximately equally accurate given verbal and numerical expressions of probability.

These assumptions and principles are not epistemologically equivalent. We have simply asserted assumption B1, but numerous corollaries that can be inferred from it are empirically supported. Similarly, there are no data to sustain assumption B2 in its full generality, but a particular instantiation of it, principle P4, is well supported. Principles P1 and P5 are well buttressed by data; whereas, in contrast, principle P3 describes a range of results, but does not have independent post hoc support. Finally, principle P2 is at this point a reasonable conjecture that is consistent with a good deal of related evidence.

B. Concluding Comments

We have attempted to review a substantial portion of the literature on linguistic probability processing, to develop a coherent theory of such processing in the form of a small set of principles, and to suggest additional necessary research. In this final section, we summarize and extend remarks relating to the third goal that have been sprinkled throughout the chapter.

An important missing link in the present development is firm support for principle P2, that membership function characteristics vary systematically with individuals and contexts. Two major obstacles must be overcome in order to achieve this goal, one technical and the other substantive. On the technical end, a relatively large number of repetitious judgments are required to obtain accurate and reliable empirical membership functions. The task is boring, it is unlikely that successive judgments are independent, and the very judgments being elicited may cause perceived meanings to change during the course of long sessions. One possible solution is to rely more heavily on parametric curve fitting for the various membership functions. In the past, we have used cubic polynomials (Wallsten et al., 1988; Zwick, 1987), but other approaches (splines, piecewise logistics) may work just as well. These techniques may require fewer points and the effects of the various factors may be detected by examining trends in the parameters of the fitted functions.

More important, however, may be the lack of a good theory of context and an appropriate framework within which to deal with individual differences. With one exception (the unequivocal distinction between the two directions of communication), most variables that have been examined (perceived base rate, severity, etc.) are continuous and their levels rely on subjective judgments. Moreover, the classification of cases into discrete "contexts" is all too often after the fact and invariably arbitrary. (Are medical scenarios describing a fractured ankle or wrist following a friendly game of tennis or basketball different or indentical contexts?) Similarly, the notion that different people may have different membership functions for the same terms in a given context, as intuitive and reasonable as it may seem, requires a good theoretical foundation. The problem is not simply to show that *likely* is understood differently by different people, but to show that this variance is systematically related to other meaningful variables. For example, Weber (1988) proposed and Weber and Bottom (1989) tested descriptive risk measures for risky prospects involving precise numerical probabilities. It would be interesting to relate these measures to parameters of membership functions for verbal probabilities involved in similar prospects.

Related to issues of context and individual differences is the intriguing possibility that membership functions will provide a means of quantifying the semantic content of probability phrases and become standard tools for dealing with important psycholinguistic problems. That goal will be reached when quantitative models of context and individual differences are developed that predict meaning as represented by membership functions. Reyna's (1981) work on negation, Reagan et al.'s (1989) analysis of complementary symmetry, and Cliff's (1959) well-known study on adjective–adverb combinations are examples of other issues that might be addressed more fully if the (point numerical) representations of the verbal stimuli were to be replaced by appropriate membership functions. An example of the possible applications of these functions is provided in a study by Zwick, Carlstein, and Budescu (1987). Subjects judged the degree of similarity between pairs of terms (say, *likely* and *good chance*) and also provided judgments for membership functions of each term. Then, a large number of mathematical measures of similarity between functions were calculated and used to predict the empirically determined proximities. The results of this study could, for example, be used to determine a mathematical definition of synonymity.

Crucial to our approach is the assumption embodied in B2, that representations are treated in their vague form whenever possible and reduced to precise values only when necessary. The idea is so reasonable, and seems so adaptive, that we would argue it must be correct at some level. The underlying processes, however, are very much open to question. We provided a stochastic mechanism in the form of principle P4 and offered limited evidence in support of it, but more research is needed to establish its credibility. Particularly important is the question of how the threshold, v, is set when making decisions based on vague inputs. In work to date (Wallsten et al., 1988), v has been left as a free parameter, but for the theory to be useful, we must learn how it depends on context, task, or individual variables.

Another topic requiring additional work concerns the relationships between decision quality, preference patterns, and linguistic processing when the number of possible outcomes exceeds two. In our judgment, we understand these issues fairly well now in the case of two alternatives. The Tsao and Wallsten (1994) data, however, suggest that matters are much more complicated in the general case.

Finally, we point out that our treatment of the role of linguistic information processing in judgment and preference is incomplete in another important way. That is, we must consider how people combine multiple inputs (say, judgments from two or more experts or consultants) to arrive at their own opinion. In section II(B), we amplified assumption B2 by stating that "people treat separate judgments in their vague form when receiving *and combining them,* but they restrict attention to a narrow range of uncertainty or to a single point value when making specific decisions" (emphasis added here). Subsequently, we amplified the latter part of that statement and ignored the former. That issue is treated to some degree, however, in a recent chapter by Wallsten, Budescu, and Tsao (1994).

Acknowledgments

The order of authors is arbitrary; both authors contributed equally. Preparation of this chapter was facilitated by an Arnold O. Beckman Research Award from the Research Board of the University of Illinois and by National Science Foundation Grant SBR-9222159. We acknowledge the contribution of many students and colleagues who have been involved in various stages and aspects of the work summarized in this chapter. In particular, we thank Brent Cohen, Jim Cox, Ido Erev, Sam Fillenbaum, Claudia González-Vallejo, Amnon Rapoport, Chen-Jung Tsao, Andrea Shapiro, and Rami Zwick for their help in shaping our thinking about these issues over the years.

References

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

Behn, R. D., & Vaupel, J. W. (1982). *Quick analysis for busy decision makers.* New York: Basic.

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

Bonissone, P. P., Gans, S. S., & Decker, K. S. (1987). RUM: A layered architecture for reasoning with uncertainty. In *Proceedings of IJCAI* (pp. 373–379). Milan, Italy.

Bortolan, G., & Degani, R. (1985). A review of some methods for ranking fuzzy subsets. *Fuzzy Sets and Systems, 15,* 1–19.

Brackner, J. W. (1985). How to report contingent losses in financial statements? *Journal of Business Forecasting, 4*(2), 13–18.

Bradburn, N. M., & Miles, C. (1979). Vague quantifiers. *Public Opinion Quarterly, 43,* 92–101.

Brun, W., & Teigen, K. H. (1988). Verbal probabilities: Ambiguous, context-dependent, or both? *Organizational Behavior and Human Decision Processes, 41,* 390–404.

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

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

Budescu, D. V., & Wallsten, T. S. (1987). Subjective estimation of precise and vague uncertainties. In G. Wright & P. Ayton (Eds.), *Judgmental forecasting* (pp. 63–81). Sussex, UK: Wiley.

Budescu, D. V., & Wallsten, T. S. (1990). Dyadic decisions with verbal and numerical probabilities. *Organizational Behavior and Human Decision Processes, 46,* 240-263.

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

Camerer, C., & Weber, M. (1992). Recent developments in modeling preferences: Uncertainty and ambiguity. *Journal of Risk and Uncertainty, 5,* 325-370.

Chesley, G. R. (1985). Interpretation of uncertainty expressions. *Contemporary Accounting Research, 2,* 179-199.

Clark, D. A. (1990). Verbal uncertainty expressions: A review of two decades of research. *Current Psychology: Research and Reviews, 9,* 203-235.

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

Cliff, N. (1959). Adverbs as multipliers. *Psychological Review, 66,* 27-44.

Cohen, B. L., & Wallsten, T. S. (1992). The effect of constant outcome value on judgments and decision making given linguistic probabilities. *Journal of Behavioral Decision Making, 5,* 53-72.

Curley, S. P., & Yates, J. F. (1985). The center and range of the probability interval as factors affecting ambiguity preferences. *Organizational Behavior and Human Decision Processes, 36,* 273-287.

Degani, R., & Bortolan, G. (1988). The problem of linguistic approximation in clinical decision making. *International Journal of Approximate Reasoning, 2,* 143-162.

DeGroot, M. H. (1970). *Optimal statistical decisions.* New York: McGraw-Hill.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representations of human judgment* (pp. 17-52). Wiley, NY.

Einhorn, H. J., & Hogarth, R. M. (1985). Ambiguity and uncertainty in probabilistic inference. *Psychological Review 92,* 433-461.

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics, 75,* 643-669.

Erev, I., & Cohen, B. L. (1990). Verbal versus numerical probabilities: Efficiency, biases, and the preference paradox. *Organizational Behavior and Human Decision Processes, 45,* 1-18.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review, 101,* 519-527.

Erev, I., Wallsten, T. S., & Neal, M. (1991). Vagueness, ambiguity, and the cost of mutual understanding. *Psychological Science, 2,* 321-324.

Farkas, A., & Makai-Csasar, M. (1988). Communication or dialogue of deafs: Pitfalls of use of fuzzy quantifiers. In *Second Network Seminar of the International Union of Psychological Science.* Amsterdam: North-Holland.

Fillenbaum, S., Wallsten, T. S., Cohen, B., & Cox, J. A. (1991). Some effects of vocabulary and communication task on the understanding and use of vague probability expressions. *American Journal of Psychology, 104,* 35-60.

Gärdenfors, P., & Sahlin, N. E. (1982). Unreliable probabilities, risk taking, and decision making. *Synthese, 53,* 361-386.

Gonzales, M., & Frenck-Mestre, C. (1993). Determinants of numerical versus verbal probabilities. *Acta Psychologica, 83,* 33-51.f

González-Vallejo, C. C., Erev, I., & Wallsten, T. S. (1994). Do decision quality and preference order depend on whether probabilities are verbal or numerical? *American Journal of Psychology, 107,* 157-172.

González-Vallejo, C. C., & Wallsten, T. S. (1992). Effects of probability mode on preference reversal. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 855–864.

Grether, D. M., & Plott, C. R. (1979). Economic theory of choice and the preference reversal phenomenon. *American Economic Review, 69,* 623–638.

Hamm, R. M. (1991). Selection of verbal probabilities: A solution for some problems of verbal probability expression. *Organizational Behavior and Human Decision Processes, 48,* 193–223.

Hammerton, M. (1976). How much is a large part? *Applied Ergonomics, 7,* 10–12.

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

Jaffe-Katz, A., Budescu, D. V., & Wallsten, T. S. (1989). Timed magnitude comparisons of numerical and nonnumerical expressions of uncertainty. *Memory & Cognition, 17,* 249–264.

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

Johnson, E. M., & Huber, G. P. (1977). The technology of utility assessment. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-7,* 311–325.

Kadane, J. B. (1990). Comment: Codifying chance. *Statistical Science, 5,* 18–20.

Keynes, J. M. (1921). *A treatise on probability.* London: Macmillan.

Knight, F. H. (1921). *Risk, uncertainty, and profit.* Chicago: University of Chicago Press.

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

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). Cambridge: Cambridge University Press.

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

López de Mántaras, R., Meseguer, P., Sanz, F., Sierra, C., & Verdaguer, A. (1988). A fuzzy logic approach to the management of linguistically expressed uncertainty. *IEEE Transactions on Systems, Man, and Cybernetics, 18,* 144–151.

Mapes, R. E. A. (1979). Verbal and numerical estimates of probability terms. *Journal of General Internal Medicine, 6,* 237.

Marshall, E. (1986). Feynman issues his own shuttle report, attacking NASA's risk estimates. *Science, 232,* 1596.

Marshall, E. (1988). Academy panel faults NASA's safety analysis. *Science, 239,* 1233.

Merz, J. F., Druzdzel, M. J., & Mazur, D. J. (1991). Verbal expressions of probability in informed consent litigation. *Journal of Medical Decision Making, 11,* 273–281.

Moore, P. G. (1977). The manager's struggle with uncertainty. *Journal of the Royal Statistical Society, 140,* 129–165.

Moore, P. G., & Thomas, H. (1975). Measuring uncertainty. *International Journal of Management Science, 3,* 657.

Mosier, C. I. (1941). A psychometric study of meaning. *Journal of Social Psychology, 39,* 31–36.

Mosteller, F., & Youtz, C. (1990). Quantifying probabilistic expressions. *Statistical Science, 5,* 2–16.

Moxey, L. M., & Sanford, A. J. (1993). *Communicating quantities: A psychological perspective.* Hove, UK: Erlbaum.

Mullet, E., & Rivet, I. (1991). Comprehension of verbal probability expressions in children and adolescents. *Language and Communication, 11,* 217–225.

Murphy, A. H., & Brown, B. G. (1983). Forecast terminology: Composition and interpretation of public weather forecasts. *Bulletin of the American Meteorological Society, 64,* 13–22.

Murphy, A. H., Lichtenstein, S., Fischhoff, B., & Winkler, R. L. (1980). Misinterpretations of precipitation probability forecasts. *Bulletin of the American Meteorological Society, 61,* 695–701.

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

National Weather Service (1984). *Weather service operations manual.* Silver Spring, MD: National Oceanic and Atmospheric Administration.

Newstead, S. E. (1988). Quantifiers as fuzzy concepts. In T. Zetenyl (Ed.), *Fuzzy sets in psychology* (pp. 51–72). North-Holland: Elsevier.

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

Parducci, A. (1968). How often is often. *American Psychologist, 23,* 828.

Pepper, S. (1981). Problems in the quantification of frequency expression. In D. W. Fiske (Ed.), *New directions for methodology of social and behavioral science.* New York: Jossey-Bass.

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

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

Rapoport, A., Wallsten, T. S., Erev, I., & Cohen, B. L. (1990). Revision of opinion with verbally and numerically expressed uncertainties. *Acta Psychologica, 74,* 61–79.

Reagan, R. T., Mosteller, F., & Youtz, C. (1989). Quantitative meanings of verbal probability expressions. *Journal of Applied Psychology, 74,* 433–442.

Reyna, V. F. (1981). The language of possibility and probability: Effects of negation on meaning. *Memory & Cognition, 9,* 642–650.

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

Schkade, D. A., & Kleinmuntz, D. N. (1994). Information displays and choice processes: Differential effects of organization, form, and sequence. *Organizational Behavior and Human Decision Processes, 57,* 319–337.

Shapiro, A. J., & Wallsten, T. S. (1994). *Base-rate effects on interpreting verbally and numerically communicated probabilities.* Unpublished manuscript. University of North Carolina.

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

Stone, D. N., & Schkade, D. A. (1991). Numeric and linguistic information representation in multiattribute choice. *Organizational Behavior and Human Decision Processes, 49,* 42–59.

Sutherland, H. J., Lockwood, G. A., Tritchler, D. L., Sem, F., Brooks, L., & Till, J. E. (1991). Communicating probabilistic information to cancer patients: Is there "noise" on the line? *Social Science and Medicine, 32,* 725–731.

Svenson, O., & Karlson, G. (1986). Attractiveness of decision alternatives characterized by numerical and non-numerical information. *Scandinavian Journal of Psychology, 27,* 74–84.

Teigen, K. H. (1988a). When are low-probability events judged to be 'probable'? Effects of outcome-set characteristics on verbal probability estimates. *Acta Psychologica, 67,* 157–174.

Teigen, K. H. (1988b). The language of uncertainty. *Acta Psychologica, 68,* 27–38.

Teigen, K. H., & Brun, W. (1993). Yes, but it is uncertain: Direction and communicative intention of verbal probabilistic terms. Presented at *Subjective Probability, Utility, and Decision Making (SPUDM) 14,* Aix-en Provence, France.

Timmermans, D. (1994). The roles of experience and domain of expertise in using numerical and verbal probability terms in medical decisions. *Medical Decision Making, 14,* 146–156.

Tsao, C. J., & Wallsten, T. S. (1994). Effects of the number of outcomes on the interpretation and selection of verbal and numerical probabilities in dyadic decisions. Unpublished manuscript.

Tversky, A., Sattath, S., & Slovic, P. (1988). Contingent weighting in judgment and choice. *Psychological Review, 95,* 371–384.

Vesely, W. E., & Rasmuson, D. M. (1984). Uncertainties in nuclear probabilistic risk analysis. *Risk Analysis, 4,* 313–322.

von Winterfeldt, D., & Edwards, W. (1986). *Decision analysis and behavioral research.* Cambridge: Cambridge University Press.

Wallsten, T. S. (1990). Measuring vague uncertainties and understanding their use in decision making. In G. M. von Furstenberg (Ed.), *Acting under uncertainty* (pp. 377–398). Norwell: Kluwer.

Wallsten, T. S., Budescu, D. V., & Erev, I. (1988). Understanding and using linguistic uncertainties. *Acta Psychologica, 68,* 39–52.

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

Wallsten, T. S., Budescu, D. V., & Tsao, C. J. (1994). *Combining linguistic probabilities.* Paper presented at the Symposium on Qualitative Aspects of Decision Making, Universität Regensburg, Regensburg, Germany, July 20, 1994.

Wallsten, T. S., Budescu, D. V., & Zwick, R. (1993). Comparing the calibration and coherence numerical and verbal probability judgments. *Management Science, 39,* 176–190.

Wallsten, T. S., Budescu, D. V., Zwick, R., & Kemp, S. M. (1993). Preference and reasons for communicating probabilistic information in numerical or verbal terms. *Bulletin of the Psychonomic Society, 31,* 135–138.

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

Weber, E. U. (1988). A descriptive measure of risk. *Acta Psychologica, 69,* 185–203.

Weber, E. U., & Bottom, W. P. (1989). Axiomatic measures of perceived risk: Some tests and extensions. *Journal of Behavioral Decision Making, 2,* 113–131.

Weber, E. U., & Hilton, D. J. (1990). Contextual effects in the interpretations of probability words: Perceived base rate and severity of events. *Journal of Experimental Psychology: Human Perception and Performance, 16,* 781–789.

Wright, G. N., Phillips, L. D., Whalley, P. C., Choo, G. T., Ng, K. O., Tan, I., & Wisudha, A. (1978). Cultural differences in probabilistic thinking. *Journal of Cross-Cultural Psychology, 9,* 285–299.

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

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

Yates, J. F., Zhu, Y., Ronis, D. L., Wang, D.-F., Shinotsuka, H., & Toda, M. (1989). Probability judgment accuracy: China, Japan, and the United States. *Organizational Behavior and Human Decision Processes, 43,* 145–171.

Zimmer, A. C. (1983). Verbal versus numerical processing of subjective probabilities. In R. W. Scholz (Eds.), *Decision making under uncertainty.* Amsterdam: Elsevier.

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

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

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.

Zwick, R., & Wallsten, T. S. (1989). Combining stochastic and linguistic inexactness: Theory and experimental evaluation of four fuzzy probability models. *International Journal of Man and Machine Studies, 30,* 69–111.