Inter-personal Communication of Precise and Imprecise Subjective Probabilities[*]

D. V. BUDESCU
University of Illinois at Urbana-Champaign, USA

T. M. KARELITZ
University of Illinois at Urbana-Champaign, USA

Abstract

We analyze communication of uncertainty among individuals as a function of the parties' preference for modes of communication. We assume that different individuals may prefer precise *Numerical* probabilities, *Ranges* of probabilities or *Verbal* descriptions of probabilities, and consider all possible pairings of communicators and receivers under this classification. We propose a general criterion of optimal conversion among the various modalities, describe several instantiations tailored to fit the special features of the various modalities, and illustrate the efficacy of the proposed procedures with empirical results from several experiments.

Keywords

subjective probability, judgment, inter-personal translation, verbal probabilities

1 Introduction

Consider a situation where two individuals communicate about stochastic events. The two are equally interested and motivated to communicate as efficiently and precisely as possible. This paper is concerned with procedures that can be employed to address the individuals' different preferences for modality of communicating probabilistic opinions. Although many decision analysts and orthodox Bayesians consider precise numerical probabilities to be *the* language of uncertainty, many people (layman and experts, alike) prefer to use probability phrases (e.g. review by Budescu and Wallsten [6]) or other imprecise variants of probability. In this paper we propose ways to achieve the highest possible level of accuracy in communication while accommodating these individual preferences.

[*] This work has been supported by the US National Science Foundation under Award No. SES 9975360. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the US National Science Foundation.

1.1 Reasons for preferences of specific communication modes

Spontaneous preferences for one particular mode may be due to several factors:

The perceived *nature of the uncertainty to be communicated*–Budescu and Wallsten [6] have speculated, and Olson and Budescu [14] have documented empirically that most individuals prefer to use precise numerical estimates to communicate uncertainty about repeated events with aleatory uncertainty, but tend to use more imprecise methods when communicating the probabilities of unique events with epistemic uncertainty.

The perceived *strength of the available information*–The responses to the survey conducted by Wallsten, Budescu, Zwick, and Kemp [18], indicate that people would gravitate towards more precise modes of communication, if they perceive the available information to be firmer, reliable and valid.

The person's *role in the communication*–In the same survey Wallsten et al. [18] have found that most people prefer to use imprecise terms when they communicate to others, but prefer others to communicate to them in precise terms, if possible (see also, Brun and Teigen [2] and Erev and Cohen [8]).

In addition to these systematic factors, preferences may be due to plain *individual differences* that reflect one's lifetime experiences in dealing with, and communicating, uncertainties.

1.2 The problem

The need to communicate probabilities arises in a variety of situations. A common case is when both individuals have prior opinions, have access to some relevant (possibly overlapping) information, and wish to exchange information to further refine their respective estimates. In this *symmetric* case the designation of communicator and receiver is arbitrary, as the two individuals can act in both capacities. For example, think of two friends who talk about the chances of their favorite team to win a game. The other prototypical case involves *asymmetric* communication: only one individual, the Forecaster (**F** for short), has access to, or possesses the necessary expertise to make sense of, the relevant information for the probability estimation. The second individual, the Decision Maker (or **DM**) needs to make a choice or decision on the basis of the F's estimate, and without the benefit of his, or her, own probability assessment. For example, think of an investor (the DM) who gets from his, or her, favorite financial advisor estimates of the likelihood that certain investment policies will succeed.

The two situations are similar in many respects but the former is more complex because a complete analysis should take into account the processes that govern the combination of one's own opinions with estimates obtained form others (Yaniv and Kleinberger [19]). To simplify the analysis, we will focus on the second case. In the same spirit, we will not consider the case where one needs to aggregate multiple forecasts from various sources (Budescu, Rantilla, Yu and

Karelitz [5], Wallsten, Budescu and Tsao [17]).

To summarize, we analyze an asymmetric dyadic communication situation where one F and one DM share a common interest in optimizing communication, but they may have different preferences for modality of communicating probabilistic opinions.

1.3 A typology of communication preferences in a dyad

We distinguish between three modes of communication: precise (point) **N**umerical probability estimates (e.g., 0.45), precise **R**anges of numerical values (e.g., 0.3 - 0.55), and **V**erbal phrases (e.g., good chance). Ranges with precise end points exclude implicit vague ranges such as ”in the forties” or ”at least 0.80”, but such expressions can be analyzed as verbal terms.

The three modes can be ranked from the most precise (N) to the most vague (V). In fact, the more precise modes can be represented as special cases of the more vague modes: clearly an N is an R where the lower and upper limit coincide, and we will show later how N and R can be viewed as special cases of V under a particular representation of the probability phrases. This typology implies 9 distinct dyadic patterns of dyadic preferences for modes of communication that will be denoted by ordered pairs, where the first character in the pair refers to the F's preference.

2 The translation process

The problem we wish to address is deceptively simple – How to *best* convert a judgment originally expressed in the F's favorite response mode (N, R or V), to an estimate in the DM's favorite mode (N, R or V).

The criterion of optimality is the level of (dis)similarity between the F's judgments translated into the DM's favorite mode, and the DM's spontaneous (and independent) judgments of the same events in his, or her, favorite mode. For example, assume that the F prefers numbers and the DM prefers verbal terms (i.e., an [N,V] dyad). If both had the same prior probability distribution and could access the same information pertaining to the target event, X_i, their spontaneous and independent judgments would be $n_F(X_i)$, and $v_{DM}(X_i)$, respectively.

Any mapping of the F's spontaneous judgment into the DM's favorite communication mode is a *translation*. For example, $v_{DM}[n_F(X_i)]$ is the verbal translation of the F's original numerical judgments. An *optimal translation* is one that maximizes the similarity between the translation of the F's term into the DM's favorite mode, and the DM's spontaneous judgment of the target event (assuming he/she has the same priors and could access the same information).

Note that (dis)similarity is measured in the scale of the target modality (i.e., the one that is favored by the DM), so it always relies on commeasurable units or entities. On the other hand, these entities vary as a function of the DM's favorite modality. Next we describe some sensible choices for the dissimilarity metrics. Our goal in this paper is to provide a general framework for the translation process and illustrate the feasibility of the approach. We make no claim of optimality, or uniqueness on behalf of these choices, and realize that other metrics could be used in this context.

Dissimilarity between two numbers, n_{DM} and n_F, is defined as the distance between them:

$$DS_n\{n_{DM}, n_F\} = |n_{DM} - n_F|. \tag{1}$$

Dissimilarity between two ranges, r_{DM} and r_F, is a function of their respective lengths, and their overlap. Consider two ranges, $r1$ (ranging from $l1$ to $u1$) and $r2$ (from $l2$ to $u2$). The width of the range over which the two overlap is $OV_{12} = Max\{0, [Min(u1,u2) - Max(l1,l2)]\}$, and the joint range of values they span is $JR_{12} = [Max(u1,u2) - Min(l1,l2)]$. We define the dissimilarity between the two ranges as:

$$DS_r\{r_{DM}, r_F\} = JR_{DM,F} - OV_{DM,F}. \tag{2}$$

This measure is zero if, and only if, the two ranges coincide. For any pair of ranges, r_{DM} and r_F, the index is maximal when they are disjoint.

Dissimilarity between two verbal terms, v_{DM} and v_F, is defined in the context of a particular representation of such phrases. Wallsten, Budescu, Rapoport, Zwick, and Forsyth [16] suggested that probability phrases are fuzzy concepts and proposed using Membership Functions (MFs) over the $[0,1]$ probability interval to represent their vague meanings (see Zadeh [20]). A phrase's MF assigns to each probability a real number that represents the (non-negative) degree of its membership in the concept defined by the phrase. These values are scaled between 0 and 1 (Norwich and Turksen [13]), 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. MFs can be estimated directly (non-parametrically) based on the participants' direct or indirect judgments (see Budescu and Wallsten [6], Wallsten et al. [16]). Alternatively, one can fit MF using specific families of functions, such as polynomials (Budescu, Karelitz and Wallsten [4]), or trapezoidal functions.

Let $\mu_{v_{DM}}(p)$ and $\mu_{v_F}(p)$ be the MFs representing the two words being compared. The similarity between the two words should reflect the closeness between their respective MFs. There are many possible single-valued indices of closeness between the two functions (see review by Zwick, Carlstein, and Budescu [21]), and we will only list two of them here (these are not necessarily monotonically related). The first measure is the total absolute distance between the two functions.

Formally, we can write[1]:

$$DS_{v_\mu}\{v_{DM}, v_F\} = \int_{p=0}^{1} |\mu_{v_{DM}}(p) - \mu_{v_F}(p)| dp. \tag{3}$$

The second index is the distance between the peaks of the two functions. Assume that both $\mu_{v_{DM}}(p)$ and $\mu_{v_F}(p)$ are single peaked (see Budescu and Wallsten [6] on this point). Let $\pi(v)$ be the probability (or the center of the range of probabilities) at which the function $\mu_v(p)$ reaches its maximal value. We define, a second measure of dissimilarity as:

$$DS_{v_\pi}\{v_{DM}, v_F\} = |\pi(v_{DM}) - \pi(v_F)|. \tag{4}$$

2.1 General comments on the measures of dissimilarity

The various measures may appear at first glance to be unrelated and, somewhat arbitrary, so a few comments and clarifications are in order. First, we should point out that all the dissimilarity indices are *distances*. In all cases they assign to every pair of (N,R or V) judgments a non-negative real number ($DS = 0$ only if the two members of the pair are identical). The measures are symmetric, satisfy the triangle inequality and induce a weak order over all pairs.

One could invoke other metrics for these comparisons. A particularly elegant approach would be to use the same metric for all modalities. Technically, this is feasible since numbers can be represented by point MFs (membership of 0 everywhere, and 1 for the chosen number) and ranges can be represented by flat MFs (membership of 0 everywhere outside, and 1 everywhere within the chosen range), and treated in the same fashion as the MFs obtained for verbal terms. However, we believe that the metrics identified above are better suited for our purposes because they are more in line with the particular level of (im)precision implied by the three modalities.

The last comment is subtler. Our definition of similarity relies on a counterfactual scenario that gives rise to a hypothetical entity - the DM's spontaneous judgment of the target event if he, or she, had the same prior probability distribution and could access the same information that was used by the F as a basis for his/her judgment. Strictly speaking, this definition is meaningful only in those cases where it makes sense to assume that a person's *judgment depends only on the specific* information presented. This implies that the relevant information is unambiguous and does not lend itself to different (subjective) interpretations. In other words, the observed variability among probabilities assigned to a target event by different individuals can be attributed solely to different response styles and/or random factors within the judges. This formulation makes perfect sense

[1] In most empirical applications the MFs are approximated by a set of n points over $[0,1]$, so a discrete version of this measure can be used to approximate it.

for repeatable and exchangeable events, but not for unique events where subjective probabilities rely on internal epistemic uncertainty that can vary systematically across individuals. (Ariely, Au, Bender, Budescu, Dietz, Gu, Wallsten and Zauberman [1] and Wallsten, Budescu, Erev and Diederich [15], discuss various facets of this key distinction).

For example, it is quite unlikely that if we were to present anti-smoking activists and tobacco lobbyists with the results of a new study on the effects of second-hand smoking, they would agree in their estimation of the probabilities that second-hand smoking has serious public health consequences. The differences between their estimates would reflect (a) their different prior probabilities, and (b) their differential assessment of the quality, reliability and validity of the new data. Clearly, no translation method can be expected to reconcile disagreements of this type. Despite these irreconcilable differences in their opinions, we can still take advantage of optimal translation schemes derived for various pairs of communicators based on their judgments of a standard set of exchangeable events. When these translation methods are applied they can reduce the effect of other sources of variability among the participants and provide *the most accurate representation of the F's assessment in the DM's favorite communication mode*, where accuracy is measured by one of the dissimilarity metrics discussed above.

2.2 Methods of translation

We return now to our original question: how to *best* convert a judgment originally expressed in the F's favorite response mode (N, R or V), to an estimate in the DM's favorite mode (N, R or V). Before we discuss translation schemes for each of the 9 cases, it makes sense to classify them into three distinct groups:

Common modalities - In three cases ([N,N], [R,R] and [V,V]) both individuals share a common preference for mode of communication, so there is no need to worry about differential precision. Conversions may be employed to account for inter-personal differences in the way the relevant terms are chosen and used.

Resolving vagueness - In three cases ([R,N], [V,R] and [V,N]) the DM prefers a more precise mode of communication than the F. Thus, the challenge is to find a translation that resolves the vagueness implicit in the F's judgment to achieve the higher level of precision required by the DM.

Imputing vagueness - In the other three cases ([N,R], [N,V] and [R,V]) the DM prefers a more vague mode of communication than the F. Thus, the challenge is to find a translation that replaces the precision implicit in the F's judgment to reflect the higher level of vagueness expected by the DM.

We will discuss the three classes separately. In each case we describe and justify a translation method designed to optimize our stated goal and, when appropriate, we review and discuss relevant results from several empirical studies that are described in the next section.

2.3 The data

Over the last two years we have conducted four experiments designed to test the efficacy and accuracy of various translation methods of probability phrases (V). The studies vary in many specific details (Budescu and Karelitz [3] and Karelitz and Budescu [11]) but they share a set of common features that allow us to analyze some of their results jointly. The focus on the N and V responses is neither accidental, nor arbitrary. Subjects rarely communicate their probabilities by means of ranges even when offered the opportunity (e.g. references in Budescu and Wallsten [6]). For example, in one of the studies analyzed below when this option was present, it was used in less that 7.5% of the cases. Thus, we will not present any empirical results concerning translations involving Rs.

The four studies involved a total of 128 individuals (all students at the University of Illinois in Urbana Champaign, and most of them native English speakers[2]). All the experiments were computer controlled, and included the following three tasks: (1) Selection of a personal verbal probability lexicon including 5-11 phrases (In a few cases some, or all, the phrases were selected by the experimenters based on previous research); (2) Elicitation of MFs for all the phrases; and (3) Numerical and verbal estimation of probabilities of a common set of events.

Subjects created their lists by selecting combinations of words and semantic operators (modifiers, intensifiers, etc.) from two lists, or typing in phrases. They were instructed to select phrases that span the whole probability range, and they tend to use regularly. Membership functions were elicited using a method validated by Budescu et al. [4]. Each phrase was presented with a set of eleven probabilities ranging from 0 to 1 in increments of 0.1. The subjects judged the degree to which the target phrase captured the intended meaning of each of the eleven numerical probabilities by using a bounded scale, anchored by the terms '*not at all*' and '*absolutely*'. In the last task, the participants saw a series of circular, partially shaded, targets. Their task was to assess the likelihood that a dart, aimed at the center of the target, would hit the shaded area. The shaded areas varied from one trial to another and covered the full (0,1) range. On separate presentations these probabilities were judged numerically (by selecting one value from a list of 21 probabilities, ranging from 0 to 1 in increments of 0.05), or verbally by selecting (in some cases up to four) phrases from their lexicons.

2.4 Common modalities

[**N**,**N**] This is the "gold standard" case of Bayesian decision analysis. Presumably numbers are universal and everyone understands, interprets and uses them in identical fashion. Therefore, no transformation is required. There is, however, evidence that people's mapping of their internal feelings of uncertainty into

[2] One of the studies was concerned with translation of probability phrases across languages and we recruited native speakers of French, German, Spanish, Russian and Turkish.

numbers is imperfect. In particular, most people over-(under-) estimate low(high) probabilities (e.g. references in Erev, Wallsten and Budescu [9]), and it is conceivable that there are systematic differences in the degree to which individuals tend to avoid (or favor) the extreme values. In principle, one could quantify this tendency and apply appropriate *stretching (or contracting) transformations*. To illustrate this point consider multiple judges $(1,2,\ldots,j,j',\ldots,J)$ who judge a set of stochastic events $(1,2,\ldots,i,\ldots,I)$. Assume that: (1) All judges have access to the same amount of information, implying that differences in their judgments are due only to (a) differences in their use of the response scales and (b) random components. (2) All judges spontaneously recognize events that are impossible (probability = 0), certain (probability = 1), and as likely as not (probability = 0.5). (3) Assume an "ideal judge" who is perfectly calibrated (no biases) and accurate (no random component). Thus his/her judgments, $p_1, p_2, \ldots p_i$, coincide with the events' "objective probabilities".

The probability assigned by judge j to event i is denoted by p_{ij}, and can be expressed as a function of the objective probability, p_i, his/her bias parameter, α_j, and the random component, e_{ij} which we assume is distributed with $\mu_e = 0$ and (finite) σ_e. We use a variation of Karmarkar's [10] model, that assumes that the logit of the judged probabilities is a linear function of the logit of their objective counterparts:

$$Log\left[\frac{p_{ij}}{(1-p_{ij})}\right] = \alpha_j \cdot Log\left[\frac{p_i}{(1-p_i)}\right] + e_{ij} \tag{5}$$

Individual differences between judges are captured by the parameter α_j, which is bounded from below by 0 (when all events are assigned a probability of 0.5). An unbiased judge should have an α_j of 1, but we expect that most individuals would have parameters between 0 and 1 that are consistent with the regressive model described above. We used a least-squares procedure to estimate the individual parameter, α_j, in model 5. The model fits the data well for almost all subjects (median $R^2 = 0.98$, median $MSE = 0.13$). The distribution of the individual parameters matches our expectations: 64 values (50%) are between 0.55 and 0.98, 45 participants (35.2%) are almost perfectly calibrated ($0.99 \leq \alpha_j \leq 1.01$), and only 19 individuals (14.8%) have parameters values above 1. To verify that these differences reflect systematic individual differences rather than pure random error, we performed two additional analyses: (a) we compared these results with a model where the parameter, α_j, was constrained to be 1 (thus the model includes only random error). A comparison of the two models in terms of R^2_{adj} favors slightly the fitted model. The modal difference (34% of the cases) is 0, but there is a clear majority (43% vs. 23%) of cases where the fitted model fits better (mean difference in fit = 0.02), even after we account for its extra parameter. Significance tests comparing the fit of the two models (separately for each subject) revealed that this differences was significant at the traditional 0.05 level, for 25.2% of the subjects. (b) We re-analyzed two of the studies in which all subjects judged all the displays twice, so we could obtain two estimates of the parameter, α_j, for each person. In both studies (involving a total of 55 subjects) the between-subjects variance component was considerably larger than the within-subject component (in fact the within-subject component was not significantly greater then 0).

In principle, one could convert numerical estimates from one person to another in an optimal fashion by applying simple stretching (contracting) transformation based on the estimates of the individual parameters, α_j, $\alpha j'$.

[**R**,**R**] The use of precise ranges instead of simple point estimates reflects one's perceived level of imprecision in his or her estimate of the probability of the target event. Clearly, the arguments invoked in the [N,N] case regarding the nature of the numbers, apply here as well. This would suggest that no transformations are indicated. It is conceivable, however, that there are systematic differences in the degree of imprecision perceived by different individuals and this would induce systematic differences in the widths of their ranges. One could quantify this tendency and apply appropriate *imprecision equating* transformations.

[**V**,**V**] This situation is, probably, the most interesting and it has been the focus of much of our recent research. This case is qualitatively different from the previous two for several reasons. There is a large literature indicating that (a) spontaneously, people tend to use highly different and diverse lexicons, and (b) the numerical meanings (as well as other forms of representation) associated with these words vary dramatically across people (e.g. review in Budescu and Wallsten [6]). Thus, one cannot assume that everyone is equally comfortable with, or interprets identically terms such as ”*likely*”, ”*poor chance*”, etc. For this case we advocate the following multi-stage procedure that is sensitive to these empirical findings: (a) each participant selects his/her own subjective lexicon; (b) MFs are elicited for all the terms in the list; (c) the MFs of the words selected by the F and the DM are placed on a common probability scale and are matched according to the criterion of choice (DS_{v_μ} or DS_{v_p}). Occasionally this procedure does not yield a unique solution, i.e., one of the F's words can be translated equally well into several of the DM's words. Of course, all these words are equally valid translations of the F's judgment. If practical considerations prevent one-to-many translations, one of them can be selected randomly (or by some other sensible tie-breaking procedure).

We have done quite a lot of empirical work documenting the efficacy of this approach (Karelitz and Budescu [11], Karelitz, Dhami, Budescu, and Wallsten [12]). In each of our studies we compared the level of agreement in assignment of verbal phrases to the same events among numerous pairs of distinct individuals. We hypothesized that the lowest level of agreement would be observed with spontaneous (un-aided), verbal discourse, and the best level of agreement would be found in the case of numerical communication. Most importantly, we expect that communication with converted phrases would be superior to un-aided verbal communication, and closer in quality to the numerical case. To quantify the level of inter-personal agreement we defined two indices of co-assignment. We use two

measures because some of the events were judged more than once and yielded different responses from the subjects. Both measures range from 0 to 1 (higher values indicate stronger agreement), and can be interpreted as measures of the accuracy of inter-personal communication of imprecise opinions.
PIA- **P**roportion of **I**dentical **A**ssignments- the proportion of *comparisons* where both participants assigned the **same** phrase to a given event.
PMA- **P**roportion of **M**inimal **A**greement - the proportion of *stimuli* for which both participants assigned **at least one** common phrase to a given event.

In the interest of brevity we only report results based on PIA, which is a more stringent measure than PMA (PIA $\leq$ PMA) because it weighs the agreement by the number of comparisons made (The PMA results are very similar in a qualitative sense). Table 1 summarizes the results of 4 studies (details in Karelitz and Budescu [11]). Each cell presents the mean (and SD) PIA in the various modes, and across all pairs of subjects analyzed.

Table 1: Summary of agreement indices from 4 studies

		Translation Criterion			
Study	No. of	VJ: Unaided	DS_{v_μ}	DS_{v_π}	NJ: Unaided
	pairs	Verbal Judgments	(Eq. 3)	(Eq. 4)	Numerical Judgments
1	306	0.05 (0.03)	0.23 (0.12)	0.22 (0.10)	0.29 (0.07)
2	90	0.04 (0.04)	0.22 (0.11)	0.19 (0.09)	0.36 (0.09)
3	86	0.04 (0.07)	0.35 (0.16)	0.35 (0.16)	0.40 (0.15)
4*	509	0.06 (0.09)	0.34 (0.15)	0.35 (0.14)	0.40 (0.13)

* Experiment 4 involves translations of words across various languages. VJ is based of the subjects’ spontaneous translation of words from their native languages to English.

The results clearly support our predictions: unaided VJ had the lowest values for both indices in all the studies and NJ had the largest values. The two translation criteria clearly outperformed the unaided verbal communication[3].

2.5 Resolving vagueness

[**R**,**N**] In principle, any sensible person should be able to infer a single N value from his/her partner range without invoking any translation scheme. The individual differences discussed in the [N,N] and [R,R] cases apply here as well. In principle, one could improve the quality of communication by (a) inferring the F’s best guess (presumably, the center of the reported range) and, if necessary, (b) applying the appropriate *stretching (or contracting) transformation*.
[**V**,**N**] Recall that every word in the F’s lexicon has a (single-peaked) MF defined over the [0,1] interval that describes the degree to which the various prob-

[3] Dhami and Wallsten [7] and Karelitz and Budescu [11] report similar results with several other translation methods.

abilities match the intended meaning of that particular word. The MF's peak, $\pi(v)$, is the single numerical probability that is most representative of the word's meaning and is the translation of choice. Occasionally, the MF does not have a unique maximum, so all probabilities within a given range can be considered to be equally good representations of the word's meaning. In these cases it is convenient to translate the word into the mid-range of these probabilities. To illustrate the potential accuracy of this approach we compared the peaks of the 977 verbal phrases used by 113 of the subjects in our experiments with the mean of their numerical judgments when judging the same events. We found a remarkable similarity between the two sets: (a) the median within-subject difference between the two is 0.006 and the median absolute difference between them is 0.097; (b) the median within-subject rank order correlation between the peaks of the words and the mean numerical judgments is 0.89; and (c) the two sets are almost perfectly related linearly with a median within-subject intercept of -0.022, a median within subject slope of 1.06, and a median R^2_{adj} of 0.90. These results indicate that the translation procedure can map with high accuracy the intended meaning of the words and predict accurately the numerical probabilities used to describe the same events.

[**V**,**R**] Every MF is, essentially, a collection of ranges since every level of membership, ν ($0 \leq \nu \leq 1$), defines a range of values, $R(\nu)$, such that $\mu(\nu) \geq \nu$. Typically, as ν increases, $R(\nu)$ becomes narrower indicating the range of values that possess that (higher) level of membership is more restrictive. Thus, the translation from a V to a R boils down to the issue of which threshold, ν, to choose. Presumably, there are systematic differences in the "typical" threshold that individuals tend to use in these circumstances, so one could quantify this tendency and identify the most appropriate range for each individual. We are not aware of any studies that have collected both verbal and upper and lower numerical bounds of the probabilities of the same events, so we are not in a position to assess the efficacy of the proposed approach.

2.6 Imputing vagueness

[**N**,**R**] If numbers are the universal language of uncertainty and everyone interprets them identically, any sensible DM would infer that the F's single N is the center of a range that describes his/her opinions, but there is no clue regarding the implied imprecision of the F's opinion. One could improve the quality of communication by reversing the procedure described for [R,N], i.e., (a) applying the appropriate *stretching (or contracting) transformation* for the DM, and (b) imputing the DM's typical band of imprecision. We are not familiar with any empirical work along these lines.

[**N**,**V**] Recall that all the words in the F's lexicon have single-peaked MFs defined over the [0,1] interval. These functions describe the degree to which any given probability matches the intended meaning of the various phrases. The proposed translation rule calls for the choice of that phrase that has the highest membership at the N in question. This procedure is not guaranteed to yield a unique solution, i.e. there could be several words with equally high membership at that probability, and all these words should be considered equally valid translations of the numerical judgment. If necessary, one of these words can be selected randomly (or by some other tie-breaking procedure). In analyzing our studies we looked at responses from 118 subjects who used an average of 14.85 distinct numerical judgments. We analyzed the verbal responses that were assigned by the subjects to the events to which they assigned a certain numerical response. On the average, each set of events that were judged to be equally probable (in the numerical mode) generated 1.81 distinct verbal phrases, and in 68% of the cases at least one[4] of these verbal responses had, indeed, the maximal membership for that probability. Another look at the same data indicates that for 59.7% of the numerical judgments at least one of the verbal terms used was predicted from the MFs. Interestingly, we found large individual differences: 30 subjects (25.4%) are at, or below a 40% success rate, while for 27 subjects (22.9%) the rate of accurate translation is greater than, or equal to 75%. Not surprisingly, the level of agreement is considerably higher for the extreme (0 and 1), and the central (0.5) numerical probabilities.

[**R**,**V**] All the words in the F's lexicon have (single-peaked and continuous) MFs defined over the [0,1] interval. For any fixed range of numerical probability these functions describe the degree to which the probabilities in that range match the intended meaning of the various phrases. The proposed translation rule calls for the choice of that verbal term that has the highest average membership over the R in question. It is possible that there would be several words with equally high membership over that range probabilities. All these words should be considered equally valid translations. We are not aware of studies that have collected the relevant data for the empirical evaluation of this procedure.

3 General Discussion

In this paper we proposed a unifying conceptual framework for optimal interpersonal translation of probabilistic information for the 9 distinct cases we identified. We discussed the 9 scenarios at different levels of details, and provided extensive empirical support for some of them. Although the cases are not encountered with similar frequency in applied settings, we decided to review all of them to illustrate the generality, feasibility and flexibility of the overall approach.

This line of research is part of an effort to create a general Linguistic Probability Translator (*LiProT*, for short) that could serve both as a useful research tool, and a general decision aid. *LiProT* would facilitate communication of subjec-

[4] In 14.5% of the cases more than one word tied for the highest membership at a given probability. The mean number of words tied for maximal membership was 1.14.

tive uncertainties between participants in various decision situations - forecasters, judges, experts and decision makers - by reducing the dangers of miscommunication of probabilities among the various members of the group.

To fix ideas consider a group of experts (physicians, intelligence officers, financial forecasters, etc.) who communicate with each other, possibly electronically form various locations. As part of this process they need to exchange probabilistic information based on the evidence available to them and reflecting their own unique expertise. If various people in this group have differential preferences for modes of communicating probabilities to others and receiving information from others, then each of the 9 cases discussed above may be relevant for some of the pairs. The procedures described and partially tested in this paper provide a foundation for such a system. Before the meeting, the participants' preferred modes of communication are ascertained, their verbal probability lexicons are mapped, and *LiProT* derives the appropriate translation scheme for each dyad. During the meeting, every probability (N, R or V) used by each of the experts is instantly converted optimally to the favorite modality (N, R or V) of each of the other participants.

For example, assume that participant *A* prefers to communicate and to receive numbers, participant *B* has a universal preference for Vs, and participant *C* prefers to communicate with V, but to receive Ns (the modal pattern according to Wallsten et al. [18]). Every uncertainty judgment provided by *A* (using Ns) will be translated by *LiProT* into the closest V in judge *B* lexicon (using the [N,V] module), and into the most appropriate N for judge *C* (using the [N,N] module). Similarly, the verbal uncertainty judgments provided by *C* will be translated into the closest N for judge *A* (using the [V,N] module), and into the most appropriate V in *B*'s lexicon (using the [V,V] module). Thus, all judges communicate their opinions and receive information in their respective preferred modes. This approach may be too restrictive, since preferences for a particular mode may vary as a function of the situation, the nature of the target event and its underlying uncertainty. A good translator should allow the receiver of the communication to choose the mode of communication. For example, judge *B* may choose to have judge *A* numerical translated by *LiProT* into the closest V in judge in most cases, but occasionally he/she may opt for a simpler, and more direct, translation into the most appropriate N.

In closing we emphasize that this work has focused on communication of uncertainty, and has not addressed the issue of the efficacy of the proposed translations in the context of specific decision situations. We are now conducting empirical work that seeks to determine the degree to which these translation rules, which were shown to improve the inter-personal communication of uncertainties, could also improve the quality of the ultimate decisions involving these uncertainties.

4 Acknowledgements

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

References

[1] D. Ariely, W. T. Au, R. H. Bender, D. V. Budescu, C. Dietz, H. Gu, T. S. Wallsten, and G. Zauberman. The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied*, 6:130–147, 2000.

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

[3] D. V. Budescu, and T. M. Karelitz Improving the quality of decisions through interpersonal translation of probability phrases. *In preparation.*

[4] D. V. Budescu, T. M. Karelitz, and T. S. Wallsten. Predicting the directionality of probability words from their membership functions. *Journal of Behavioral Decision Making*, in press, 2003.

[5] D. V. Budescu, A. K. Rantilla, H. Yu, and T. M. Karelitz. The effects of asymmetry among advisors on the aggregation of their opinions. *Organizational Behavior and Human Decision Processes*, in press, 2003.

[6] D. V. Budescu, and T. S. Wallsten. Processing linguistic probabilities: General principles and empirical evidence. In *Decision Making from a Cognitive Perspective*, J. Busemeyer, D. L. Medin, and R. Hastie (Eds.), 275-318, 1995. San Diego, CA: Academic Press.

[7] M. K. Dhami, and T. S. Wallsten. Interpersonal comparison of subjective probability and subjective probability phrases. *Submitted for publication*.

[8] I. Erev, and B. L. Cohen. Verbal versus numerical probabilities: Efficiency, biases, and the preference paradox. *Organizational Behavior and Human Decision Processes*, 45:1–18, 1990.

[9] I. Erev, T. S. Wallsten, and D. V. Budescu. Simultaneous over- and under-confidence: The role of error in judgment processes. *Psychological Review*, 101:519–527, 1994.

[10] U. S. Karmarkar. Subjectively weighted utility: A descriptive extension of expected utility model. *Organizational Behavior and Human Performance* 21:61–72, 1978.

[11] T. M. Karelitz, and D. V. Budescu. You say probable and I say likely: Improving interpersonal communication with probability phrases. *Submitted to publication*.

[12] T. M. Karelitz, M. K. Dhami, D. V. Budescu., and T. S. Wallsten. Toward a universal translator of verbal probabilities. In *Proceedings of the 15th International Florida Artificial Intelligence Research Society (FLAIRS) Conference*, 298-503, 2002.

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

[14] M. J. Olson, and D. V. Budescu. Patterns of preference for numerical and verbal probabilities. *Journal of Behavioral Decision Making*, 10:117–131, 1997.

[15] T. S. Wallsten, D. V. Budescu, I. Erev, and A. Diederich. Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making*, 10:189–209, 1997.

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

[17] T. S. Wallsten, D. V. Budescu, and C-Y. Tsao. Combining linguistic probabilities. *Psychologische Beitraege*, 39:27–55, 1997.

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

[19] I. Yaniv, and E. Kleinberger. Advice taking in decision-making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, 83:260–281, 2000.

[20] L. A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.

[21] R. Zwick, E. Carlstein, and D. V. Budescu. Measures of similarity among fuzzy concepts: A comparative analysis. *International Journal of Approximate Reasoning*, 1:221–242, 1987.

David V. Budescu is a Professor in the Quantitative division of the Department of Psychology, University of Illinois, Urbana-Champaign, IL, USA 61820. E-mail: dbudescu@uiuc.edu

Tzur M. Karelitz is a graduate student in the Quantitative division of the Department of Psychology, University of Illinois, Urbana-Champaign, IL, USA 61820. E-mail: karelitz@uiuc.edu