Comparing the Calibration and Coherence of Numerical and Verbal Probability Judgments

Thomas S. Wallsten • David V. Budescu • Rami Zwick
Department of Psychology, CB# 3270, Davie Hall, University of North Carolina, Chapel Hill, NC 27599-3270
Department of Psychology, University of Illinois, Champaign, Illinois 61820
Department of Marketing, The Pennsylvania State University, University Park, Pennsylvania 16802-9976

Despite the common reliance on numerical probability estimates in decision research and decision analysis, there is considerable interest in the use of verbal probability expressions to communicate opinion. A method is proposed for obtaining and quantitatively evaluating verbal judgments in which each analyst uses a limited vocabulary that he or she has individually selected and scaled. An experiment compared this method to standard numerical responding under three different payoff conditions. Response mode and payoff never interacted. Probability scores and their components were virtually identical for the two response modes and for all payoff groups. Also, judgments of complementary events were essentially additive under all conditions. The two response modes differed in that the central response category was used more frequently in the numerical than the verbal case, while overconfidence was greater verbally than numerically. Response distributions and degrees of overconfidence were also affected by payoffs. Practical and theoretical implications are discussed.

(*Subjective Probability; Judgment; Calibration; Verbal Probabilities; Coherence; Additivity*)

1. Introduction

Probability judgments by experts, or by decision makers themselves, often play an important role in decision analysis and research. For at least two reasons decision analysts and researchers virtually always require the judgments to be numerical (e.g., 80% or 4:1 odds) rather than linguistic (e.g., *very likely*). First, it is commonly thought and widely argued (Behn and Vaupel 1982; von Winterfeldt and Edwards 1986) that numerical expressions are precise, unambiguous communications that allow expected value or expected utility calculations, while natural language is vague, subject to different interpretations by different people, and not useful for meaningful calculations. Second, precisely for the reasons just described, the quality of numerical expressions can be evaluated, whereas that of linguistic ones cannot.

In contrast to this view is a growing interest in the use of verbal probability expressions to communicate opinion. In part, this interest stems from the development of expert systems that use linguistic phrases to represent differentially precise and hedged degrees of uncertainty (e.g., Clark 1988; Dutta 1985; Fox, Barber and Bardhan 1980; Zadeh 1975). More generally, people often resist expressing in quantitative form their personal judgment about the chances that an event will occur or a statement is true. Reasons frequently cited in support of the preference for verbal estimates are, the fact that they are perceived as more natural, easier to understand and communicate, and that they convey the vagueness, or softness, of one's opinions (e.g., Budescu and Wallsten 1985, 1987; Wallsten and Budescu 1990).

The assumed wide-spread preference for verbal over numerical communication has been confirmed in a recent survey of 442 people (Wallsten et al. in press). In that study, 65% of the respondents stated that they prefer to communicate uncertainty verbally to other people, while 70% preferred to receive it numerically. These results are consonant with Erev and Cohen's (1990)

0025-1909/93/3902/0176$01.25

finding that most people actually chose to communicate verbally and to receive information numerically in a decision-making task.

The fact that a majority of people prefer communicating their opinions verbally and a substantial minority do not mind receiving them in the same mode does not negate the possibility that the use of linguistic expressions in decision situations entails serious problems of decision quality and forecast evaluation. The present paper suggests a method that may be useful in real situations for obtaining verbal probability estimates and for assessing their calibration or other quantitative measures of goodness. The method is evaluated by means of an experiment that compares properties of numerical judgments with those of verbal ones obtained in the proposed fashion.

The remainder of this introduction first addresses the issue of decision quality to set the stage for considering the complementary problem of forecast evaluation. Next the evaluation of forecasts from the perspective of calibration measures is considered. Subsequently, our procedure for obtaining and evaluating verbal probability judgments is described. The introduction concludes with an overview of the experiment, which also introduces coherence as an additional evaluation measure.

1.1. Possible Effects of Verbal Probabilities on Decision Quality

We have compared the pattern and quality of decisions given verbal and numerical probability estimates of the same events in a number of studies involving both choice and bidding paradigms (Budescu and Wallsten 1990; Budescu et al. 1988; Erev and Cohen 1990; González-Vallejo et al. in press; González-Vallejo and Wallsten 1992). Remarkably, very little difference appeared in the overall quality of the decisions under the two conditions, despite subtle, but important, differences in their patterns. Reviews of much of this work are provided by Wallsten (1990a; 1990b). Other research (Jaffe-Katz et al. 1989; Rapoport et al. 1990) suggests overriding similarities in judgment processes given the two types of information.

If decision quality were grossly inferior given verbal than numerical probability judgments, then it would be unnecessary to worry about any other evaluation of verbal forecasts. On the other hand, the fact that on average decision quality is roughly equivalent under the two modes does not imply that the evaluation problem is solved. In any particular situation, one mode may lead to distinctly better decisions than another, and in some cases (e.g., forecasting long range behavior of countries or of markets) interest is in the quality of the forecast itself prior to the formation of a decision. Thus, it is necessary to ask how verbal forecasts might be evaluated.

1.2. Evaluating Verbal Forecasts by Means of Calibration Measures

Beyth-Marom (1982) speculated that one reason forecasters prefer using verbal judgments is their perception that the quality of verbal forecasts cannot be assessed in the same way the quality of numerical ones can. In reviewing the literature on probability encoding Wallsten and Budescu (1983) identified five evaluation criteria—reliability over time, internal consistency or coherence, external validity (i.e., correspondence with other probability measures of the same events), construct validity (i.e., the degree of invariance of judgments across different elicitation methods), and calibration. The last of these criteria refers to the degree to which the subjective judgments of events' probabilities are confirmed by the eventual observation of the true state of nature. Calibration was referred to by Wallsten and Budescu (1983) as the "common standard" for the analyst, and it undoubtedly is the most researched and closely scrutinized measure of goodness of judgment.

Calibration is achieved when, in the long run, $P\%$ of the events which are judged to have a probability of $P/100$, are true, or occur (e.g., Lichtenstein, et al. 1982). The definition, however, does not induce an explicit operational measure. For this purpose, calibration is often described in terms of a bivariate plot of the judged probabilities and the relative frequencies of occurrence of the events accorded various estimates (or accorded estimates within specified intervals). The plot is technically known as a "reliability diagram" (e.g., Yates 1982), and when its points are connected they generate a "calibration curve" (e.g., Lichtenstein et al. 1982).

It is generally understood that perfect calibration entails the curve falling on the diagonal. Global measures of calibration attempt to quantify the discrepancy between the empirical curve and the main diagonal of the diagram, and can be decomposed into subscores reflecting various aspects of the judgments. These decompositions and their interpretations are discussed and illustrated by Yates (1982).

Most behavioral research has focused on the calibration of answers to questions regarding discrete unrelated facts (almanac items). The "state of the art" up to 1980 was summarized by Lichtenstein et al. (1982), and the main results have not been seriously challenged. In general, people tend to be overconfident and this tendency is related to the item's difficulty, with greater degrees of overconfidence for more difficult items. Overconfidence tends to be less extreme when predicting future events (e.g., Fischhoff and MacGregor 1982; Ronis and Yates 1987; Wright 1982), and calibration generally is better when the events are related than when they are unique (Keren 1987). With certain very marked exceptions, (e.g., Beck et al. 1985; Daan and Murphy 1982; Murphy and Winkler 1977; and Wallsten and Budescu 1983) the pattern of results is similar with experts. A thorough recent review and analysis of a broad range of calibration studies has been provided by Keren (1991).

The only previous attempt to empirically quantify the quality of verbal judgments is by Zimmer (1983, 1984). In his work, 90 drafted German soldiers were administered a test of 150 political knowledge items. After answering each item, the subjects selected a verbal estimate from a fixed set of five available terms. Zimmer reported good correspondence at the group level between the relative frequencies correct and the median numerical values associated with the phrases used. The study is a good first step and the results are promising, but they are far from definitive because of shortcomings in the design. Aggregation across a large number of subjects and a small number of terms is most likely to mask large individual differences in the understanding and use of the selected terms (e.g. Wallsten et al. 1986). In addition, restricting the vocabulary to a relatively small number of terms preselected by the experimenter may have biased the results. Finally, the study did not include a control condition in which the same stimuli were judged numerically.

1.3. A Procedure for Obtaining and Evaluating Verbal Probability Estimates

Hamm (1991) has suggested that when verbal probabilities are required, they be selected from a fixed list and, optionally, be assigned numerical values by the forecaster. Our proposal is similar, but differs in important ways. First, subject to a constraint to be described below, we propose that forecasters or analysts choose a priori their individual limited vocabularies—we set the number of phrases at 11 in this study. This recommendation stems from the observation by Budescu and Wallsten (1990), Budescu et al. (1988), and Rapoport et al. (1990) that when forecasters are allowed to select verbal expressions either freely or from a large list, the terms they employ vary enormously over individuals, but most people use no more than 11 to 15 phrases (but see Zwick and Wallsten 1989). This result suggests that (1) no information would be lost at the individual level if analysts are limited to a fixed number of phrases (sufficient to span the full probability range), and (2) that each individual should be allowed to select his or her own vocabulary to span the full range of probability. The constraint on the selection, imposed for reasons that will be immediately obvious, is that the vocabulary contain anchor phrases for the two end points and the middle of the probability continuum (such as *impossible*, *tossup* and *certain*).

Once an individual's vocabulary is selected, it still must be converted to numbers for purposes of evaluation (and possibly of expected utility or other calculations). Numerous methods for converting probability phrases to standard numerical values have appeared over the years (e.g., Bass et al. 1974; Beyth-Marom 1982; Kadane 1990; Mosteller and Youtz 1990), but all are flawed because they rely on group average judgments. As Clark (1990), Wallsten and Budescu (1990), and other commentators on the Mosteller and Youtz (1990) article have pointed out, variability over individuals in interpreting such phrases is large and consistent.

The second part of our proposal is to use either of two methods for numerical conversion, each based on an individual's own judgments. One procedure takes advantage of the fact that individuals have stable vague interpretations of phrases within a context. In particular people rank order such phrases very consistently over time and methods (Beyth-Marom 1982; Budescu and Wallsten 1985; Johnson 1973). Thus, one recommendation is simply to have each individual rank order his or her selected terms from *impossible* to *certain*. Following the ranking, assign probability values of 0, 0.5, and 1 to *impossible*, *tossup*, and *certain*, respectively. Assume that phrases between *impossible* and *tossup* are equally spaced, as are phrases between *tossup* and *certain*, and assign probability values accordingly to the nonanchor terms uniquely selected by each person. We call this the "modified equal spacing method." This scaling procedure is arbitrary, but cannot be too far off when the number of phrases is sufficiently high.

The alternative scaling procedure is theoretically more justifiable, but also more arduous. It relies on techniques that were recently proposed and successfully validated by Wallsten et al. (1986) and Rapoport et al. (1987), who showed that the meanings of probability terms can be represented by membership functions over the [0, 1] probability interval. Each function takes its minimum value (usually zero) for probabilities not at all in the vague concept represented by the term, its maximum value (generally one) for probabilities definitely in the concept, and intermediate values otherwise. The derived membership functions have interpretable and reasonable shapes. They are generally single peaked, although for all practical purposes the functions for extreme expressions can be considered monotonic (decreasing for low probability terms and increasing for high probability ones) (Budescu and Wallsten 1990; Jaffe-Katz et al. 1989; Rapoport et al. 1987; Wallsten et al. 1986). Numerous indices have been proposed for ranking membership functions or for converting them to single numerical values. Because the single-value indices perform very well (Borolan and Degani 1985) we propose converting each function to a central value in the manner suggested by Yager (1979). Specifically, letting $\mu_v(p)$ be the membership value of probability p for phrase v, Yager suggested that a location, or probability value, be derived for each phrase from its membership function by means of

$$W_v = \frac{\sum_{i=1}^{m} \mu_v(p_i)p_i}{\sum_{i=1}^{m} \mu_v(p_i)} \qquad (1)$$

where $i = 1, \ldots, m$ indexes the specific probability values for which membership values were derived. Note that W_v is analogous to the mean of a distribution.

2. The Experiment

The goal of the experiment was to compare the quality of numerical probability judgments with that of verbal judgments obtained from individuals using their own vocabularies, which subsequently were converted to numerical values by each of the two methods described above. Subjects in the study provided verbal and numerical probability estimates that each of 300 general knowledge items is true. Each item, in fact, appeared on different days in a true or a false form, which allowed additivity (a requirement of coherent estimates) to be compared between the two response modes. Thus, quality was compared by contrasting the two modes of responding in terms of calibration and coherence.

For three reasons, we elected to use unique almanac-type questions rather than either future oriented or related events for the current study. The empirical foundation is greatest in this domain, allowing the most thorough comparison between verbal and numerical judgments. It was easiest to generate the required large number of items whose truth values were already known to us. Finally, the events that are forecasted in most real-world situations are unique, not related.

A payoff scheme was imposed and manipulated both to motivate careful responding and to determine whether it had any differential effect on the use of the two response scales in the present context. We used the spherical scoring rule

$$S = a + bN^{-1} \sum_{i=1}^{N} p_i \Big/ [p_i^2 + (1 - p_i)^2]^{0.5} \qquad (2)$$

where p_i equals the judged probability of true if event i is actually true or the judged (or implied) probability of false if the event is actually false, a is a constant, b is a positive constant, and the sum is taken across all N events. This is one of a number of *strictly proper scoring rules*, all of which share the property that the optimal strategy on the part of a respondent is to provide probability estimates that match one's true subjective probabilities (Stäel von Holstein 1970).

Note that S is maximized by assigning all true events a probability of true equal to 1 and assigning all false events a probability of true equal to 0, and that in general higher scores are better. Equation (2) was used to reward judgments because it is sensitive to deviations from the optimal strategy and because in previous studies (Wallsten 1976) manipulation of the constants a and b profoundly affected the distribution of numerical responses without altering their ordinal characteristics.

2.1. Method

2.1.1. Subjects. Twenty-one subjects, all summer school students at the University of North Carolina at

Chapel Hill, provided data. Subjects were randomly assigned to each of three groups (C, N, and P) differing in terms of the payoff scheme, which will be explained below, used to motivate careful responding.

Subjects received $32 for five sessions. In addition, subjects in each of groups P and N were eligible for bonuses of $8, $6, $4 and $2, respectively, for the first through the fourth highest score. No bonuses were available for subjects in group C.

2.1.2. Materials. A total of 300 factual claims were derived from 500 binary questions originally developed by researchers at Decision Research. The original material consisted of 100 historical items (of which we used 60) asking which of two events occurred earlier, 111 items (of which we used 67) asking which of two cities was closer to a target city, and 289 items (of which we used 173) asking which of two continents, countries, states, or cities was more populous in 1970. In addition, 30 items were selected for practice, 10 of each type.

We cast each item into the form of two complementary claims, one true and one false, in order to use a full-range response method and to allow tests of response additivity. For example, one of the historical items became Claim A: "The Monroe Doctrine was proclaimed before the Republican Party was founded," and Claim B: "The Republican Party was founded before the Monroe Doctrine was proclaimed." Thus the 300 items used for data collection yielded 600 claims, of which half were true and half were false. Subjects were required to give their confidence, or subjective probability, that each claim is true, responding numerically in some sessions and verbally in others.

2.1.3. Procedure. The experiment was run on PC's, with subjects working in separate cubicles. Subjects served for five sessions, generally on consecutive days. The first four sessions consisted of judging the factual claims described above. Sessions 1 and 2, which included instruction and practice periods, generally lasted one and a half to two hours, while sessions 3 and 4 generally lasted 30 minutes less. Responses within a given session were either numerical (N) or verbal (V), with subjects roughly evenly distributed over the orders NVNV, NVVN, VNVN and VNNV.

Each item was judged in its true and false version, both numerically and verbally. Thus, the first numerical and the first verbal session each contained all 300 items, 150 in the form of true claims and 150 as false claims. The remaining numerical and verbal sessions contained the 300 complementary claims. The first numerical and first verbal session each began with extensive instructions. For the numerical sessions subjects were instructed that, "We are interested in your best estimate of the chances from 0% to 100% that each claim is true. If you are sure that the statement is true, indicate this by typing in 100. If you are confident that the statement is false, type in 0. . . . If you have no information relevant to judging the truth of the claim, and you feel it is as likely to be false as it is to be true, type in 50. In all other cases type in a number between 0 and 100 to indicate your confidence in the truth of the statement." The numerical sessions began with nine practice trials. No practice was offered on the second numerical session, but subjects were allowed to read the instructions again before beginning.

For the verbal sessions, subjects were instructed, "We are interested in your best estimate of the chances that each claim is true. You will provide this estimate by selecting the most appropriate phrase from a list of such terms. If you are sure that the statement is true, indicate this by typing in *certain*. If you are confident that the statement is false, type in *impossible*. If you have no information relevant to judging the truth of the claim, and you feel it is as likely to be false as it is to be true, type in *tossup*. In all other cases select that phrase from the list that best indicates your confidence in the truth of the statement. In addition to the three words *certain*, *impossible* and *tossup*, you will have the chance to select eight additional phrases describing various levels of probability. Therefore, you will be operating with a list of 11 phrases, which should be sensitive enough to allow you to distinguish between various levels of confidence."

The first verbal session began with each subject selecting and rank ordering a vocabulary of 11 probability phrases constrained to include the phrases *impossible*, *tossup*, and *certain*. They chose eight additional terms to span the full range from *impossible* to *certain* by selecting from a set of 64 phrases formed by using the terms *sure*, *good chance*, *likely*, *probable*, *improbable*, *unlikely*, *slight chance*, and *doubtful*, alone or in combination with the modifiers *highly*, *very*, *quite*, *rather*, *pretty*, *fairly*, and *somewhat*. Phrases subsequently were ranked with the constraint that *impossible* be ranked 1 and *certain* be ranked 11.

After the rank ordering the computer initiated the nine practice trials. In each case a claim appeared at the top of the screen and the 11 phrases appeared under it in random sequence, but identified by sequential numbers from 1 to 11. The instructions were available, but no practice was provided on the second verbal session.

2.1.4. Payoffs. No payoffs were employed for group C (Control group). Instead, the instructions said, "Our main interest is in your ability to express accurately your degree of knowledge. In each case, provide your most precise judgment, given your knowledge, of the chances of the claim being true."

Subjects in groups P (Positive payoff) and N (Negative payoff) won or lost points according to each judgment and to whether the claim was true or false. Scoring was accomplished by means of the spherical scoring rule, Equation (2). The judged probability of true, p, was derived from a numerical response by dividing it by 100. Probabilities were derived from the verbal responses, by means of the modified equal spacing method described earlier.

For group P, $a = -87$ and $b = 137$ in Equation (2), causing responses of 50 or *tossup* to guarantee 10 points regardless of the truth of the claim (hence the label *positive* or P for that group). For group N, $a = -155$ and $b = 205$, causing 50 or *tossup* to guarantee -10 points (hence the label *negative* or N). For both groups, the maximum score for correctly identifying the truth or falsity of a claim was 50 points. The minimum score, obtained by responding 100 or *certain* to a false claim, or 0 or *impossible* to a true one was -137 points for group P and -155 points for group N. Subjects were not given the scoring rule or the constants. Instead, extensive instructions for the numerical and verbal cases, complete with illustrative tables, were used to explain the consequences of different response selections. No feedback was provided; subjects learned their scores only at the end of each session.

2.1.5. Membership Functions. The fifth session lasted about 45 minutes and was devoted to obtaining membership functions for each subject's eight selected phrases plus *tossup*. The method of graded pair-comparisons, described in detail by Wallsten et al. (1986), was employed for this purpose. On each trial one of the phrases appeared at the top of the screen, with two probability wheels below it, one on the left and one on the right. Below the two wheels was a response line extending from the left to the right side of the screen with an arrow centered on it. The subject was asked to place the arrow on the response line so that its relative position corresponded to how much better the phrase described one probability rather than the other.

This pair-comparison procedure requires that the subjects judge all pairs of a set of probability values for each phrase. The responses are then analyzed according to a difference model, which assumes that the relative placement of the arrow on the response line is proportional to the difference in the degree to which each of the two probabilities is described by the phrase. The methodology and scaling procedure have been tested extensively within the framework of conjoint-measurement by Wallsten et al. (1986) and Rapoport et al. (1987).

Based on the results of the Wallsten et al. and the Rapoport et al. studies, as well as that of Budescu et al. (1988), different probability values were selected for use with the various phrases. Specifically, membership values were assessed for *tossup* at the probability values of 0.40, 0.45, . . . , 0.60. For the high terms (*sure*, *good chance*, *likely*, and *probable*) and the hedged high terms (those modified by *quite*, *rather*, *pretty*, and *fairly*) the probability values were 0.38, 0.48, . . . , 0.98, for the intensified high values (those modified by *very*, *highly*, or *extremely*) they were 0.84, 0.88, . . . , 1.00, for the low terms (*improbable*, *unlikely*, *slight chance*, and *doubtful*) they were 0.02, 0.07, . . . , 0.32, for the hedged low terms they were 0.02, 0.09, . . . , 0.44 and for the intensified low terms the probabilities were 0.00, 0.04, . . . , 0.16. Two replicates of all probability pairs were presented for each phrase. Phrases appeared in random order, and the second replication of the pair-comparisons immediately followed the completion of the first replication.

2.2. Results

2.2.1. Selection of the Verbal Phrases and Numerical Responses. Sixty different verbal phrases were selected by 21 subjects, of which 20 (33%) were chosen by only one person, 28 (47%) by two to four people, and only eight (13%) by more than five people. None appeared in more than half the lists. Table 1 shows the eight phrases selected by at least five subjects plus the three anchor terms, along with the distributions of their assigned ranks. With the exception of the anchor terms,

Table 1 **Distribution of Eight Most Frequently Selected Words and Three Anchor Terms Across Ranks**

	Rank											
Phrase	1	2	3	4	5	6	7	8	9	10	11	Total
impossible	21											21
improbable		2	4	2								8
unlikely		2	4	1	1							8
doubtful		3	0	4	1							8
slight chance			1	3	3	3	1					11
tossup					8	13						21
good chance					1	3	1	1				6
probable						1	3	1	2			7
likely						1	1	3	2			7
pretty sure								3		5		8
certain											21	21

Note: Impossible and *certain* were constrained to be ranked 1 and 11, respectively.

these phrases span from three to five rank positions. Even *tossup* is not uniformly placed in the center rank. The considerable degree of individual difference in use of language was expected and mandates that most analyses be done at the individual subject level.

In session 5 subjects made graded pair-comparisons to provide data for the derivation of membership functions for each of their selected phrases plus *tossup*. The judgments were replicated twice. The pooled reliability correlation over all subjects and terms is 0.75. Given the generally good levels of reliability, responses were averaged over the two replications and scaled in the manner described by Wallsten et al. (1986). The result is a membership function $\mu_v(p)$ for each of the eight selected phrases plus *tossup* for each subject, from which locations W_v were calculated according to Equation (1). The vector of W_v values for a given subject will be denoted $\mathbf{W}$.

The subjects also explicitly rank ordered the 11 phrases in their personal vocabularies from *impossible* to *certain*, which in conjunction with the modified equal spacing assumption yields a second set of rank values we call R_V, with the vector referred to as $\mathbf{R}$. The pooled rank correlation between $\mathbf{R}$ and $\mathbf{W}$ over nine terms (excluding the anchors *impossible* and *certain*) is 0.81. At the individual level, the lowest correlation is 0.67 and the highest is 1.00.

Of the 12,600 numerical judgments, 84% were multiples of 0.05 and 65% were multiples of 0.10. The two extremes (0 and 1.0) accounted for 23% of the judgments and the middle point (0.5) for another 20%. The mean number of distinct numerical judgments used by a subject was 36.5. However, when all values used less than six times (i.e., 1% of the time) were eliminated this figure dropped to 17.1.

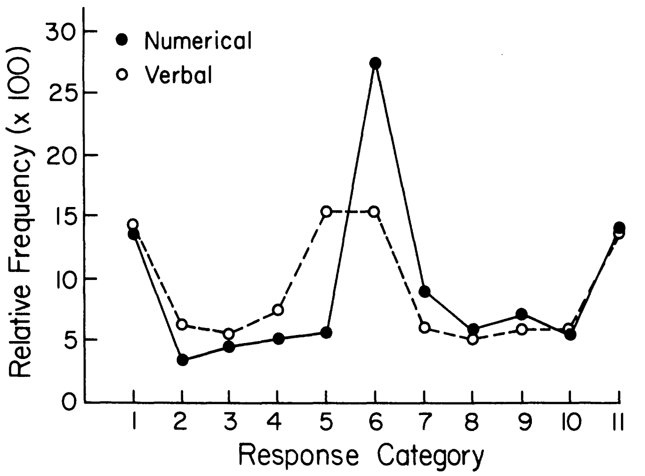
2.2.2. Distributions of Responses. A possible difference between the two modes of responding that may affect the other comparisons is in the distributions of responses. To render this and certain of the subsequent comparisons meaningful, the probability judgments were grouped into 11 classes centered at 0.025, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 0.975.

An indication of the relative equality of use of the response classes within a modality is obtained from the variance in the frequencies of their usage. If a subject distributed responses uniformly among the 11 categories, the resulting variance in frequency is zero. And the more unequal was the use of response categories, the greater is the variance in frequency of usage. A 3 × 2, group by response mode, repeated-measures ANOVA run on the log variances, yielded significant effects of response mode ($F(1, 18) = 12.76$, $p < 0.05$) and of group ($F(2, 18) = 6.01$, $p < 0.05$), but not of their interaction ($F(2, 18) < 1$). Tukey's HSD test shows the group effect to be due to a significant difference ($p < 0.05$) between groups P and N, while neither group differs significantly from C.

The differences in response frequencies are illustrated in Figures 1 and 2. Figure 1 plots the mean relative frequency (×100) of verbal and numerical responses in each of the 11 categories averaged over the three groups. It is evident that variability in the use of response categories is greater for the numerical than the verbal case, because the numerical response category 6 (centered at 0.50) was used much more frequently than the other numerical categories. On the average, 0.50 was used 20 times more than *tossup* and this difference is very close to significant ($F(1, 18) = 3.70$, $p = 0.07$). Also, 14 subjects (67%) used 0.50 more frequently than *tossup*.

Figure 2 shows the mean relative frequency (×100) of responses in each category for Groups P and N, averaged over the verbal and numerical modes. Group C fell generally between the two and is omitted from the graph so that the significant difference between Groups

Figure 1 **Distribution of Responses Over the 11 Response Categories Separately for the Numerical and Verbal Modes**



P and N can be appreciated. Note that categories 1, 5, 6, and 11 were used more frequently by Group P than by Group N, while the reverse was true for the remaining categories, so that overall, categories were used somewhat more equally by Group N than by Group P. It must be noted that three of the six Group P subjects ranked *tossup* fifth and two of the seven Group N subjects did so, so that category 5 was "central" for a greater proportion of the P than the N subjects.

2.2.3. Additivity Analyses. Each subject was asked on separate occasions to provide a personal probability that each of two complementary claims is true. Perfectly coherent judgments of the two claims should add to 1.00. This prediction was checked by using the actual (unrounded) judgments for the numerical responses and by means of the two scales, **R** and **W**, for the verbal responses. Table 2 shows the mean sums of the judgments of complementary claims for each subject under numerical responding and under each of the two conversions of the verbal responses. For most subjects, the mean sums are rather close to 1.0 under both response modes, although the patterns are not identical. T-tests indicate that the mean sums do not differ significantly

Figure 2 **Distribution of Responses Over the 11 Response Categories Separately for Group P (Positive) and Group (N) (Negative)**

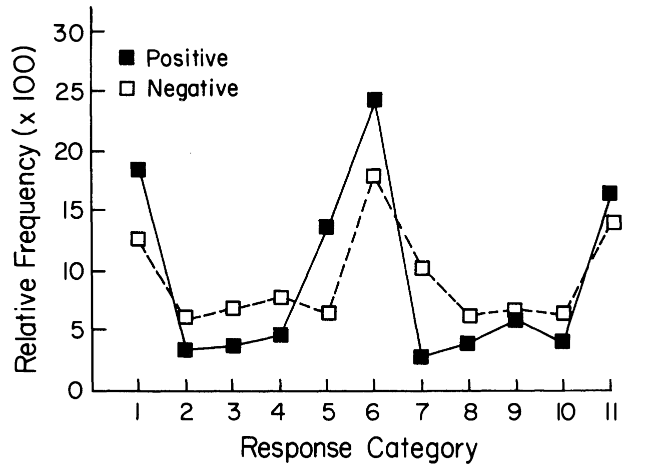


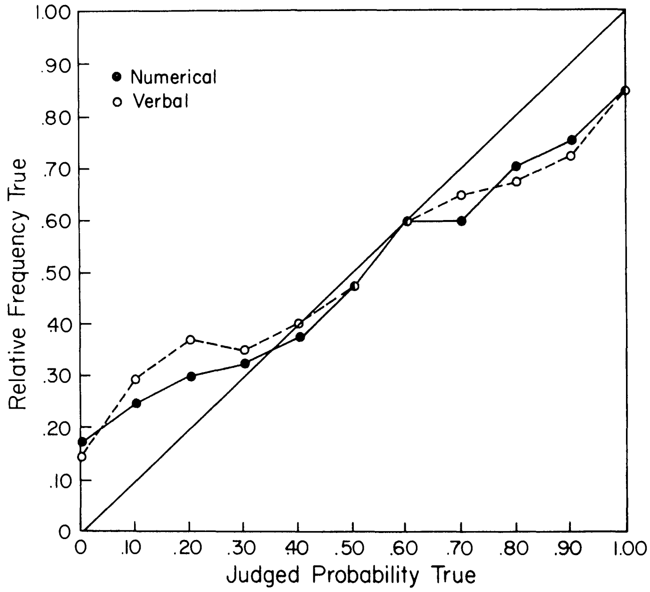
Table 2 **Sums of Judgments of Complementary Claims**

		Verbal	
Subject	Numerical	R	W
	Group N		
1	1.02	1.02	1.02
2	1.07	1.03	0.93
3	0.98	0.96	0.90
4	1.04	1.00	0.97
5	1.02	1.02	1.16
6	1.00	0.99	0.99
7	1.00	0.97	0.95
Mean	1.02	1.00	0.99
	Group C		
8	0.98	0.98	1.00
9	1.04	1.00	1.04
10	1.00	1.00	0.99
11	1.01	1.04	0.92
12	1.26	1.14	1.23
13	1.03	0.99	0.96
14	0.97	0.99	0.87
15	1.00	1.05	0.97
Mean	1.04	1.02	1.00
	Group P		
16	0.99	1.02	0.92
17	1.01	0.97	1.00
18	1.04	1.02	0.97
19	1.03	1.04	1.01
20	0.89	0.92	0.86
21	1.00	1.02	1.04
Mean	0.99	1.00	0.97
Grand Mean	1.02	1.01	0.99

from 1.0. Thus, additivity is essentially satisfied for both the verbal and the numerical judgments. Interestingly, the mean sums for the numerical and verbal responses correlate strongly over subjects ($r = 0.85$ and 0.74 between the numerical and the **R** or **W** representations, respectively). Thus the degree to which additivity is satisfied is consistent within subjects over response modes.

It can be seen in Table 2 that the grand mean sum in the numerical mode exceeds that under **R** in the verbal mode by a mere 0.01. This difference is not significant. A slightly different pattern emerges when the verbal judgments are represented by **W**. In this case there is a significant mode effect ($F(1, 18) = 5.39$, $p < 0.05$), with a grand mean difference of 0.03. A similar pattern emerges when additivity is inspected at the level of individual items. Each of the 300 items was classified for each subject according to the sum of the two judgments as either perfectly additive (sum = 1.0), subadditive, or superadditive. (Only the verbal analysis using **R** is reported here—**W** results are identical.) An index of super versus subadditivity is obtained by taking the ratio of the percentages of superadditive items to subadditive items for each subject. A repeated-measures ANOVA on the logs of the ratios again showed neither a mode effect, a group effect, nor an interaction. For the numerical responses, the ratios of 11 subjects exceed 1.0 and the overall geometric mean is 0.99, which is not significantly different from 1.0. (Actually, the hypothesis that the mean log of the ratio equals zero was tested.) Under verbal responding, the ratios of 10 subjects exceed 1.0, and the overall geometric mean is 0.98, which is not significant by the same test. Thus, both response modes yield judgments that are not systematically super or subadditive.

Figure 3 **Numerical and Verbal Calibration Curves for the 13 Subjects Who Ranked *Tossup* Sixth**



2.2.4. Calibration. Calibration curves were constructed for each subject based on the numerical responses and on the two transformations of the verbal responses, **W** and **R**. In the numerical case the relative frequency of true items in each of the 11 intervals formed by rounding judgments to the nearest multiple of 0.10 was plotted as a function of the interval midpoints (except using 0 rather than 0.025 and 1.0 rather than 0.975 for the endpoints).[1] In the verbal case the relative frequencies of true items to which each of the 11 phrases was applied was plotted as a function of **W** and also of **R**. Perfect calibration would be represented by the 11 points on a graph falling on the diagonal. Inspection of the calibration curves for individual subjects revealed large differences in their smoothness and approximation to the diagonal, but considerable overlap in the verbal and numerical curves for a given subject. Mean numerical and verbal (using the **R** metric) calibration curves are shown in Figures 3 and 4. Very similar curves emerge using **W**. Figure 3 is averaged over those subjects who ranked *tossup* in the middle of the 11 phrases, and Figure 4 is averaged over those who ranked it fifth. The similarity of the two curves in both cases is substantial.

One set of measures by which the data in Figures 3 and 4 can be compared statistically derive from the Brier (1950) score, or mean probability score:

$$B = N^{-1} \sum_{i=1}^{N} (p_i - d_i)^2 \tag{3}$$

where N is the number of items judged; p_i is the probability assigned to item i; and $d_i = 0$ (if the item is false)

[1] A reviewer pointed out that the relative frequencies more properly are plotted as a function of the weighted mean judgments of the categories. However, due to the distributions of judgments, the weighted means never differed from the midpoints by more than 0.01, and the calibration curves are virtually identical plotted either way.

or 1 (if it is true). Sanders (1963) has proposed a useful decomposition of B based on the partition of the $[0, 1]$ interval into J distinct intervals. Yates (1982) has shown in the discrete case that Sander's decomposition of Equation (3) yields

$$B = N^{-1} \sum_{j=1}^{J} N_j \bar{d}_j (1 - \bar{d}_j) + N^{-1} \sum_{j=1}^{J} N_j (f_j - \bar{d}_j)^2 \qquad (4)$$

where f_j is the jth probability value (or in our case the midpoint of the jth probability interval), and $\bar{d}_j$ is the fraction of items assigned f_j or to the jth interval that are true. The first term on the right-hand side of Equation (4) is a measure of the *resolution* of the probability judgements. Note that its value is minimized at 0 by using probability intervals, such that all the events in each interval are either true or false. In general, for any number of probability intervals, the closer to 0 or 1 are the fractions of true items in each interval, the lower (and therefore better) is the resolution score, and this is so regardless of the labels attached to the intervals.

The second term on the right-hand of Equation (4) is a measure of the quality of the numerical labels, or responses, assigned to each probability interval, and is called *reliability-in-the-small* by Yates (1982). For simplicity, we will use the term *calibration* here. Lower calibration scores are better, and in fact the score is minimized at 0 when the assigned probability values exactly match the fraction of true items in each interval.[2]

Figure 4 **Numerical and Verbal Calibration Curves for the Eight Subjects Who Ranked *Tossup* Fifth**

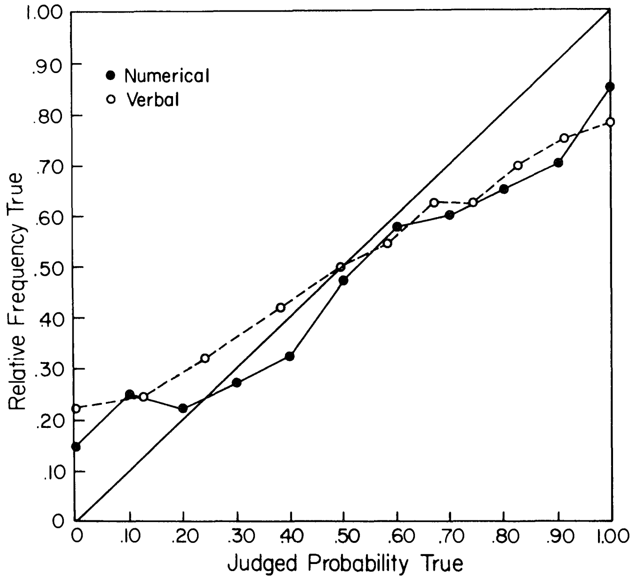


Table 3 **Sanders Decomposition of Total Probability Scores (×100)**

	Resolution		Calibration		
Subject	Numeric	Verbal	Numeric	Verbal (**R**)	Verbal (**W**)
			Group N		
1	17	17	00	01	01
2	22	23	03	02	02
3	19	18	02	03	03
4	19	20	00	01	03
5	20	23	01	06	11
6	16	17	00	00	00
7	13	14	01	00	00
Mean	18	19	01	02	03
			Group C		
8	17	17	02	02	02
9	19	18	01	01	02
10	16	16	00	00	00
11	24	24	03	05	06
12	24	23	06	05	04
13	19	21	00	01	02
14	17	18	01	01	02
15	19	19	01	01	01
Mean	19	20	02	02	03
			Group P		
16	24	24	01	02	03
17	19	19	02	02	02
18	24	24	02	06	06
19	23	23	01	02	02
20	21	21	05	04	04
21	19	20	02	03	03
Mean	21	22	02	03	03
Grand Mean	19	20	02	02	03

Table 3 presents the resolution and calibration scores (×100) by subject and group for the two response modes. Note that there is a single resolution score for the verbal responding, because its calculation does not depend on a specific quantification of the response categories. However, there are separate calibration scores

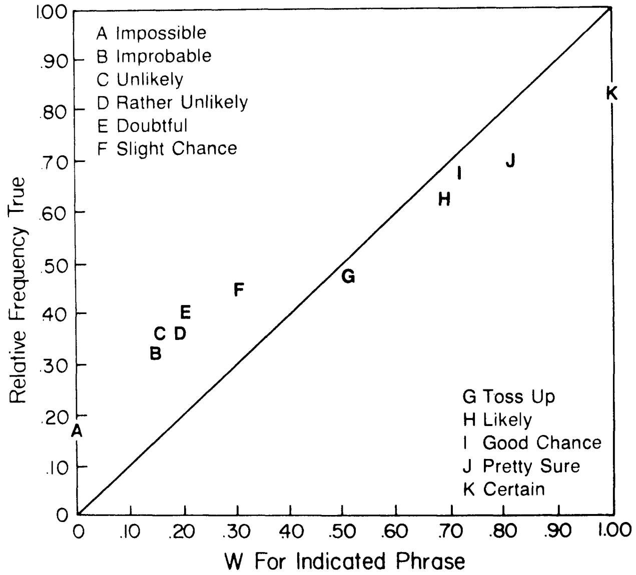
[2] Most analyses of calibration use a different decomposition due to Murphy (1973). This decomposition further breaks down the resolution score into an outcome index and resolution. Since we compare performance of the same subjects on a fixed set of items the two resolution scores are unique up to a linear transformation.

for **R** and **W**. Total probability scores in all cases are obtained by summing the two component scores. The verbal and numerical resolution scores are highly correlated ($r = 0.94$), and the calibration scores are moderately so ($r = 0.63$ and 0.30 between the numerical and the **R** or **W** metrics, respectively). Repeated measures ANOVAs failed to uncover any effect of group, response mode, or their interaction on the calibration scores, nor a group or an interaction effect on the resolution score. The mode effect on resolution was significant ($F(1, 18) = 6.43$, $p < 0.05$), albeit very small (0.19 versus 0.20).

An alternative way to average verbal responses is to do so over the 11 phrases used most frequently by the subjects rather than over the 11 ranks regardless of the phrases. Now, of course, subjects contribute unequally to the results, but one can see how individual phrases are used at a group level. The resulting calibration curve, based on mean **R** values, is shown in Figure 5, where it can be seen that the essential pattern is repeated.

2.2.5. Overconfidence. The calibration curves illustrate the subjects' overconfidence. Items given a high probability of being true were not true with the predicted relative frequency, as indicated by the calibration curves dipping below the diagonal at judgments greater than 0.50. Correspondingly, items given a low probability of being true (high probability of being false) were not false frequently enough, as indicated by the calibration curve rising above the diagonal for judgments less than 0.50.

Figure 5 **Calibration Curve Based on the 11 Most Frequently Used Phrases**



To compare the overconfidence under the two conditions we calculated a mean over/under confidence score for each subject under each condition:

$$O = J^{-1}\left[\sum_{f_j<0.5} (\bar{d}_j - f_j) + \sum_{f_j>0.5} (f_j - \bar{d}_j)\right]. \qquad (5)$$

In summing deviations from the diagonal, Equation (5) distinguishes between two cases. For low probabilities (under 0.5) deviations above the diagonal are treated as positive and deviations below the diagonal are treated as negative. The reverse is true for high probabilities (over 0.5). The equiprobable category (0.5 or *tossup*) is not included. Thus positive scores indicate overconfidence and negative ones reflect underconfidence.

The mean overconfidence score ($O \times 100\%$) for the numeric judgments is 8.8%. There are two scores for the verbal conditions, based on the two scalings of the terms. Both mean values (12% for **R** and 14.1% for **W**) are significantly larger than the mean numeric score ($F(1, 18) = 5.77$ and 6.23 respectively, $p < 0.05$) indicating a greater degree of overconfidence for verbal judgments. This result also holds for 16 of the 21 subjects.

When verbal responses are scaled by **R**, we also found a significant effect for the payoff method ($F(2, 18) = 4.06$, $p < 0.05$). The P group was the most overconfident (mean $O = 17.8\%$) and the N and C groups were practically identical (mean values of $O = 7.3\%$ and 7.6%, respectively). The payoff effect was not replicated when the words were scaled according to **W**.

Because overconfidence cannot be defined unambiguously for the central category (50% or *Tossup*), we excluded it from the calculation of our index. In order to reject any possible artificial interpretation of our previous results, we compared the percentage of true statements in this category across the various conditions. On the average, subjects were very accurate in their use of this category: 47.6% of the items in the 50% category and 48.8% of those classified as *tossup* are, in fact, true. This percentage shows no significant effect of the payoff scheme.

Table 4 **Transformed Spherical Scores (×100)**

Subject	Numeric	Verbal
	Group N	
1	81	80
2	73	72
3	77	77
4	78	76
5	77	67
6	82	81
7	85	84
Mean	79	77
	Group C	
8	79	79
9	78	78
10	81	81
11	70	67
12	66	69
13	78	75
14	80	79
15	78	77
Mean	76	76
	Group P	
16	71	70
17	77	77
18	70	66
19	73	73
20	72	74
21	76	76
Mean	73	73
Grand Mean	76	75

2.2.6. Payoffs. Table 4 shows the earned scores according to Equation (2) calculated for each subject after setting $a = 0$ and $b = 1$. Different constants were used to score subjects in groups P and N, but it is only by equating the constants over the groups that meaningful comparisons can be made. Of course, scores can be calculated for subjects in group C despite the fact that the subjects knew nothing about the scoring system. The verbal and numerical scores are positively correlated over subjects ($r = 0.87$, $p < 0.001$) and, a repeated-measures ANOVA on the scores yielded no effects of group ($F(2, 18) = 1.95$, ns), of response mode ($F(1, 18) = 3.96$, ns), or of their interaction ($F(2, 18) = 1.19$, ns).

3. Conclusions

Perhaps the most obvious conclusion to be drawn from this study is that the quality of verbal forecasts can be evaluated, at least when the analyst employs a limited, self-selected vocabulary. Under this constraint it is meaningful to assign numerical values to verbal expressions, as attested to by the facts that (a) two different methods, the modified equal spacing method and the derivation of central tendencies from membership functions, resulted in highly similar numerical values, and (b) subjecting these values to a multitude of analyses resulted in systematic and interpretable results. Moreover, the implication of these analyses is that neither the numerical nor the verbal mode of assessing uncertainty is uniformly better than the other, although the two differ in interesting ways.

Before discussing similarities and differences between the two response modes, we must emphasize the use of limited individualized vocabularies. As has occurred in other previous studies, the subjects collectively selected a large number of phrases (60 for the 21 individuals), yet each worked very well with his or her own small set. Free choice of language would have made it very difficult to rank order an individual's vocabulary or to obtain membership functions, although not necessarily impossible (see Zwick and Wallsten 1989). Possibly, another benefit of selecting phrases as was done here, or in the manner advocated by Hamm (1991), is that subjects were encouraged to think carefully about the phrases' meanings prior to using them. We cannot claim that 11 categories is the ideal number to use, but that number did work well in this study.

Because response mode and payoffs never interacted in affecting any of the dependent variables, the remainder of this discussion treats the two factors separately. We consider first similarities and differences in the two response modes and then the effects of payoffs. We conclude with some recommendations.

3.1. Similarities in Response Modes

Based on analyses at the level of individual subjects, the two response modes did not differ systematically in terms of the usual measures of quality. In terms of both the components of the Brier score and the spherical score used to pay subjects, virtually no differences emerged between the two response modes. Neither calibration nor amount earned was affected by the type of judgment. Resolution was very slightly but significantly better in the numerical mode.

An unusual feature of this study was to provide true and false statements regarding the same fact at different points in time, and to require subjects to give their subjective confidence regarding the truth of the statement in both cases. This allowed a measure of coherence, or additivity, that to our knowledge has not been employed before in the evaluation of probabilistic forecasts. Coherence was equal and excellent for both response modes. Regardless of the analysis used, there was no indication of either superadditivity or subadditivity in judgments.

3.2. Differences Between the Two Response Modes

The two modes did differ in two very interesting ways. Dividing the numerical continuum into 11 categories for purposes of comparison with the verbal mode allowed the observation that judgments were more unevenly spread over the categories in the numerical than in the verbal case. This difference was due primarily to much greater use of the 50% category in the numerical case than of the *tossup* category in the verbal case. Despite the differential frequency of use of the central categories, their accuracy was equal and excellent under the two modes.

An interesting possibility is that individuals are prone to mapping imprecise or vague feelings of confidence into the 50% category because it is equally defensible regardless of the final outcome. In contrast, a verbal response scale clearly allows the expression of vague judgments that need little defense regardless of their location along the probability interval. Consequently, people may feel comfortable providing intermediate judgments verbally, but not numerically. If this interpretation is correct, it suggests that the verbal scale may encourage somewhat greater honesty than the numerical one because individuals are not compelled to map their judgments into a response scale strategically in order to indicate level of confidence in a probability estimate.

The second difference between the two modes of responding is in the degree of overconfidence. Overconfidence was demonstrated in both modes, however, its magnitude systematically was greater given verbal than numerical responding. Continuing the interpretation from above, perhaps the greater honesty encouraged by the verbal mode allows us to see that actual overconfidence is even greater than is apparent given numerical judgments. All of this is speculative, of course, and worthy of further research.

3.3. Effects of Payoffs

Payoffs did not affect the quality of the judgments as indexed by the components of the Brier score or by the spherical score. They did, however, affect the distribution of responses and the degree of overconfidence in judgments. In particular, response distributions were more unequal in the positive than in the negative payoff condition (with the control condition between the two). Subjects tended to use the central and the extreme categories more often and the intermediate categories less often in the positive than the negative case. These results are consistent with those obtained by Wallsten (1976) in the context of a Bayesian revision of opinion task employing two payoff groups similar to the present groups P and N. In that study, the ordinal response properties of the two groups were identical over various information conditions, but the responses of group P were considerably more extreme than those of group N. Those results were interpreted as showing equivalent information processing in the two cases, but a differential mapping of judgment onto the response scale. The same conclusion might be appropriate here and, at the very least, demonstrates that an individual's overt probability estimate depends on the (explicit or implicit) payoff matrix as well as on his or her judgment of the event in question. Consistent with a more extreme use of the response scale, subjects in the positive group displayed more overconfidence in their judgments than did subjects in the other two payoff groups.

3.4. Recommendations

There is no basis in the present results for suggesting that either numerical judgments or verbal judgments using a constrained, self-selected vocabulary provides the better medium for issuing probabilistic forecasts. Advantages of the numerical mode are that everyone uses the same vocabulary and resolution is a bit better. The contrasting advantage for the verbal mode is that individual forecasters may feel more comfortable providing estimates in that format. As long as their selected vocabularies are known to the decision makers who must use the forecasts, it does not seem that information will be lost. The primary tradeoff between the two modes is that the 50% category is used more frequently numerically, while overconfidence is greater verbally. To the degree that systematic overconfidence (departure from the diagonal in the calibration plot) can be corrected by means of a suitable transformation (see Clemen and Murphy 1990), this tradeoff would seem to favor the verbal response mode.

Verbal forecasting in the manner we have suggested requires that an analyst select and scale his or her vocabulary, and that it be known to the users of the forecasts. We did not study the selection process per se, but in light of the well-established fact that the meanings of probability phrases depend on the range of expected probabilities in the situation (Clark 1990; Tiegen 1988; Wallsten et al. 1986), it seems reasonable to suggest that the selection and quantification take place with some knowledge of the range of events for which judgments ultimately will be sought, and that the users be advised of this range. There is no evidence that the context affects the ranking of terms, but it very likely affects the specific vocabulary comfortably used, the relevant anchor terms, and the numerical conversions. This issue is open, of course, but in the absence of research directed specifically to it, this consideration seems appropriate.

Two numerical conversion methods were employed in this study. The results were very similar for both and neither showed a systematic advantage in any of the evaluation measures we employed. It is certainly easier to have analysts rank order a selected set of terms, to assign fixed values to anchor terms, and to assume that the remaining terms are equally spaced between the anchors, than it is to establish membership functions. For many purposes the former procedure may be sufficiently accurate. However, because the membership functions allow a more complete representation of an individual's understanding of a phrase within a particular context, it may, in the long run, and for additional purposes not investigated here, be the preferred method.[3]

[3] This research was supported by National Science Foundation Grant BNS8608692 and BNS8908554. We thank Michele Termotto for assistance in running subjects, Sarah Lichtenstein, Martha Neal, and Amos Tversky for comments on earlier drafts, and Donald MacGregor at Decision Research in Eugene, Oregon, for supplying the items from which the stimuli were developed.

References

Bass, B. M., W. F. Cascio and E. O'Connor, "Magnitude Estimations of Frequency and Amount," *Journal of Applied Psychology*, 69 (1974), 313–320.

Beck, P. J., I. Solomon and L. A. Tomassini, "Subjective Prior Probability Distributions and Audit Risk," *Journal of Accounting Research*, 23 (1985), 37–56.

Behn, R. D. and J. W. Vaupel, *Quick Analysis for Busy Decision Makers*, Basic Books, New York; 1982.

Beyth-Marom, R., "How Probable is Probable? A Numerical Translation of Verbal Probability Expressions," *Journal of Forecasting*, 1 (1982), 257–269.

Borolan, G. and R. Degani, "A Review of Some Methods for Ranking Fuzzy Subsets," *Fuzzy Sets and Systems*, 15 (1985), 1–19.

Brier, G. W., "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review*, 78 (1950), 1–3.

Budescu, D. V. and T. S. Wallsten, "Consistency in Interpretation of Probabilistic Phrases," *Organizational Behavior and Human Decision Processes*, 36 (1985), 391–405.

—— and T. S. Wallsten, "Subjective Estimation of Precise and Vague Uncertainties," in G. Wright and P. Ayton (Eds.), *Judgmental Forecasting*, Wiley, New York; (1987), 63–81.

—— and T. S. Wallsten, "Dyadic Decisions with Numerical and Verbal Probabilities," *Organizational Behavior and Human Decision Processes*, 46 (1990), 240–263.

——, S. Weinberg and T. S. Wallsten, "Decisions Based on Numerically and Verbally Expressed Uncertainties," *Journal of Experimental Psychology: Human Perception and Performance*, 14 (1988), 281–294.

Clark, D. A., *Psychological Aspects of Uncertainty and Their Implications for Artificial Intelligence*, unpublished Ph.D. Dissertation, University of Wales Institute of Science and Technology, Cardiff, Wales, 1988.

Clark, H. H., "Comment on Mosteller and Youtz's Quantifying Probabilistic Expressions," *Statistical Science*, 5 (1990), 12–16.

Clemen, R. T. and A. H. Murphy, "The Expected Value of Frequency Calibration," *Organizational Behavior and Human Decision Processes*, 46 (1990), 102–117.

Daan, H. and A. H. Murphy, "Subjective Probability Forecasting in the Netherlands: Some Operational and Experimental Results," *Meteorologishe Rundshau*, 35 (1982), 99–112.

Dutta, A., "Reasoning with Imprecise Knowledge in Expert Systems," *Information Sciences*, 37 (1985), 3–24.

Erev, I. and B. L. Cohen, "Verbal versus Numerical Probabilities: Efficiency, Biases, and the Preference Paradox," *Organizational Behavior and Human Decision Processes*, 45 (1990), 1–18.

Fischhoff, B. and D. MacGregor, "Subjective Confidence in Forecasts," *Journal of Forecasting*, 1 (1982), 155–172.

Fox, J., D. C. Barber and K. D. Bardhan, "Alternative to Bayes? A Quantitative Comparison with Rule-based Diagnostic Inference," *Methods of Information in Medicine*, 19 (1980), 210–215.

González-Vallejo, C. C., I. Erev and T. S. Wallsten, "Do Decision Quality and Preference Order Depend on Whether Probabilities are Verbal or Numerical?," *American J. Psychology* (in press).

—— and T. S. Wallsten, "Effects of Probability Mode on Probability Reversal," *Journal of Experimental Psychology: Learning, Memory and Cognition*, 18 (1992), 855–864.

Hamm, R. M., "Selection of Verbal Probabilities: A Solution for Some Problems of Verbal Probability Expression," *Organizational Behavior and Human Decision Processes*, 48 (1991), 193–223.

Jaffee-Katz, A., D. V. Budescu and T. S. Wallsten, "Timed Magnitude Comparisons of Numerical and Nonnumerical Expressions of Uncertainty," *Memory and Cognition*, 17 (1989), 249–264.

Johnson, E. M., *Numerical Encoding of Qualitative Expressions of Uncertainty*, Technical Paper 250, U.S. Army Research Institute for the Behavioral and Social Sciences, Arlington, VA, 1973.

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

Keren, G., "Facing Uncertainty in the Game of Bridge," *Organizational Behavior and Human Decision Processes*, 39 (1987), 98–114.

——, "Calibration and Probability Judgments: Conceptual and Methodological Issues," *Acta Psychologica*, 77 (1991), 217–273.

Lichtenstein, S., B. Fischhoff and L. D. Phillips, "Calibration of Probabilities: The State of the Art to 1980," In D. Kahneman, P. Slovic and A. Tversky (Eds.), *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge, England, 1982, 306–334.

Mosteller, F. and C. Youtz, "Quantifying Probabilistic Expressions," *Statistical Science*, 5 (1990), 2–16.

Murphy, A., "A New Vector Partition of the Probability Score," *Journal of Applied Meteorology*, 12 (1973), 595–600.

—— and R. L. Winkler, "Reliability of Subjective Probability Forecasts of Precipitation and Temperature," *Journal of the Royal Statistical Society*, 26 (1977), 41–47.

Rapoport, A., T. S. Wallsten and J. A. Cox, "Direct and Indirect Scaling of Membership Functions of Probability Phrases," *Mathematical Modelling*, 9 (1987), 397–417.

——, ——, I. Erev and B. L. Cohen, "Revision of Opinion with Verbally and Numerically Expressed Uncertainties," *Acta Psychologica*, 74, (1990), 61–79.

Ronis, D. L. and J. F. Yates, "Components of Probability Judgment Accuracy: Individual Consistency and Effects of Subject Matter and Assessment Method," *Organizational Behavior and Human Decision Processes*, 40 (1987), 193–218.

Sanders, F., "On Subjective Probability Forecasting," *Journal of Applied Meteorology*, 2 (1963), 191–201.

Stäel von Holstein, C., "Measurement of Subjective Probability," *Acta Psychologica*, 34 (1970), 146–159.

Teigen, K. H., "When Are Low-probability Events Judged to Be Probable? Effects of Outcome-set Characteristics on Verbal Probability Estimates," *Acta Psychologica*, 68 (1988), 157–174.

von Winterfeldt, D. and W. Edwards, *Decision Analysis and Behavioral Research*, Cambridge University Press, Cambridge; 1986.

Wallsten, T. S., "Using Conjoint Measurement Models to Investigate a Theory About Probabilistic Information Processing," *Journal of Mathematical Psychology*, 14 (1976), 144–185.

—— and D. V. Budescu, "Encoding Subjective Probabilities: A Psychological and Psychometric Review," *Management Science*, 29 (1983), 151–173.

——, "Measuring Vague Uncertainties and Understanding Their Use in Decision Making," In G. M. von Furstenberg (Ed.), *Acting Under Uncertainty*, Kluwer, Norwell, MA, 1990a, 377–398.

——, "The Costs and Benefits of Vague Information," In R. Hogarth (Ed.), *Insights in Decision Making. A Tribute to the Late Hillel Einhorn*, University of Chicago Press, Chicago, IL, 1990b, 28–43.

—— and D. V. Budescu, "Comment on Mosteller and Youtz' Quantifying Probabilistic Expressions," *Statistical Science*, 5 (1990), 23–26.

——, ——, A. Rapoport, R. Zwick and B. Forsyth, "Measuring the Vague Meanings of Probability Terms," *Journal of Experimental Psychology: General*, 115 (1986), 348–365.

——, ——, R. Zwick and S. Kemp, Preferences and Reasons for Communicating Probabilistic Information in Verbal or Numerical Terms, *Bulletin of the Psychonomic Society* (in press).

——, S. Fillenbaum and J. A. Cox, "Base Rate Effects on the Interpretations of Probability and Frequency Expressions," *Journal of Memory and Language*, 25 (1986), 571–587.

Wright, G., "Changes in the Realism and Distribution of Probability Assessments as a Function of Question Type," *Acta Psychologica*, 52 (1982), 165–174.

Yager, R. R., "A Note on Probabilities of Fuzzy Events," *Information Sciences*, 18 (1979), 113–129.

Yates, J. F., "External Correspondence: Decompositions of the Mean Probability Score," *Organizational Behavior and Human Performance*, 30 (1982), 132–156.

Zadeh, L. A., "The Concept of a Linguistic Variable and Its Application to Approximate Reasoning," Parts 1, 2, and 3, *Information Science*, 8 (1975), 199–249; 8, 301–357; 9, 43–96.

Zimmer, A. C., "Verbal vs. Numerical Processing of Subjective Probabilities," In R. W. Scholz (Ed.), *Decision Making Under Uncertainty*, North-Holland Publishing Company, Amsterdam, 1983, 159–182.

——, "A Model for the Interpretation of Verbal Predictions," *International Journal of Man-Machine Studies*, 20 (1984), 121–134.

Zwick, R. and T. S. Wallsten, "Combining Stochastic Uncertainty and Linguistic Inexactness: Theory and Experimental Evaluation of Four Fuzzy Probability Models," *International Journal of Man-Machine Studies*, 30 (1989), 69–111.

Accepted by Gregory W. Fischer; received January 25, 1991. This paper has been with the authors 2 months for 1 revision.