Journal of Experimental Psychology: Applied
2004, Vol. 10, No. 1, 25–41

1076-898X/04/$12.00 DOI: 10.1037/1076-898X.10.1.25

You Say "Probable" and I Say "Likely": Improving Interpersonal Communication With Verbal Probability Phrases

Tzur M. Karelitz and David V. Budescu
University of Illinois at Urbana–Champaign

When forecasters and decision makers describe uncertain events using verbal probability terms, there is a risk of miscommunication because people use different probability phrases and interpret them in different ways. In an effort to facilitate the communication process, the authors investigated various ways of converting the forecasters' verbal probabilities to the decision maker's terms. The authors present 3 studies in which participants judged the probabilities of distinct events using both numerical and verbal probabilities. They define several indexes of interindividual coassignment of phrases to the same events and evaluate the conversion methods by comparing the values of these indexes for the converted and the unconverted judgments. In all the cases studied, the conversion methods significantly reduced the error rates in communicating uncertainties.

Decision making under uncertainty invariably involves the use of likelihood assessments. People often rely on external forecasters (experts or laypeople) to provide these assessments and to help them adjust to the uncertain environments. These assessments are vital to all levels of decision processes. They are used in answering mundane questions such as "What are the chances of rain tomorrow?" as well as making complex business, financial, academic, judicial, medical, and political decisions.

Likelihood judgments can be communicated by using numerical probabilities (e.g., there is a 40% chance that *X* will occur) or by using probability phrases (e.g., it is not very probable that *X* will occur). Throughout this article, the generic term *probability phrase* refers to probability words such as *likely* and more complex probability statements such as *extremely high chance.* Although probability phrases are typically considered to be the natural method of choice (Wallsten, Budescu, Zwick, & Kemp, 1993), they suffer from a common deficiency; people tend to interpret them in different ways. The different meanings that people associate with verbal probabilities can result in misunderstandings and errors in communication. Consider for example the communication between a financial analyst and an investor. If the analyst asserts that a certain company is likely to exceed its yearly projected profits, the investor, who interprets "likely" to mean the 80%–85% vicinity, might consider adding this company to her portfolio. The investor may be surprised, however, to learn that in the analyst's mind this word describes probabilities in the 55%–65% range.

There are quite a few well-documented examples of disastrous decision errors caused by differential understanding and usage of verbal probabilities. Marshall (1986) describes the process of creating a verbal probability scale for the engineers at the National Aeronautics and Space Administration to communicate risk assessments to their supervisors. Marshall suggested that this process led to unrealistic assessment of space shuttle parts failure. One such part was the infamous O-ring that caused the Challenger's explosion and the death of seven astronauts. Similarly, Wyden (1979) discussed the costly consequences of the different interpretations of the term *fair chance* by the U.S. Joint Chiefs of Staff and the Central Intelligence Agency in the Bay of Pigs invasion in 1961.

The present research explores ways to reduce the prevalence and magnitude of such communication errors. The main thesis of our work is that it is possible, and indeed feasible, to derive conversion schemes that would facilitate and improve the quality of communication among people who use different lexicons and interpret the probability terms in different ways. These conversions take advantage of previous work on the psychological principles underlying the representation, use, and communication of probability terms (see Budescu & Wallsten, 1995; Wallsten, Budescu, & Tsao, 1997).

Next we summarize some of these results, with special attention to those aspects that are directly relevant to the conversion schemes; discuss the details of the conversions; and report the results of three experiments designed to test the efficacy of these conversion methods.

General Principles Underlying the Use of Vague Probabilistic Terms

In reviewing the empirical literature on how humans process vague, especially linguistic, information about uncertainty, Budescu and Wallsten (Budescu & Wallsten, 1995; Wallsten et al.,

Tzur M. Karelitz and David V. Budescu, Department of Psychology, University of Illinois at Urbana–Champaign.

This study was supported by National Science Foundation Award NSF SES 9975360. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation.

We thank Professor Thomas S. Wallsten for many insightful comments on an earlier version of this article.

Correspondence concerning this article should be addressed to David V. Budescu, Department of Psychology, University of Illinois, 603 East Daniel Street, Champaign, IL 61820. E-mail: dbudescu@cyrus.psych.uiuc.edu

1997) formulated the following two background assumptions and six principles.

Background Assumptions

1. B1. Except in very special cases, all representations of uncertainties are vague to some degree in the minds of the originators and in the minds of the receivers. This assumption implies that people consider both the event definition and the database when deciding how to represent their own judgment and when interpreting a communication from someone else. A corollary of B1 is that numbers should also show some degree of imprecision (e.g., Budescu, Weinberg, & Wallsten, 1988; Mullet & Rivet, 1991; Windschitl & Weber, 1999).

2. B2. People use the full representation of a phrase 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 in part whether, and how, an individual resolves representational vagueness. Put crudely, one can have imprecise opinions but cannot take imprecise actions.

Principles

1. P1. Probability phrases can be meaningfully scaled by means of membership functions. These functions (see details below) provide meaningful representations of an individual's understanding of a phrase within the context that it is being used and allow predictions of behavior (e.g., Budescu, Karelitz, & Wallsten, 2003; Budescu & Wallsten, 1990; Jaffe-Katz, Budescu, & Wallsten, 1989; Wallsten, Budescu, Rapoport, Zwick, & Forsyth, 1986).

2. P2. The location, spread, and shape of membership functions vary over individuals and depend on context and communication direction. Context effects on the interpretation of probability terms are pervasive. Many studies have examined the effects of single contextual variables on the meaning of target probability phrases. For example, Fischer and Jungermann (1996), Pepper and Prytulak (1974), Wallsten, Fillenbaum, and Cox (1986), and Weber and Hilton (1990) have all shown that the numerical interpretation of a term used to describe the chances of a given event occurring covaries positively with the event's perceived base rate. Furthermore, Weber and Hilton found that the mean probability interpretations of phrases were positively correlated with event severity. With regard to direction of communication, Budescu and Wallsten (1990), and Fillenbaum, Wallsten, Cohen, and Cox (1991) observed that recipients of the verbal forecasts interpreted probability phrases as being less extreme and more imprecise than intended by the communicators.

3. 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. Most people prefer to communicate their opinions verbally under conditions of imprecision (e.g., Brun & Teigen, 1988; Erev & Cohen, 1990; Wallsten et al., 1993). However, most people prefer receiving numerical information (Brun & Teigen, 1988; Erev & Cohen, 1990; Fischer & Jungermann, 1996; Wallsten et al., 1993). In fact, the modal pattern of responses in Wallsten et al.'s (1993) survey and Erev and Cohen's (1990) study was preference for communicating to others verbally and receiving information from others numerically. The reverse preference pattern was virtually never reported. Those who prefer to communicate verbally claim that it is more natural, more personal, and easier, and those who prefer to receive numerical information claim that it is more precise (Wallsten et al., 1993).

The three remaining principles are less relevant to the present research, as they deal primarily with the processes used by individuals when making decisions with imprecise probabilities and less with the understanding, interpretation, and communication of forecasts. The interested reader should consult the chapters by Budescu and Wallsten (1995) and Wallsten et al. (1997).

The Sources of Interindividual Communication Errors

These principles highlight three major reasons for problems in communication with verbal probabilities: (a) People prefer to express their uncertainties verbally; (b) people use diverse lexicons to describe uncertainties; and (c) people vary in their (numerical) interpretations of the linguistic terms.

Preference for Using Verbal Terms

Undoubtedly, one could reduce the communication errors that result from the different meanings people attribute to probability phrases by avoiding words and using only numerical probabilities. In fact, this is a standard recommendation in the field of decision analysis (e.g., Behn & Vaupel, 1982; von Winterfeldt & Edwards, 1986). However, as described in P3, most people prefer expressing their uncertain beliefs with verbal probabilities. Verbal communication is preferred unless the underlying opinions are based on solid quantitative evidence about aleatory events (e.g., Olson & Budescu, 1997) or there is some clear incentive to be precise (Erev, Wallsten, & Neal, 1991).

Several researchers (e.g., Beyth-Marom, 1982; Hamm, 1991; Mosteller & Youtz, 1990) have suggested that using a fixed list of terms could reduce errors in communication of uncertainty and could consequently improve decision outcomes. Many studies have disputed this claim by showing the effects of context on the interpretation of probability terms (see P2). Another drawback of standardized verbal scales is the difficulty of most people to suppress the meanings they normally associate with these terms. For example, Wallsten, Fillenbaum, and Cox (1986) demonstrated that National Weather Service forecasters could not transfer the imposed meanings of some words from one domain to another. These forecasters were trained to use a fixed set of phrases to communicate their estimates for meteorological events. However, when the same phrases were presented in a different domain (interpreting medical advice), the forecasters did not interpret them

as they were trained to, indicating that one cannot legislate meaning any more than imposing a certain communication mode.

Diversity of the Verbal Lexicon

Over their lifetime, people develop preferences for specific terms and tend to avoid others. Consequently, when they need to choose terms to describe uncertainty, different individuals will spontaneously pick different words. Budescu et al. (1988) reported that the 20 participants in their study produced 111 distinct probability phrases when asked to describe 11 different, graphically displayed probabilities (see also Erev & Cohen, 1990; Zwick & Wallsten, 1989). In a recent review of over 25 published studies, we found that the researchers in this domain have used more than 100 different probability phrases (and a similar number of frequency phrases)! The only way to accommodate this diversity is to let people use their favorite terms.

Variability in the Meaning of Verbal Probabilities

Numerous researchers have studied how different phrases are mapped and converted into numerical probabilities. Participants were asked to rank order, compare, or simply convert phrases into numbers, and vice versa, in various ways and under various contexts (see Budescu & Wallsten, 1995, for a partial list of these studies). The most robust finding is that between-individuals variability is larger than the variance observed within individuals when judging the same terms (Beyth-Marom, 1982; Budescu & Wallsten, 1985; Clarke, Ruffin, Hill, & Beamen, 1992; Johnson, 1973; Mullet & Rivet, 1991; Reagan, Mosteller, & Youtz, 1989). This suggests that most people perceive the meanings of verbal probabilities consistently and reliably but differently from each other. Budescu and Wallsten (1985) claimed that this state of affairs is most likely to induce an illusion of valid communication. Numerous studies have found considerable interpersonal variability in interpreting probability phrases not only among lay people but among experts within their professional domains, such as military intelligence (Beyth-Marom, 1982), accounting (Chesley, 1985), and especially medicine (Bryant & Norman, 1980; Kong, Barnett, Mosteller, & Youtz, 1986; Mapes, 1979; Merz, Druzdzel, & Mazur, 1991; Nakao & Axelrod, 1983; Sutherland, Lockwood, Tritchler, & Sem, 1991). In fact, many of these studies caution against the use of probability phrases in their respective domains.

Reducing Interindividual Communication Errors

In the present experiments we examined a new way to reduce the errors in communication of uncertainties by devising methods to convert one person's lexicon to another person's. This framework allows people to use their favorite phrases. The recipients of these estimates do not necessarily receive the communicators' original phrases but rather a mapping of these phrases to terms in their own lexicon. The intermediate conversion module is analogous to a French-English dictionary; French speakers can use it to translate their opinions into English and thus be understood by English speakers. We intend to show that converting verbal probabilities facilitates communication of uncertainties.

Interpersonal Conversion of Probability Phrases

Many of the conversion methods that we implemented and tested rely on the assumption that the meaning of probability phrases can be represented by their membership functions. Before reviewing the specific methods, we describe this view.

Membership functions. Our key theoretical assumption is that probability phrases are vague concepts and that different numerical probabilities in the [0–1] range are represented to various degrees in these concepts. Wallsten, Budescu, et al. (1986) and Rapoport, Wallsten, and Cox (1987) used the notion of membership functions, adopted from fuzzy set theory (Zadeh, 1965) to implement this approach. The membership function of any given phrase assigns a number to each value on the probability line [0,1] that represents its degree of membership in the concept defined by the phrase.

The membership function approach to probability phrases was carefully validated and applied successfully in several instances (Budescu et al., 2003; Budescu & Wallsten, 1990; Fillenbaum et al., 1991; Jaffe-Katz et al., 1989). Most functions are single peaked, but there are also a few monotonic functions, decreasing from probabilities close to 0 for low probability phrases or increasing to probabilities close to 1 for high probability phrases. Membership values are bounded (Norwich & Turksen, 1984; Wallsten et al., 1986) such that memberships of 0 denote probabilities that are absolutely not in the concept, and memberships of 1 denote elements that are perfect exemplars of the concept. All other positive values represent intermediate degrees of membership. Membership functions are nonnegative, but there are no special constraints on the shape, area under the function, or other properties, and membership functions are not probability density functions. Most functions cover a wide range of probabilities and can be skewed in one direction or another. The probability (or the mean of the range of probabilities) at which the function reaches its maximum value is the membership function's peak. The peak will play an important role in some of the methods studied below. Figure 1 presents three hypothetical functions varying in peak location, shape, skewness, height, and spread.

Membership functions can be used to obtain phrase-to-number mappings of verbal probability lexicons. Whereas every person may use a completely different lexicon to express uncertainty, the functions map their vocabularies onto a common scale—the probability line. This common representation can be used to match phrases from different lexicons.

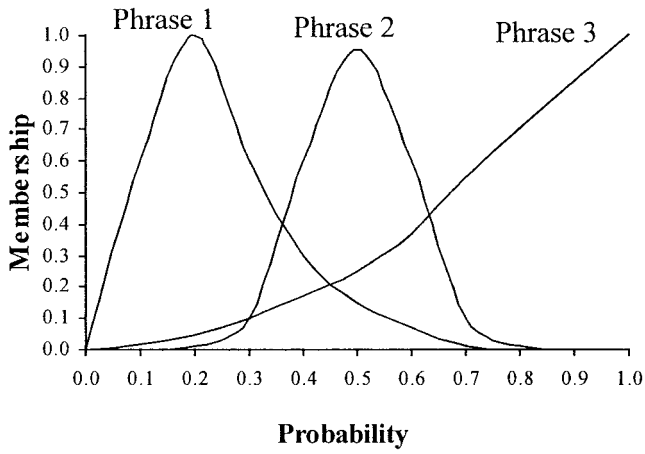

Figure 1. Membership functions for three hypothetical phrases.

Conversion methods. Conversion methods are designed to match a phrase from one person's lexicon to every phrase in his or her partner's lexicon on the basis of a specific well-defined criterion. Consider two people labeled *i* and *j*. Sometimes the mapping is from *i* to *j*, and sometimes vice versa. Because some of the conversion methods are asymmetric, this distinction is important in some instances, and we will indicate when this is the case. Specific phrases from a certain lexicon are labeled w_{jm} (phrase *m* of person *j*) or w_{in} (phrase *n* of person *i*). In this article we study the following four criteria (conversion methods) for selecting a phrase from person *j*'s (the communicator) lexicon to match a specific phrase from person *i*'s (the recipient) lexicon.

1. *ABSDEV (ABSolute DEViation):* Calculates the sum of absolute deviations between the membership values of w_{in} and the membership values of each phrase in person *j*'s lexicon and selects w_{jm}, the phrase for which the sum is smallest.

2. *PRO (Peak Rank Order):* Chooses w_{jm} such that its rank order (as inferred from the peaks of person *j*'s membership functions) matches the rank order of w_{in} (as inferred from the peaks of person *i*'s membership functions). In cases where two or more phrases peak at the same probability, they are all assigned the same mean rank.

3. *DPEAK (Distance between PEAKs):* Chooses w_{jm} such that the distance between the location of its peak and that of w_{in} is less than that of any other phrase in person *j*'s lexicon.

The last method does not require elicitation of membership functions but requires people to rank order the words in their subjective lexicons.

4. *RANK:* Chooses w_{jm} such that its rank (as given by person *j*) matches the rank of w_{in} (as given by person *i*).

These descriptions apply to the cases in which people use the same number of phrases in their lexicon. When people use different numbers of phrases, some of the methods require further refinements. In adjusting the RANK method we divide each rank given by an individual by the total number of phrases in that individual's lexicon and rescale all ranks between 0 and 1. Consequently, phrases from a list of size 9 are ranked 1/9, 2/9, 3 9 ,. . . , 9/9; whereas phrases from a list of size 11 are ranked 1/11, 2/11, 3/11, and so on. Accordingly, the adjusted conversion method, ADJRANK (ADJusted RANK), chooses w_{jm} such that its adjusted rank (as given by person *j*) is the closest (smallest absolute difference) to the adjusted rank of w_{in} (as given by person *i*).

Another rank-based method is peak rank order (PRO), in which the adjusted ranks of the membership function peaks are treated as continuous rather than discrete. For example, if a person uses five phrases, we define five equal intervals: 0–0.2, 0.2–0.4, 0.4–0.6, 0.6–0.8, and 0.8–1. Similarly, phrases from a lexicon of size three are mapped into three intervals: 0–0.33, 0.34–0.67, and 0.68–1. Finally, phrases are matched according to the level of overlap between the intervals covered by their adjusted peak ranks over the full [0–1] interval. Accordingly, the adjusted conversion method, ADJPRO (ADJusted peak rank order), chooses w_{jm} such that its adjusted rank order (as inferred from the peaks of person *j*'s membership functions) has maximal overlap with the adjusted rank order of w_{in} (as inferred from the peaks of person *i*'s membership functions).

The criteria underlying the various methods do not necessarily lead to unique solutions. If more than one phrase matches the criterion set by a certain method, the result is a one-to-many mapping. In such cases, one phrase from person *i*'s lexicon is mapped into more than one phrase in person *j*'s lexicon. A one-to-many mapping implies that phrase w_1 can be substituted, for instance, by w_2, w_3, or both. Of course, one-to-many mappings are also very common in translation between languages as one can easily verify by browsing any dictionary.

Some conversion methods (e.g., RANK) are symmetric, whereas others (e.g., ABSDEV and DPEAK) are asymmetric. A method is symmetric if w_{im}, the *m*th phrase used by person *i*, is mapped to w_{jn}, the nth phrase used by person *j*, and w_{jn} is mapped to w_{im}. A method is asymmetric if w_{im} is mapped to w_{jn} but w_{jn} is mapped to w_{ik} (where $w_{ik} \neq w_{im}$). Asymmetric conversions are also common in regular translation between languages. Thus, converting phrases from person *i* to person *j* may lead to different results than converting from person *j* to person *i*.

The Present Experiments

Our primary goal was to test whether the decision makers understand the forecasters better when the forecasters' estimates are converted to the decision makers' lexicons. In other words, we tested whether the decision makers interpret the forecasters' estimates similarly to what the forecasters had in mind when giving these estimates. We claim that if two people describe the likelihood of an event using the same phrase, they interpret its underlying uncertainty in similar fashion. It follows that if we can develop systematic methods to convert one person's estimates to phrases that match the other persons' estimates of the same events, we can increase the accuracy of their communication.

We hypothesized that communication errors could be reduced by converting the probability lexicon of person *i* to that of person *j*. Consider an imaginary continuum of communication quality. Its lower bound is the level of error observed in unconverted, verbal discourse, and the upper bound is the level of error observed when communicating with numerical probabilities. Our expectation was that communication with converted phrases would be superior to unaided communication and closer in quality to the numerical case. Three experiments were designed to test this hypothesis. Each participant was presented with a set of stimuli, and we collected participants' verbal and numerical probability judgments for the occurrence of a common set of aleatory events (see Wallsten, Budescu, Erev, & Diederich, 1997). We also elicited membership functions for the verbal probabilities used by the participants in the judgment task. These functions were used as the main tool for matching the phrases of person *i* to those of person *j*, generating the conversion tables. A unique feature of the third experiment in this series is the use of speakers of six languages who made their judgments in their respective native languages. This experiment illustrates the generality and flexibility of the proposed methodology.

The quality of conversion was evaluated by comparing the number and magnitude of errors in the converted communication

with their hypothesized lower and upper bounds (verbal and numerical judgments, respectively). To this end, we calculated several coassignment indexes of phrases (or numbers) to events for each pair of participants (to be described in detail in the *Results* section). These indexes measured the degree of similarity between the phrases or numbers used by person *i* and those used by person *j*. The second step was to convert the lexicon of every individual to the lexicon of all the other individuals and to calculate the same indexes by using the converted terms. The efficacy of the conversion methods was assessed by the degree to which these coassignment indexes matched the hypothesized pattern.

Experiment 1

Method

Participants

Twenty people volunteered to participate in this experiment (2 participants were discarded because of corrupted data). All were students or employees of the University of Illinois at Urbana–Champaign and were paid $8 for their participation. The participants were native English speakers, 5 women and 13 men (*M* age = 24.30 years, *SD* = 7.33).

Materials and Procedure

The experiment was computer controlled, lasted about 1 hr, and consisted of three tasks:

1. Selection of a verbal probability lexicon and ranking of the words.
2. Elicitation of membership functions for the selected phrases.
3. Numerical and verbal likelihood estimation of precise, graphically displayed events.

First, participants were asked to create a list of 6–11 phrases by selecting combinations of words and semantic operators (modifiers, quantifiers, negators, intensifiers, etc.) from two lists or by simply typing in phrases. Participants were instructed to select phrases that spanned the whole probability range and that they also use in their daily lives (some participants used frequency phrases instead of probabilities, we used these frequency phrases in the analysis, nevertheless). The list of phrases and semantic operators was derived from an intensive review of papers concerning verbal probabilities. Three phrases were preselected for all participants: Certain, Even odds, and Impossible.

After the selection of their lists, participants were asked to rank the phrases they chose in ascending order. In the second task, membership functions were elicited using a multistimuli method validated by Budescu et al. (2003). Each phrase from the participants' lists was presented with a set of 11 probabilities ranging from 0% to 100% in increments of 10%. For each phrase, participants judged the degree to which the target phrase captured the intended meaning of each of the 11 numerical probabilities. All judgments were made using a bounded scale, ranging from *not at all* to *absolutely*. If the participants believed that a target phrase was a perfect description of a specific numerical probability, they were told to move the pointer to *absolutely*. If they believed the number did not describe the phrase at all, they were told to move the pointer to *not at all*.

In the third task participants had to judge the likelihood of several events using numerical probabilities in one presentation and verbal probabilities in the other. The events were defined by two sets of 19 circles presented to each participant in random order. Each set covered all probabilities from 5% to 95% in 5% increments but used different configurations of shaded sections. For each trial, participants saw a circular target, partially shaded. Their task was to assess the likelihood that a dart aimed at the center of the circle would hit the shaded area. The numerical judgments were made by selecting one value from a list of 21 probabilities, ranging from 0% to 100% in increments of 5%. Participants made the verbal judgments by selecting up to four phrases from their lexicons. They were instructed to order the phrases according to the degree of their appropriateness in describing the event. Thus, the most appropriate phrase to describe the likelihood of each stimulus was considered the first-choice phrase.

Results

Descriptive Statistics of the Numerical and Verbal Judgments

The average number of phrases per participant, including the 3 fixed phrases was 9.6 (10 participants chose a set of 11 phrases, and the rest chose between 7 and 10 phrases). Overall, the 18 participants chose 71 different phrases! In fact, 40 phrases were chosen by only 1 participant and 28 phrases were shared by 2 participants or more. A list of these phrases appears in the Appendix.

Each participant made 38 verbal and 38 numerical judgments, resulting in 684 judgments per modality. In 445 of the 684 (65%) occasions in which verbal judgments were requested, participants used more than one phrase to describe an event. In most cases, participants used only two phrases, but in 17% of the cases they included more. On the average, participants used 77% of the phrases on their lists as their first choice and they did not use 12% of the selected phrases.

The reliability of the judgments can be measured by the degree of consistency between the repeated judgments of the same events. For the numerical judgments the mean absolute deviation between the repeated judgments was .053 (*SD* = .006). For the verbal judgments the reliability estimation is complicated by the fact that on the two separate occasions the participants could have used different numbers of phrases. We calculated the mean consistency of use of each phrase, that is, the average fraction of phrases in the longer of the two lists that was also included in the shorter list (when the two lists were equal in size the designation of short and long was arbitrary). This measure is bounded by 0 (no common phrases in the two lists) and 1 (the two lists are identical in length and composition). The mean consistency was 0.65 (*SD* = 0.16), and 54% of the first-choice phrases selected were identical.

The participants' judgments were accurate in both modalities. The mean absolute deviation between the numerical judgments and the proportion of shaded area of the circle was 0.039 (*SD* = 0.064). To compare the accuracy of the two response modes, we calculated the (Kendall τ_b) rank-order correlation between the shaded areas and the judgments (in the verbal judgment case we used the rank of the phrase as given by the participants in the first task). The median correlation for the numerical judgment was almost perfect (.96) with very little variation among participants (*SD* = .04). The median correlations for the verbal judgment were also very high (.90 for the first word chosen and .89 across all words used), but there was considerably more variation among participants (*SD* = .30 and .25, respectively).

Assessment of Conversion Methods

Prior to assessing the quality of conversion, we examined two measures that reflect, at least indirectly, the quality and feasibility of the conversion methods. The first is the prevalence of one-to-many mappings. The middle column in Table 1 includes the average number of mappings per phrase. This index is 1, if and only if, a conversion method converts each phrase of the communicator into a unique phrase of the receiver (as in ABSDEV). Higher values indicate various levels of one-to-many mapping occurrences. The second index measures the degree to which the receiver's lexicon is used by the conversion method. When a method is mapping all the communicator's words into a subset of the receiver's lexicon, some of the receiver's phrases are not used in the phrase-matching process. The right column in Table 1 shows the average proportion of the receiver's phrases used in the conversion of the communicator's complete lexicon. This measure peaks at 100% if for each pair of participants a method uses all possible phrases (like ADJPRO).

Table 1 shows that methods that did not produce one-to-many mappings (e.g., ABSDEV) failed to use the full lexicon. Methods that used the full lexicon (e.g., ADJPRO) yielded a large number of one-to-many mappings. However, ADJRANK had the desirable features of both a wide conversion range and almost no one-to-many mappings. For this analysis we combined the responses of the two replications of the same stimuli. In 46% of the cases, participants selected distinct first-choice phrases. In those cases, both phrases were considered first choices. When combining responses from the two replications, we discarded duplicate phrases. For example, if w_1 was selected to describe event X in Set 1, and w_1 and w_2 were selected to describe the same event X in Set 2, the combined response set included w_1 and w_2. Furthermore, both w_1 and w_2 were considered first choices. For 92% of the stimuli, participants had three phrases or less in the combined set.

The translation methods were evaluated by comparing the verbal or numerical responses of each pair of participants with the same stimulus at two levels: first pair level—comparison of the first pairs of (verbal or numerical) judgments (i.e., only first-choice phrases); all pairs level—comparison of all pairs of judgments without any consideration of their appropriateness rankings.

Table 1
One-to-Many Mappings and Conversion Range Indexes for Experiment 1

Method[a]	Average number (and *SD*) of mappings per phrase		Conversion range index (%)[b]
ABSDEV	1.00	(0.07)	60
ADJPRO	2.9	(1.30)	100
DPEAK	1.6	(0.77)	83
ADJRANK	1.0	(0.18)	93

[a] Conversion methods are based on the following criteria: ABSDEV matches phrases by minimal absolute deviation between membership functions; ADJPRO matches phrases with identical adjusted peak ranks; DPEAK matches phrases by minimal distance between their peaks' locations; and ADJRANK matches phrases with identical adjusted subjective rank order. [b] Average percentage of phrases in the receivers' lexicon used in the conversion (per participant).

The first level of analysis is stricter than the second level because fewer comparisons are made, thus there is a smaller chance of matching a pair of phrases. The analysis was performed by using each method and by designating one participant in each pair as the communicator and the other as the recipient. We repeated the procedure after reversing the roles within each dyad. Thus, each person served both as the communicator and as the recipient, resulting in 306 (18×17) pairs of participants. All the phrases used by the communicator to describe each stimulus were converted to the recipient's lexicon. The comparison was done between the recipient's original responses and the communicator's converted responses for each stimulus. To quantify the level of interpersonal agreement we defined two indexes of coassignment:

1. PIA: Proportion of identical assignments—the proportion of comparisons in which both participants assigned the same phrase.

2. PMA: Proportion of minimal agreement—the proportion of stimuli for which both participants assigned at least one common phrase.

Both the PIA and PMA indexes ranged from 0 to 1, such that higher values indicated stronger agreement. PIA is stricter than PMA (i.e., PIA $\leq$ PMA) because it weights the agreement by the number of comparisons made (see more on the attributes of these indexes in the *Discussion* section). To understand the calculation of the two measures and to fully appreciate the differences between them, consider the responses of two imaginary judges (A and B) across all the presentations of a spinner that was 30% shaded. The numerical judgments of Participant A on both presentations were .30. Her responses on the first verbal presentation were *improbable*, *unlikely*, and *poor chance*, in this order, and on the second occasion she responded with *unlikely* and *poor chance*. When presented with the same stimulus, Judge B responded by using the numbers .30 and .35, and on the two verbal occasions he used *unlikely* and *improbable* for Set 1 and *unlikely*, *improbable*, and *doubtful* for Set 2. When calculating agreement at the first pair level, we considered two pairs: improbable (Judge A) versus *unlikely* (Judge B), and *unlikely* (Judge A) versus *unlikely* (Judge B). Only one of these two was identical, so PIA = 1/2. However, PMA = 1 because there was at least one identical assignment. When calculating agreement at the all-pairs level, we analyzed all nine distinct combinations of the three distinct words used by Judge A (improbable, unlikely, and good chance) and the three words selected by Judge B (improbable, unlikely, and doubtful). Because the two judges used at least one common term, PMA = 1. However, out of the nine distinct combinations, only 2 were identical (*improbable* and *unlikely*), therefore PIA = 2/9. When comparing the numerical judgments, we considered two pairs: .30 (Judge A) versus .30 (Judge B) and .30 (Judge A) versus .35 (Judge B). Thus PIA = 1/2 and PMA = 1.

The combination of the two levels of aggregation (first pair and all pairs) and the two coassignment indexes (PIA and PMA) yielded four different measures. Among them, PIA at the first pair level is the strictest, and PMA at the all-pairs level is the most lenient, with all-pairs PIA and first-pair PMA in between. The average coassignment indexes (and their *SD*s) for the numerical

judgments, the unconverted verbal judgments, and the four conversion methods are presented in Table 2.

We analyzed each of the coassignment indexes, treating the communication mode (unaided verbal or numerical discourse and aided verbal discourse) as repeated measures. First we averaged the values of the relevant coassignment index for each forecaster across all decision makers. These averages served as dependent variables in the respective analysis of variance (ANOVAs) with the individual forecaster as the unit of analysis. The global F tests were significant for all four indexes, as evident in Table 3. We tested the significance of all 15 pairwise differences among the various conversion methods by means of the Tukey's honestly significant difference (HSD) procedure. We found that the coassignment indexes of the unaided verbal communication were significantly lower than all other conversion methods in all cases. Coassignment indexes for numerical communication were significantly higher than all conversion methods only for the all-pairs PIA index. In general, the indexes of the various conversion methods were not significantly different from each other with the exception of PRO, which had significantly higher PMA indices than all other methods, and RANK, which had significantly higher first-pair PIA indices. Power analysis (e.g., Cohen, 1998) indicates that these tests have power greater than .90 of detecting medium-size effects of $d = 0.5$, at the $\alpha = .05/15 = .0033$ level.

The results support our predictions; verbal discourse was consistently the most error-prone communication method, and in most cases numerical discourse was the least error prone. Most important, all the conversion methods significantly outperformed the unconverted communication at all levels of analysis, and in all cases the differences uncovered were large according to Cohen's classification. (In fact, $d > 1$ for all differences between conversion methods and the unaided verbal communication.)

Table 2
Average (and Standard Deviation) of Coassignment Indexes for Experiment 1

Level	ABSDEV	ADJPRO	DPEAK	ADJRANK	VJ	NJ
			PIA			
FP	.23	.19	.22	**.27**	.05	.29
	(.12)	(.07)	(.10)	(.12)	(.05)	(.07)
AP	.21	.18	.21	**.23**	.05	.29
	(.09)	(.06)	(.08)	(.09)	(.03)	(.07)
			PMA			
FP	.39	**.76**	.54	.52	.11	.68
	(.18)	(.18)	(.19)	(.22)	(.11)	(.14)
AP	.67	**.90**	.76	.78	.26	.68
	(.23)	(.13)	(.20)	(.21)	(.17)	(.14)

Note. $N = 306$ participant dyads. The indexes are as follows: PIA = proportion of identical assignment; PMA = proportion of minimal agreement. Cells in bold letters indicate the best results for each analysis. Levels of analysis are as follows: FP, comparison between the pair of first choice responses; AP, comparison between all pair of responses. Conversion methods are based on the following criteria: ABSDEV matches phrases by minimal absolute deviation between membership functions; ADJPRO matches phrases with identical adjusted peak ranks; DPEAK matches phrases by minimal distance between their peaks' locations; ADJRANK matches phrases with identical adjusted subjective rank order. The baselines are as follows: VJ, unaided verbal judgments; NJ, unaided numerical judgments.

Discussion

The results of Experiment 1 illustrate that the proposed conversion methods are highly beneficial and can significantly improve the quality and precision of interpersonal communication. The procedure implemented required each participant to select his or her own favorite set of phrases at the beginning of the session and to use them throughout the experiment. This was done to accommodate the large individual differences in people's spontaneous preferences for phrases. Of course, we do not mean to imply that people could not use other words. Nor do we believe that they would necessarily always select exactly the same phrases. We simply assume that this option provides them maximal flexibility and convenience. To achieve this goal, we allowed them to select phrases from a long list (which was based on numerous previous studies) or to include their own idiosyncratic choices.

All the conversion methods yielded higher agreement indexes than the unconverted judgments. On average, the agreement indexes of the conversion methods were about four times higher than those observed with unconverted communication. Although numerical discourse yielded the best PIA indexes, in most cases the conversion methods outperformed the numerical judgments in terms of their PMA values. Two conversion methods had better results than the others; ADJRANK had the largest PIAs, and ADJPRO had the largest PMA indexes.

The differences between the first-pair and the all-pairs levels are due to the fact that the latter involves a much larger number of comparisons. This affects the number of matching and nonmatching responses. It is interesting that the PIA indexes were highly similar at the two levels of aggregation, but the PMA indexes were consistently higher at the all-pairs level. When calculating PIA, the increase in the number of matching responses was offset by a similar increase in the number of nonmatching responses. In other words, PIA penalizes for a larger number of comparisons in the all-pairs level, so the PIA values were generally similar for the first-pair and the all-pairs levels. The effects of the larger number of comparisons in the all-pairs level were more pronounced in the PMA measure that does not penalize for the increased proportion of nonmatching responses. Therefore, the average PMA index increased for every stimulus to which a pair of participants gave at least one identical response.

Experiment 2

The primary goal of Experiment 2 was to replicate the findings of Experiment 1 and to address several methodological issues. The first two issues are related to the choice of phrases for the participants' lexicons, and the other two issues are related to the implementation of the conversions methods.

In the first study the participants were free to choose any number of phrases, and indeed, there were differences in the sizes of their lists. This required the adjustment of some conversion methods and increased the number of one-to-many mappings. In Experiment 2, participants were instructed to choose exactly 11 phrases (the modal set size in Experiment 1). The second issue concerns the use of predetermined phrases in the lexicons. Originally, we used these phrases to anchor the list of phrases to be used by the participants. This restriction might have affected the level of agreement and biased the results. Therefore, in Experiment 2, no predetermined phrases were imposed on the participants.

Table 3
Analysis of Variance for Communication Mode for Experiment 1

	First-pair level PIA			All-pairs level PIA			First-pair level PMA			All-pairs level PMA		
Source	*df*	*F*	Cohen's *f*	*df*	*F*	Cohen's *f*	*df*	*F*	Cohen's *f*	*df*	*F*	Cohen's *f*
Mode	5	85.59*	2.00	5	102.49*	2.18	5	138.60*	2.53	5	134.06*	2.50
Error	85	(0.001)		85	(0.001)		85	(0.007)		85	(0.007)	

Note. Values in parentheses represent mean square errors. PIA = proportion of identical assignment. PMA = proportion of minimal agreement.
* $p < .05$.

The third issue concerns the strategy of matching pairs. When all possible pairs of participants are considered, the global level of agreement is based on highly interdependent measures. Experiment 2 was designed to overcome this problem by (a) increasing the total sample size and (b) creating several small teams of participants (four people in each). This team size reflects more realistically the number of decision makers in many real-life committees and other interacting decision-making teams. Indexes of agreement were calculated within the teams and were then averaged across them.

Finally, although one-to-many mappings make perfect sense in any translation process, they nevertheless pose technical difficulties and impede fluent discourse. The simplest way to deal with this problem is to randomly select one of many possible translations, implicitly assuming that all these translations are interchangeable. Experiment 2 tested this randomized selection and investigated its implications.

Method

Participants

Thirty-two people volunteered to participate. They participated in the study individually, but for the purpose of the analysis they were divided into eight teams of four people (one participant was discarded because of corrupted data, so one of the teams had only three people). All participants were native English speakers, 20 were women and 11 were men (M age = 20.20 years, *SD* = 2.70). The participants were paid $12 to complete two sessions, 3–7 days apart.

Materials and Procedure

The procedure used was similar in most respects to the one used in the first experiment. In Session 1, the participants were first asked to select and rank order exactly 11 phrases (with no predetermined phrases). The second task was to elicit the membership functions for the 11 selected phrases. The last task was verbal and numerical likelihood judgments of 21 graphical stimuli depicting probabilities that ranged from 0% (*almost*) to 100% (*almost*) in steps of 5% (Experiment 1 did not include these two endpoints). Seven randomly chosen stimuli were repeated to test the consistency of the participants' judgments. The tasks in Session 2 are not relevant to this article and therefore are not discussed further.

Results

Descriptive Statistics of the Numerical and Verbal Judgments

In total, participants chose 146 different phrases. Eighty-one phrases were each chosen by no more than 1 participant, 20 phrases were selected by 2 participants each, and 45 phrases were used by more than 2 participants.

Each participant made 28 verbal judgments and 28 numerical judgments, resulting in 868 judgments per modality. Sixty-one percent of the 868 verbal judgments included more than one phrase to describe the target stimulus, and 23% used more than two phrases. On average, participants used 91% of the phrases in their lexicons as the first choice and did not use 2% of the selected phrases at all. These values indicate that although the lexicon size was imposed, participants used almost all the phrases in their judgments.

The reliability of the judgments is measured by the degree of consistency between the repeated judgments of the same events. For the numerical judgments, the mean absolute deviation between the repeated judgments was .045 (*SD* = .087). The verbal mean consistency of use, that is, the average fraction of phrases in the longer of the two response lists that was also included in the shorter list, was .62 (*SD* = 0.17). Moreover, 59% of the first-choice phrases selected were identical, and in 82% of the stimuli at least one of the phrases used in the first response list was also used in the second.

Participants were accurate in both modalities. The average absolute difference between the numerical judgments and the proportion of shaded area on the circle was .039 (*SD* = .066). We calculated the (Kendall τ_b) rank-order correlation between the shaded areas and the judgments. The median correlation for the numerical judgment was almost perfect (.96) with very little variation among participants (*SD* = .04). The median correlations for the verbal judgment (using the rank order given by the participants) were also very high: .90 (*SD* = .06) for the first word chosen and .89 (*SD* = .07) across all words used. These values are similar to those from the first study but the variance among participants is much lower, presumably because of the elimination of the fixed anchor terms.

Assessment of Conversion Methods

Table 4 presents the index of one-to-many mappings, the average number of mappings per phrase, and the measure of conversion range, the percentage of phrases in the recipient lexicon (out of 11) used in the conversion. Notice that RANK has desirable indexes on both measures. This is due, in part, to the fixed lexicon size guarantees perfect one-to-one mappings in RANK. This analysis includes comparisons of verbal, numerical and, converted responses to the same stimuli at the first-pair level and all-pairs level. The comparisons were performed among all pairs of partic-

Table 4
One-to-Many Mappings and Conversion Range Indexes for Experiment 2

Method[a]	Average number (and *SD*) of mappings per phrase		Conversion range index (%)[b]
ABSDEV	1.01	(0.10)	52
PRO	2.95	(1.99)	100
DPEAK	1.70	(0.86)	79
RANK	1.00	(0.00)	100

[a] Conversion methods are based on the following criteria: ABSDEV matches phrases by minimal absolute deviation between membership functions; ADJPRO matches phrases with identical adjusted peak ranks; DPEAK matches phrases by minimal distance between their peaks' locations; ADJRANK matches phrases with identical adjusted subjective rank order. [b] Average percentage of phrases in the receivers' lexicon used in the conversion (per participant).

ipants within each of the teams, for a total of 90 pairs, 7 teams of 4 ($7 \times 4 \times 3$) and one team of 3 ($1 \times 3 \times 2$).

We calculated the measures of agreement, PIA, and PMA for each conversion method, verbal judgment, and numerical judgment. We also calculated a new, empirical baseline for comparison. Given all the data (the actual judgments), we constructed a pseudo-conversion table that maximizes the level of the agreement indexes by matching the phrases that were coassigned to the same stimuli the highest number of times. We refer to this baseline as BEST. It is important to recognize that this value is not independent of the participants' judgments and it cannot be derived directly and independently from the phrases' membership functions or ranks. Therefore, it is not a conversion method but instead is the ultimate upper bound of the agreement indexes, given the data collected (thus the name BEST). The average coassignment indexes for the four conversion methods, the numerical judgments, the unconverted verbal judgments, and the BEST are presented in Table 5.

We analyzed each of the coassignment indexes, treating the different conversion methods and communication baselines as repeated measures. We treated the averaged coassignment indexes for each forecaster as dependent variables in the respective ANOVAs. The global *F* tests were significant for all four indexes, as evident in Table 6. We tested the significance of all 15 pairwise differences among the various conversion methods by means of the Tukey's HSD procedure. Power analysis (e.g., Cohen, 1998) indicated that these tests have almost perfect power (>.98) of detecting medium-size effects of $d = 0.5$, at the $\alpha = .05/15 = .0033$ level.

We found that the coassignment indexes of the unaided verbal communication were significantly lower than all other conversion methods in all cases. Coassignment indexes for numerical communication were significantly higher than all conversion methods only for the all-pairs PIA index. In general, the indexes of the various conversion methods were not significantly different from each other with the exception of PRO having a significantly higher first-pair PMA index than all other methods, RANK having significantly higher PIA indexes, and PRO and RANK having significantly higher all-pairs PMA indexes.

These results support our hypotheses. Across all agreement indexes, the verbal discourse had the lowest coassignment values and the numerical had the highest values in most cases. Most important, all the conversion methods significantly outperformed the unaided verbal communication, and in all cases the difference between them was large according to Cohen's (1998) classification of effect sizes. In fact, the agreement indexes for the most successful conversion methods were close to BEST, suggesting that these conversion methods achieved almost maximal levels of agreement.

To eliminate the one-to-many mappings, we implemented a randomized selection algorithm. In every case in which a conversion method mapped one of the communicator's phrases into multiple phrases in the recipient's lexicon, the algorithm randomly selected one of these phrases. To test this procedure, we repeated this process 15 times for every set of phrases that exhibited one-to-many mappings. On every replication, one of these phrases was selected, and we recalculated the agreement indices. Table 7 presents the average indexes (and their *SD*s) across the 15 independent replications (the values of the baseline indexes, verbal judgment, numerical judgment, and BEST are identical to those in Table 5 and are presented for comparative purposes).

Generally, PIA indexes were unaffected by the randomized selection, whereas PMA indexes decreased (as explained earlier, PMA is inflated by increasing the number of comparisons). It is also of interest that the variance across the randomized selection is small. For ABSDEV and RANK, indexes were unaltered because these methods rarely produce one-to-many mappings. RANK had the largest agreement indexes when randomized selection was used.

Table 5
Average (and Standard Deviation) of Coassignment Indexes for Experiment 2

Level	ABSDEV	PRO	DPEAK	RANK	VJ	NJ	BEST
			PIA				
FP	.22 (.11)	.2 (.08)	.19 (.09)	**.33** (.14)	.04 (.04)	.36 (.09)	.34 (.12)
AP	.2 (.07)	.17 (.05)	.16 (.06)	**.24** (.07)	.03 (.03)	.36 (.09)	.29 (.08)
			PMA				
FP	.27 (.13)	**.53** (.16)	.37 (.17)	.41 (.15)	.05 (.05)	.49 (.1)	.58 (.16)
AP	.56 (.18)	**.78** (.16)	.61 (.2)	.76 (.16)	.11 (.1)	.49 (.1)	.79 (.15)

Note. $N = 90$ participant dyads. The indexes are as follows: PIA = proportion of identical assignment; PMA = proportion of minimal agreement. Cells in bold letters indicate the best results for each analysis. Levels of analysis as follows: FP, comparison between the pair of first choice responses; AP, comparison between all pair of responses. Conversion methods are based on the following criteria: ABSDEV matches phrases by minimal absolute deviation between membership functions; ADJPRO matches phrases with identical adjusted peak ranks; DPEAK matches phrases by minimal distance between their peaks' locations; and ADJRANK matches phrases with identical adjusted subjective rank order. The baselines are as follows: VJ, unaided verbal judgments; NJ, unaided numerical judgments; BEST, the maximal possible agreement indexes for aided discourse given the data.

Table 6
Analysis of Variance for Communication Mode for Experiment 2

	First-pair level PIA			All-pairs level PIA			First-pair level PMA			All-pairs level PMA		
Source	*df*	*F*	Cohen's *f*	*df*	*F*	Cohen's *f*	*df*	*F*	Cohen's *f*	*df*	*F*	Cohen's *f*
Mode	5	130.43*	1.87	5	251.53*	2.57	5	160.86*	2.06	5	262.40*	2.65
Error	150	(0.003)		150	(0.002)		150	(0.006)		150	(0.007)	

Note. Values in parentheses represent mean square errors. PIA = proportion of identical assignment. PMA = proportion of minimal agreement.
* $p < .05$.

Discussion

The results from the second experiment replicate, complement, and further reinforce the results of the first experiment. First, and most important, conversion of personalized subjective probability lexicons is feasible and beneficial, in the sense that it can reduce the rate of errors in communication of verbal uncertainties. In both studies all the conversion methods outperformed the unaided verbal judgment, according to both agreement indexes at the two levels of analysis. The second conclusion is that no conversion method systematically outperformed all the others. In fact, different agreement indexes favored different conversion methods. For example, a method that had the largest value on one index could have had the smallest value on another index (such as ADJPRO on PMA and PIA in Experiment 1). Finally, we have shown that effective adjustments can be made to account for different lexicon sizes and for one-to-many mappings.

Table 7
Average (and Standard Deviation) of Coassignment Indexes Using Randomized Selection for Experiment 2

Level	ABSEDV	PRO	DPEAK	RANK	VJ	NJ	BEST
			PIA				
FP	.22 (.000)	.22 (.007)	.2 (.005)	**.33** (.000)	.04	.36	.34
AP	.20 (.000)	.18 (.004)	.16 (.003)	**.24** (.000)	.03	.36	.29
			PMA				
FP	.27 (.001)	.28 (.008)	.23 (.006)	**.41** (.000)	.05	.49	.58
AP	.56 (.000)	.59 (.010)	.48 (.009)	**.76** (.000)	.11	.49	.79

Note. $N = 90$ participant dyads. The indexes are as follows: PIA = proportion of identical assignment; PMA = proportion of minimal agreement. Cells in bold letters indicate the best results for each analysis. Levels of analysis are as follows: FP, comparison between the pair of first choice responses; AP, comparison between all pair of responses. Conversion methods are based on the following criteria: ABSDEV, matches phrases by minimal absolute deviation between membership functions; ADJPRO, matches phrases with identical adjusted peak ranks; DPEAK, matches phrases by minimal distance between their peaks' locations; and ADJRANK matches phrases with identical adjusted subjective rank order. The baselines are as follows: VJ, unaided verbal judgments; NJ, unaided numerical judgments; BEST, the maximal possible agreement indexes for aided discourse given the data.

Agreement Indexes

The quality of conversion methods was examined using two indexes, PIA and PMA, at the first-pair and all-pairs levels. These measures illustrate that agreement between two individuals can be defined in many ways. PIA is a strict measure of agreement that determines the rate of matching assignments of terms, compared with all possible distinct pairings. However, PMA is a permissive index, which requires only a minimal level of coassignment (at least one common phrase) to declare agreement between two participants. We used both measures not to compare them and to identify the more appropriate but rather to show the universal superiority of the conversion methods and to demonstrate that it does not depend on the choice of a particular (and possibly controversial) measure of agreement.

The fact that methods were ranked differently by the various measures suggests that these measures are sensitive to distinct features. Thus, decision makers might find one conversion method more attractive than the others in a particular context on the basis of the definition of coassignment they consider most relevant for that context. For example, consider a group of forecasters preparing a report to be delivered to a decision maker, say, their supervisor. The report consists of a table whose entries are the individual estimates of the forecasters for multiple uncertain events. The advisors wish to convert their estimates to the lexicon used by the decision maker. If they chose a conversion method such as RANK (high PIA), they are guaranteed to achieve high agreement among different forecasters within each event, but not in all events. If they use a method such as PRO (high PMA), they are guaranteed that in many cases at least two advisors will agree. Therefore, the choice of conversion method may vary depending on the type of consistency they like to present to their supervisor.

The Best Conversion Method

All conversion methods outperformed the unaided verbal discourse on all agreement indexes, but two methods had better results than the rest: PRO and RANK (and their adjusted versions, ADJPRO and ADJRANK). Both methods are based on the same principle, matching phrases according to their rank, but they vary with respect to the criterion used to rank order the phrases; PRO is based on the rank ordering inferred from the membership functions peaks, and RANK is based on the participants' subjective ordering of the phrases (see also Dhami & Wallsten, 2003). Both methods have a wide conversion range but differ on the number of one-to-

many mappings induced. PRO has the highest number of one-to-many mappings and RANK has the lowest.

It is tempting to single out RANK as the method of choice. Its advantage over the other methods is its simplicity and, more specifically, the fact that it does not require membership functions. The process of membership function elicitation calls for approximately 10 min of reading instructions and practice before the actual task, which takes between 30–80 s per phrase. In contrast, subjective rank ordering of phrases does not require much practice and takes little time. The main problem of the subjective ranks is that they are relative to, and depend on, the other terms in the selected set. Thus, the same word could be ranked differently by the same judge on various occasions, as a function of the other phrases the judge chooses to use. However, the membership function of any given word is an absolute and stable representation of that phrase and is relatively independent of the other words in one's lexicon. Thus, even if a judge chooses different lists of words on various occasions, the membership functions of the common words would be more or less invariant. Thus, we conclude this discussion by highlighting the performance of ADJPRO and ADJRANK but without endorsing either of them over the other.

Experiment 3

The two previous experiments focused on converting probability phrases of people who speak the same language. However, decision makers sometimes consult with forecasters from foreign countries and are forced to speak a language different from their own. For example, suppose an English-speaking lawyer is working on an international case in which she has to communicate daily with several nonnative English speakers from different cultures. The type of information she receives from her foreign colleagues involves likelihood estimates of possible outcomes and has implications on the way she approaches the case. The lawyer may change her strategy in court according to these estimates. Thus, the outcome of the lawsuit is affected by the way the likelihood estimates are translated into English.

Verbal probabilities can be translated from one language to another in two ways. One can apply a direct translation of the target phrase from Language A to Language B, by using an appropriate dictionary or a bilingual translator who is fluent in both languages. The other option is to apply an indirect translation, in which the target phrase is translated to an intermediate Language, C, which is spoken by both communicators and then translated into Language B. We suspect that this form is much more prevalent in most applications. In fact, as English has become the effective lingua franca of science, commerce, and Internet communication, it is reasonable to speculate that in most cases it serves as the intermediate language. In this study, we focus only on indirect translations.

Indirect translations can be based on the participants' familiarity with the languages, or they can rely on translation tools such as dictionaries, computer-aided translation systems or professional translators. We refer to direct translation as *spontaneous translation* (ST) and to indirect translation as *aided translation* (AT). In either approach, the mapping of words from one language to another is on the basis of a set of rules and conventions that does not take into account the subjective meanings of the verbal probabilities. Much like the communication of uncertainties within the same language, when probability phrases are treated as having the same meanings across individuals, misunderstandings can occur.

Communication of uncertainty between languages introduces another source of error that enhances miscommunication. This point is illustrated by several studies that have examined the meanings assigned to probability terms in various languages. Cohn, Cortés, and Álvarez (2003) reported significant differences between the numerical values attached to supposedly identical (according to their dictionary definitions) verbal probabilities in English and Spanish by Mexican and American students as well as by professional interpreters and translators. In a similar spirit, Shying (2000) has documented systematic differences in the numerical translations of probability words by auditors from several countries.

Another type of problem that affects literal translation between languages is described by Teigen and Brun (1999):

> The Norwegian term is *en viss mulighet,* which can be literally translated as *a certain possibility*. However, the word for certain (*viss*) used here is etymologically and semantically distinct from the word for certainty (*sikkerhet*), so *a certain possibility* has no connections suggesting certainty and is best translated as *some possibility* (although this expression may be even more vague than the Norwegian term). (p. 163)

The conversion methods we advocate are based on the membership functions or the subjective ranking of probability phrases and, as such, transcend language barriers by matching phrases on the basis of their subjective numerical, rather than literal, meaning. This suggests that using these conversion methods would improve communication of uncertainties between different languages. We predicted that the proposed conversion methods will be more accurate than SA or AT.

The next experiment was designed to test this hypothesis. We used a similar methodology but with multilingual participants who operated in their native languages. Accuracy is the ability to convert a phrase that person *i* used (in his or her language) to describe event *X* to the correct phrase person *j* used to describe the same event (in his or her native language).

Method

Participants

Thirty-five volunteers, who spoke six different native languages, received $8 to participate in a 1-hr study. There were 5 English, 5 Turkish, 6 Spanish, 6 Russian, 6 German, and 7 French native speakers. Overall, there were 17 women and 18 men (M age = 24.30 years, SD = 3.70).

Procedure

The experiment was similar in most respects to the first two experiments. It consisted of four computerized tasks: (a) selection and ranking of a personal verbal probability lexicon in one's native language, (b) elicitation of membership functions of the selected phrases, (c) numerical and verbal likelihood estimations of graphically displayed events, and (d) spontaneous translation of the selected verbal probabilities to English (only for nonnative English speakers).

In the first task, participants created lists of 5–11 phrases and ranked them. Participants typed the phrases in their native language, using the English keyboard (special characters were ignored). In the second task, the

membership functions of the selected phrases were culled. Next, participants judged the likelihood of the occurrence of several events by using numerical and verbal probabilities. Participants judged 63 targets in the verbal judgment task; each one of the 21 targets (from 0% to 100%, in increments of 5%) was repeated three times under different rigid (90°) rotations. Participants responded by selecting only one probability phrase (from the phrases they had selected in their native languages) to describe each target. The 21 targets were presented only once in the numeric judgment task. In the final task, participants were asked to spontaneously translate their phrases into English.

To obtain the baseline literal translations, we recruited five additional native speakers of the languages used in the study. They were asked to translate the phrases selected by the participants to English, using the relevant dictionary. Then, they were given all the English probability terms used in the experiment (either from the participants' ST or from the other translators) and were asked to translate them back to their native languages using the relevant dictionary. Thus, we obtained both the STs and the ATs required for comparison with the conversion methods. When the back-translated phrases were compared with the original phrases from the same language, we found a high level of agreement, indicating that the translators' work was accurate. The translators also assisted us in identifying and correcting phrases that were spelled differently by different participants because of negligence or discrepancies in converting special characters (e.g., in the case of the Russian-speaking participants, transliteration from Cyrillic to Latin characters).

Results

Descriptive Statistics of Lexicon Size and the Numerical and Verbal Judgments

Table 8 summarizes different aspects of the 266 phrases selected by the participants. The table shows a great deal of variability in lexicon size (the rows of the table are ordered accordingly) and communality of terms. For example, of the 42 words selected by the 6 Russian-speaking participants, only one phrase was selected by more than 1 person! However, the 6 Spanish-speaking participants selected 58 phrases and only 60% were unique.

Each participant made 63 verbal judgments, resulting in 2,205 judgments across all participants, and 21 numerical judgments—a total of 735 judgments. Only 5 of the 266 words selected were never used to describe a given target. The verbal judgments were quite reliable; in 62.5% of the cases the same phrase was used over the three presentations of a given display, in 32% of the cases the same phrase was used twice, and in only 5.5% of the cases participants used completely different phrases for each presentation. The judgments were also quite accurate; the median rank-order correlation between shaded areas and the numerical judgments was .94 (SD = .03), and the corresponding value for the verbal judgment was .87 (SD = .07).

Assessment of Conversion Methods

In all subsequent analyses, we distinguished between two types of conversions: *Within language* refers to the interpersonal conversions involving all the participants that speak a certain language, and *between language* refers to the interpersonal conversions among speakers of different languages. Table 9 presents indexes of one-to-many mappings and conversion range. Generally, the conversion methods seem to produce similar indexes when they are applied between and within languages. We compared the verbal and numerical judgments of every stimulus for each pair of participants, using the PIA and PMA measures at the first-pair level (participants could only use one numerical or one verbal response per stimulus). We continued to look at the measures calculated for numerical discourse as the hypothetical upper bounds for communication but used different baselines for the within- and between-languages cases. The regular verbal index applies only to the within-language communication with unconverted judgments. Four different agreement baselines were calculated for the between-languages conversions as illustrated as dotted lines in Figure 2. ST1 describes communication between a nonnative English forecaster and a nonnative English decision maker who converse spontaneously in English. ST2 describes a situation in which English text written by a nonnative English speaker forecaster is translated to the language of the decision maker, using a dictionary. AT1 refers to the situation in which a non-English-speaking forecaster and a non-English-speaking decision maker converse through English translators. Finally, AT2 is the situation in which the forecasters' estimates are first translated to English and then into the decision maker's language, using different dictionaries. Generally, ST1 and AT1 compare the English translations of both phrases, whereas ST2 and AT2 compare the translations of a phrase from Language A to Language B, with the original phrase in Language B.

Table 10 presents the between- and within-language agreement indexes. Because the four baselines yielded similar results, we

Table 8
Summary of the Number of Selected Phrases and Common Phrases per Language for Experiment 3

	No. of phrases					Common phrases[a]				
Language	*M*	*Mdn*	*SD*	Min	Max	1	2	3	4	*N*
Turkish	5.80	5.0	1.30	5	8	14	6	1	—	5
English	7.00	6.0	1.87	5	9	27	4	—	—	5
Russian	7.00	6.5	2.10	5	11	40	1	—	—	6
French	7.57	8.0	1.27	5	9	31	8	2	—	7
German	8.17	8.0	1.72	6	11	23	6	2	2	6
Spanish	9.67	10.0	1.03	8	11	35	10	1	—	6

Note. Min = minimum; Max = maximum.
[a] Number of phrases common to *n* participants, where *n* equals the number in the top of the column.

Table 9
One-to-Many Mappings and Conversion Range Indexes for Experiment 3

Method[a]	Average number (and *SD*) of mappings per phrase: Between languages		Within languages		Conversion range index (%)[b]: Between languages	Within languages
ABSDEV	1.01	(0.12)	1.01	(0.11)	64	66
ADJPRO	1.42	(0.74)	1.49	(0.89)	89	90
DPEAK	1.33	(0.64)	1.32	(0.65)	83	83
ADJRANK	1.07	(0.25)	1.06	(0.24)	90	93

[a] Conversion methods are based on the following criteria: ABSDEV matches phrases by minimal absolute deviation between membership functions; ADJPRO matches phrases with identical adjusted peak ranks; DPEAK matches phrases by minimal distance between their peaks' locations; and ADJRANK matches phrases with identical adjusted subjective rank order. [b] Average percentage of phrases in the receivers' lexicon used in the conversion (per participant).

report only the highest among them (ST1). The most important result is that all the conversion methods clearly outperformed the relevant baselines. It is interesting to note that the conversion methods tended to produce similar indexes within and between languages as well as similar levels of improvement relative to the relevant baselines.

We analyzed both indexes derived from the between-languages conversions in separate two-way mixed ANOVAs treating the forecaster's language as a between-subjects factor and the communication mode as a within-subject factor. We found a significant effect of the communication mode (numerical, translated–verbal, or converted verbal). The results are summarized in Table 11. No significant effect was found for the forecaster's native language or the interaction effect between language and communication mode. These tests are not very powerful; the probability of detecting effects of size $f = 0.4$ (designated as a large effect in Cohen's 1998 classification) is .35 for the language and .42 for the interaction, but they do not pertain to our major research hypotheses.

We tested the significance of all pairwise differences among the various conversion methods by means of the Tukey's HSD procedure. We found that the coassignment indexes of the translated verbal baseline (ST1) were significantly lower than all other conversion methods. Coassignment indexes for the numerical communication were significantly higher than all conversion methods only for the PIA index. In general, the indexes of the various conversion methods were not significantly different from each other. Power analysis (e.g. Cohen, 1998) indicates that these tests have almost perfect power (>.99) of detecting medium-size effects of $d = 0.5$, at the $\alpha = .05/15 = .0033$ level.

A randomized alternative selection algorithm was carried out on the conversion tables to eliminate one-to-many mappings. The results were similar to the original analysis supporting the validity of the randomized selection process and reinforcing the conclusion regarding the benefits of the conversion methods over the aided and spontaneous translation.

Discussion

The results of Experiment 3 replicated the findings of Experiment 1 and Experiment 2 for intralanguage communication of uncertainties. We documented the superiority of our conversion methods over four distinct alternative methods of interlanguage communication. Remarkably, the conversion methods were equally effective whether the forecaster and the decision maker used the same language or different languages. Although the various conversion methods yielded highly similar results,

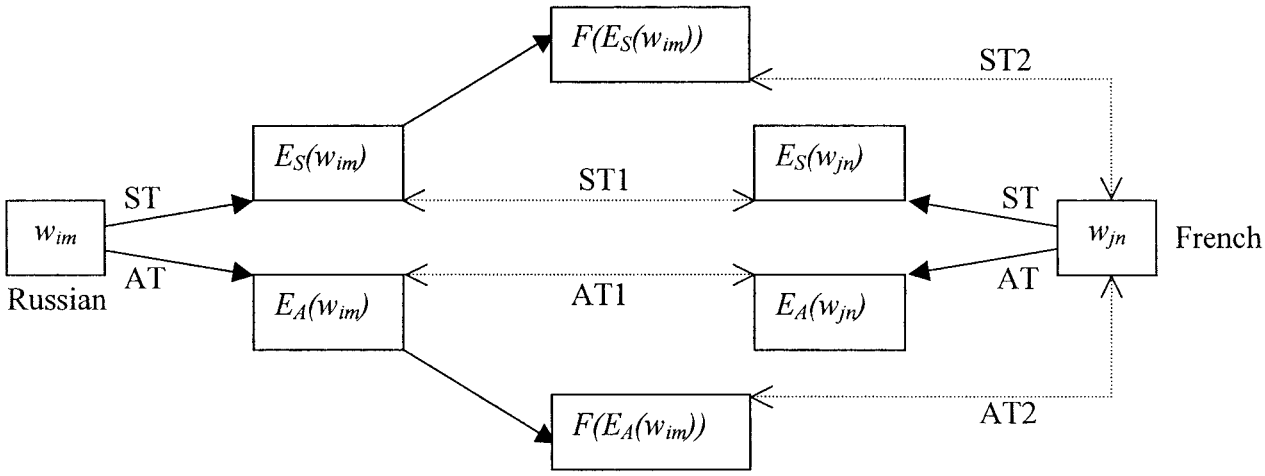

Figure 2. Schematic presentation of the four baselines of communicators' agreement. Suppose w_{im} is the phrase used by participant *i* (say, a Russian speaker) and w_{jn} is the phrase used by participant *j* (say, a French speaker). Each gave a spontaneous translation to English of their phrase, which is denoted by $ES(w)$. Their phrases were also translated into English by independent translators. These translations are denoted by $EA(w)$. Finally, $ES(w_{im})$ and $EA(w_{im})$ were translated into French by the French translator. These translations are denoted by $F(ES[w_{im}])$ and $F(EA[w_{im}])$. The dotted lines represent four possible communication paths–Spontaneous Translation 1 and 2 (ST1 and ST2) and Aided Translation 1 and 2 (AT1 and AT2).

Table 10
Average (and Standard Deviation) of Coassignment Indexes Between and Between Languages for Experiment 3

Index	ABSDEV	ADJPRO	DPEAK	ADJRANK	BEST	ST1	VJ	NJ
				Between languages				
PIA	.34	.35	.35	**.37**	.53	.06	—	.40
	(.15)	(.16)	(.14)	(.14)	(.13)	(.09)		(.13)
PMA	.57	**.73**	.71	.67	.91	.12	—	.40
	(.18)	(.18)	(.18)	(.19)	(.08)	(.15)		(.13)
				Within languages				
PIA	.35	.34	.35	**.40**	.53	—	.04	.40
	(.16)	(.17)	(.16)	(.15)	(.13)		(.07)	(.15)
PMA	.59	**.72**	.71	**.72**	.91	—	.07	.40
	(.18)	(.17)	(.19)	(.17)	(.07)		(.12)	(.15)

Note. *N* for within languages is 86 dyads. *n* for between languages is 509 dyads. The indexes are as follows: PIA = proportion of identical assignment; PMA = proportion of minimal agreement. Cells in bold letters indicate the best results for each analysis. Conversion methods are based on the following criteria: ABSDEV matches phrases by minimal absolute deviation between membership functions; ADJPRO matches phrases with identical adjusted peak ranks; DPEAK matches phrases by minimal distance between their peaks' locations; ADJRANK matches phrases with identical adjusted subjective rank order. The baselines are as follows: BEST, the maximal possible agreement indices for aided discourse given the data; ST1, unaided discourse in English by two nonnative English speakers; VJ, unaided verbal judgments; NJ, unaided numerical judgments. There is no difference between PIA and PMA for the NJ because there was only one judgment of each stimulus.

ADJPRO and ADJRANK were again slightly better than their competitors. The use of randomized selection process to handle one-to-many mappings hardly affected the quality of the conversion. ADJRANK had the largest agreement indexes when such a process was used.

General Discussion

The ultimate goal of this line of research is to develop a general Linguistic Probability Translator (LiProT) that would serve both as a useful research tool and a general decision aid and would facilitate communication of subjective uncertainties between the multiple participants in various decision situations—forecasters, judges, experts, and decision makers. The hope and expectation is that by improving the interpersonal communication of uncertainties, LiProT would also improve the quality of the ultimate decisions. In the introduction we discussed the major obstacles to efficient and accurate interpersonal communication of uncertainty—preference for verbal terms, diversity in personal lexicons, and variability in the interpretation of the terms that make up these lexicons. The procedure implemented and tested in this article addresses all three concerns. More specifically, all participants were allowed to choose their own vocabulary of uncertainty, addressing the first two points. Finally, the LiProT system minimizes the dangers of misinterpretation, by using the conversion methods documented here. The major contribution of the studies reported here is to illustrate the feasibility, flexibility, and most importantly, the efficiency of the different methods for converting probability lexicons.

Table 11
Analysis of Variance for Communication Mode and Forecasters' Language for Experiment 3

	PIA			PMA		
Source	*df*	*F*	Cohen's *f*	*df*	*F*	Cohen's *f*
	Between participants			Between participants		
Language	5	0.95	0.15	5	2.07	0.25
Error	29	(0.011)		29	(0.013)	
	Within participants			Within participants		
Mode	5	243.04*	2.45	5	435.35*	3.25
Mode × Language	25	1.06	0.37	25	0.99	0.36
Error	145	(0.002)		145	(0.004)	

Note. Values in parentheses represent mean square errors. PIA = proportion of identical assignment. PMA = proportion of minimal agreement.
* $p < .05$.

We consider this series of studies a demonstration of the methods' feasibility rather than the ultimate and authoritative answer to the problem of miscommunication. We realize that one could use other conversion methods, and we make no claims of superiority or optimality on behalf of those we tested. Also, as we have pointed out earlier, we did not attempt to identify the best or most appropriate criterion of agreement between two sets of assignments. Instead, we pointed out the robustness of our approach—all conversion methods clearly improved on unaided verbal communication across all the measures of agreement considered. Future research should address the relevant optimality issues.

Although this article was concerned with judgments and interpersonal communication of probabilities, it is impossible to ignore the issue of the quality of the proposed conversion schemes. Given that each probability judgment provided by a forecaster, as well as its translations into the various decision makers' lexicons, is related to a well-defined event with a known probability (a given target), we address this question by examining the external predictive validity of the original forecasts and their various conversions. The results of this analysis for the three experiments are summarized in Table 12. For each forecaster we calculated the (Kendall τ_b) rank-order correlation between the objective probabilities (the shaded areas of the spinners) and their verbal judgments (represented by the peaks of the membership functions of the words selected by forecasters to describe the events). These values are presented in the second column of Table 12, labeled Forecaster. The next four columns in the table include the mean rank correlations between the same events and the peaks of the decision makers' functions for the translated phrases according to the various conversion methods. These means are calculated across all the relevant decision makers. For example, Table 12 shows that in Experiment 1 the validity of the forecasters' judgments was .77 and the mean validity of the translations (across all the 17 decision makers) according to the ABSDEV criterion was .76. Three key results stand out and require no statistical tests: (a) The validity of the four conversion methods are almost identical, (b) they are similar to the validities of the original forecasts, and finally, (c) these patterns hold in all three experiments. In other words, the phrases selected by our conversion schemes from the decision makers subjective lexicons correlate with the actual events (almost) as highly as the words chosen by the forecasters when they saw the stimuli.

Table 12
Mean External Predictive Validity and (Standard Deviation) of Original Forecasts and Conversions

Experiment	Forecaster	ABSDEV	PRO	DPEAK	RANK
1 (n = 18)	.77 (.24)	.76 (.24)	.73 (.27)	.77 (.24)	.68 (.21)
2 (n = 31)	.80 (.15)	.73 (.13)	.79 (.14)	.80 (.14)	.76 (.10)
3 (n = 35)	.88 (.07)	.80 (.07)	.85 (.03)	.87 (.06)	.86 (.07)
Total (n = 84)	.83 (.15)	.76 (.14)	.81 (.15)	.82 (.15)	.79 (.14)

Note. Conversion methods are based on the following criteria: ABSDEV matches phrases by minimal absolute deviation between membership functions; ADJPRO matches phrases with identical adjusted peak ranks; DPEAK matches phrases by minimal distance between their peaks' locations; ADJRANK matches phrases with identical adjusted subjective rank order.

One possible criticism of LiProT is that using conversion methods in discourse in the same language is artificial. However, we believe that modern online communication is a comfortable setting to implement these methods without disturbing the flow or interfering with the content of discourse. Furthermore, we believe that most decision makers would be more than happy to trade off a slight inconvenience for the increase in the level of interpersonal coassignment of meaning documented in the three experiments. Imagine, for example, a group of decision makers, say a team of three doctors from Chicago, who are about to conduct a complex operation. The team is concerned about possible complications they might encounter during the operation. They have never performed such a procedure, and they would like to decrease the uncertainties before entering the operation room. They decide to consult with two internationally renowned experts who happen to live in Paris and in Moscow. An online meeting is organized, and the five doctors discuss the case. The experts provide likelihood estimates (in Russian and in French) for possible complications, on the basis of their experience and the patient's condition.

Before the meeting, the verbal probability lexicon of all the doctors is mapped by LiProT. During the meeting, all the physicians discuss the case using chat software that interfaces with LiProT. Every probability phrase used by the physicians (in English, French, or Russian) is instantly converted to the corresponding phrases in the lexicons of all the other doctors in their native language. At the conclusion of the meeting, much of the uncertainty would be resolved, and the team of doctors should be better prepared to handle the operation. The current experiments have advanced us a step closer to realizing this scenario. They suggest that the construction of LiProT is feasible and is likely to be beneficial in increasing the effectiveness of interpersonal communication of uncertainty.

References

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

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

Brun, W., & Teigen, K. H. (1988). Verbal probabilities: Ambiguous, context-dependent, or both? *Organizational Behavior and Human Decision Processes, 41,* 390–414.

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

Budescu, D. V., Karelitz, T. M., & Wallsten, T. S. (2003). Determining the directionality of probability words from their membership functions. *Journal of Behavioral Decision Making, 16,* 159–180.

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. (1990). Dyadic decisions with numerical and verbal probabilities. *Organizational Behavior and Human Decision Processes, 46,* 240–263.

Budescu, D. V., & Wallsten, T. S. (1995). Processing linguistic probabilities: General principles and empirical evidence. In J. Busemeyer, D. L. Medin, & R. Hastie (Eds.), *Decision making from a cognitive perspective* (pp. 275–318). San Diego, CA: Academic Press.

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

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

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.

Cohen, J. (1998). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohn, L., Cortés, M., & Álvarez, A. (2003). *Quantifying health risks likelihood terms in Spanish and English.* Manuscript submitted for publication.

Dhami, M. K., & Wallsten, T. S. (2003). *Interpersonal comparison of subjective probability and subjective probability phrases.* Manuscript submitted for publication.

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., & Neal, M. (1991). Vagueness, ambiguity, and the cost of mutual understanding. *Psychological Science, 2,* 231–324.

Fillenbaum, S., Wallsten, T. S., Cohen, B. L., & 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.

Fischer, K., & Jungermann, H. (1996). Rarely occurring headaches and rarely occurring blindness: Is rarely = rarely? The meaning of verbal frequentistic labels in specific medical contexts. *Journal of Behavioral Decision Making, 9,* 153–172.

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

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 Rep. No. 250). U.S. Army Research Institute for the Behavioral & Social Sciences. Arlington, VA: Author.

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.

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

Marshall, E. (1986, June 27). Feynman issues his own shuttle report, attacking NASAs risk estimates. *Science, 232,* 1596.

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.

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

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

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

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.

Olson, M. J., & Budescu, D. V. (1997). Patterns of preference for numerical and verbal probabilities. *Journal of Behavioral Decision Making, 10,* 117–131.

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.

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

Shying, M. (August, 2000). Auditors interpretations of "in-isolation" verbal probability expressions: A cross-national study. In M. Kennelley, *The Influence of Environmental and Cultural Factors in International Accounting.* Symposium conducted at the American Accounting Associations (AAA) Annual Meeting, Philadelphia.

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

Teigen, K. H., & Brun, W. (1999). The directionality of verbal probability expressions: Effects on decisions, predictions, and probabilistic reasoning. *Organizational Behavior and Human Decision Processes, 80,* 155–190.

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

Wallsten, T. S., Budescu, D., Erev, I., & Diederich, A. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making, 10,* 243–268.

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

Wallsten, T. S., Budescu, D. V., & Tsao, C.-Y. (1997). Combining linguistic probabilities. *Psychologische Beitraege, 39,* 27–55.

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

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

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.

Windschitl, P. D., & Weber, E. (1999). The interpretation of "likely" depends on the context but "70%" is 70%—right? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1514–1533.

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

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

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

Appendix

Phrases Selected by Participants in Experiment 1 (Listed in Alphabetical Order)

Phrase	*N*	Phrase	*N*	Phrase	*N*
a chance	1	high likelihood	2	pretty doubtful	2
a possibility	1	high possibility	1	pretty impossible	1
absolute certainly	2	highly improbable	1	pretty unlikely	2
against the odds	1	highly unlikely	2	probable	4
almost certain	3	impossible	18	quite certain	1
almost impossible	2	improbable	1	quite doubtful	1
almost unfeasible	1	likely	3	quite possible	3
always	1	little chance	1	rather likely	1
certain	18	little likely	1	rather unlikely	2
definitely	2	little uncertain	1	slight chance	1
doubtful	4	maybe	2	slight possibility	1
even odds	18	medium likelihood	1	slight probability	2
extreme doubtful	1	more likely	1	small chance	1
faint possibility	1	most definitely	3	sure thing	2
fair chance	2	most likely	3	toss-up	3
fair possibly	1	most of the time	1	uncertain	1
fair probability	1	never	1	unlikely	1
fairly certain	1	no chance	1	usually	1
fifty-fifty chance	2	no possible	1	very certain	1
good chance	4	not likely	3	very doubtful	4
good possibility	1	not very feasible	1	very feasible	1
good probability	2	not very likely	1	very likely	5
great likelihood	1	possible	4	very unlikely	4
great possibility	1	pretty certain	1		

Note. *N* represents the number of participants who selected each word in their lexicon.

Received April 4, 2003
Revision received October 2, 2003
Accepted October 5, 2003 ■