Journal of Behavioral Decision Making, Vol. 6, 283–296 (1993)

An Evaluation of the Reliability of Probability Judgments Across Response Modes and Over Time

KATHLEEN M. WHITCOMB
University of South Carolina, USA

DILEK ÖNKAL
Bilkent University, Ankara, Turkey

P. GEORGE BENSON
Rutgers University, USA

SHAWN P. CURLEY
University of Minnesota, USA

ABSTRACT

Despite the importance of probability assessment methods in behavioral decision theory and decision analysis, little attention has been directed at evaluating their reliability and validity. In fact, no comprehensive study of reliability has been undertaken. Since reliability is a necessary condition for validity, this oversight is significant. The present study was motivated by that oversight. We investigated the reliability of probability measures derived from three response modes: numerical probabilities, pie diagrams, and odds. Unlike previous studies, the experiment was designed to distinguish systematic deviations in probability judgments, such as those due to experience or practice, from random deviations. It was found that subjects assessed probabilities reliably for all three assessment methods regardless of the reliability measures employed. However, a small but statistically significant decrease over time in the magnitudes of assessed probabilities was observed. This effect was linked to a decrease in subjects' overconfidence during the course of the experiment.

KEY WORDS Reliability Probability assessment Probability judgment Subjective probability

In probability assessment tasks, the concept of reliability is concerned with the internal consistency of assessments. The reliability of a probability assessment method is typically evaluated by measuring the strength of the linear relationship between two independent encodings for the same set of events, made by the same assessor with the same level of relevant knowledge. High correlations indicate that the assessments are 'reliable' in the sense that they are relatively free from random error (Wallsten and Budescu, 1983). A stronger form of reliability requires assessments to be reasonably free from both random and systematic error. In this case, a probability assessment method is reliable to the extent that a bivariate plot of the second set of assessments versus the first falls close to the identity line (Wallsten and Budescu, 1987). Perfect reliability is achieved when the bivariate plot forms a line with unit slope and zero intercept.

0894–3257/93/040283–14$12.00

Received 4 December 1992
Accepted 19 April 1993

The question of whether an assessment method is reliable is absolutely fundamental. If it is not reliable, a method is neither valid nor can it transmit meaningful information (Wallsten and Budescu, 1983). Despite its importance, the evaluation of the reliability of probability assessment methods has received little attention. Only a few studies have reported reliability results and in each of these reliability was, at best, a secondary concern in the design of the study. Moreover, most of the studies reported reliability as correlational reliability. To date, no comprehensive studies have been conducted that were specifically designed to determine the extent to which probability judgments are free from both random and systematic error. Nor have studies been carried out to determine which methods of assessment are reliable, or perhaps whether some methods are more reliable than others.

In this paper we present the results of an experiment that was conducted to evaluate and compare the reliability of probability measures derived from three response modes: numerical probabilities, pie diagrams, and odds. The next section reviews the results of previous research pertaining to reliability. The third section describes the experimental methodology employed in this study and the measures used to evaluate the reliability of probability assessments. The results are presented in the fourth section and discussed in the final section.

PRIOR RESEARCH IN RELIABILITY

In this section we describe the findings and experimental methods of four studies providing reliability results. In so doing, we identify various limitations in the design of previous studies and establish a basis for evaluating the results of the current study. Additional studies related to the reliability of probability assessments do exist. For example, Wallsten *et al.* (1983) reported high correlational reliabilities for the endpoints of subjectively assessed probability intervals (mean value for 8 subjects = 0.903), but noted that these values were in part attributable to the imposed coherency requirements. Carroll (1971) and Carroll and Lamendella (1974) evaluated reliabilities for subjectively assessed word and phoneme frequencies, observing reliabilities of 0.801 and 0.479, respectively. However, the reliability measure used in these studies did not evaluate test–retest correlational reliability for each subject, but reflected the total variation in the assessed frequencies 'explained' by subject differences and item (word or phoneme) differences (Ebel, 1951). Only the four studies detailed below provide information on the reliability of probability judgments as defined in the previous section.

The first of these studies was conducted by Peterson *et al.*, (1965). The primary aim of their experiment was to measure the extent to which subjectively assessed probabilities conformed to the multiplicative law of probability. Twelve introductory psychology students assessed subjective probabilities for twenty different characteristics of individuals such as 'good', 'dishonest', 'witty' etc. Direct judgments on a scale of 0 to 100 were made first for the unconditional characteristics, and then for the characteristics conditional upon the other characteristics. Subjects gave two different assessments for each of the unconditional probabilities, separated over time, thereby providing test–retest data from which reliability measures were computed. Only linear correlations were reported. The highest and lowest correlations were 0.91 and 0.53, respectively, with a mean of 0.72.

Beach (1966) investigated whether subjects revise their subjective probabilities according to Bayes's Theorem. Thirty-six subjects assessed the probability that a card belonged to a particular class of cards, based on a set of cues. The subjects were not told the relative frequencies with which the different cues co-occurred with each class, but were allowed to view each set of cards twice. Each set consisted of a large number of cards so that subjects would not be able to estimate frequencies accurately. After seeing one cue, the subjects assessed the probability that the card belonged to

each of the possible classes. Probability judgments were encoded using a partitioning method in which subjects slid markers along an unmarked metallic bar. The distances to the left and right of the marker corresponded to the probabilities for the event of interest and the complementary event, respectively.

The subjects were then shown a second cue and asked to revise their probability that the card belonged to each of the possible classes in light of the new information. Recognizing that proper evaluation of the subjects' revised probabilities required knowledge of whether the assessments were stable, Beach re-tested each subject using five of the original set of test cards. Correlations between the first and second set of five probability assessments were computed for each subject. Only the results for the 15 subjects with the highest correlations were reported. The other 21 subjects were presumably considered too unreliable to be included in the remaining analyses. The mean correlation for the 15 'reliable' subjects was 0.82 with maximum and minimum values of 0.99 and 0.68, respectively.

Branthwaite (1974) studied the inter-response correlational validity of three methods of probability assessment — decision time, times-out-of-ten, and numerical probabilities expressed as percentages (confidence ratings). Twelve subjects were asked to make probability judgments as to whether or not a ball could be rolled through a gap of varying width (eight widths in all). For each width, subjects were asked to make assessments in each of three ways. In Method 1, subjects responded 'yes' or 'no' to whether they could roll a ball through the gap. The time to make a decision was recorded. In Method 2, subjects were asked how many times out of ten they could roll a ball through a hole. Finally, subjects were asked how confident they were that they could roll a ball through the hole. After they completed the experiment, they waited 5 minutes and repeated the experiment using a different random order of presentations for the widths of the gaps. Correlations were calculated for the test–retest data for each method of assessment. Branthwaite found that the mean correlations between the first and second set of assessments using decision time, times-out-of-ten, and confidence ratings were 0.73, 0.96, and 0.89, respectively.

Goodman (1973) summarized the results of a series of experiments conducted at the Engineering Psychology Laboratory at the University of Michigan. Three of the experiments contained test–retest data that could be used to compute reliability measures. The experiments investigated a variety of assessment methodologies — 13 in all. Subjects made judgments using either cumulative or non-cumulative likelihood ratios or cumulative or non-cumulative odds. They either provided a verbal report of their uncertainty judgments or marked their responses on log paper. The experiments used both between-subjects and within-subjects designs. The test–retest data were obtained during the same assessment session, or on two separate days.

In computing reliability measures, Goodman pooled responses across subjects within each of the assessment methodologies for each experiment. Accordingly, 13 (6 + 3 + 4) group measures of reliability were computed. The group correlational reliability was generally high, with the exception of three groups where subjects were required to use verbal reports of odds and likelihood ratios. The mean group correlation for the 13 cases was 0.883 with maximum and minimum values of 0.977 and 0.657, respectively. Mean signed deviations between the test and retest data for each group ranged from −0.139 to 0.17.

While all the studies described above indicate that correlational reliability is reasonably high for a wide range of experimental situations, they do not establish the extent to which deviations in probability assessments obtained in different sessions are random versus systematic in nature. Systematic differences could be due to factors such as experience with the assessment method or increased knowledge regarding the events for which probabilities are assessed.

The current study addresses this issue by measuring both random and systematic deviations in probabilities assessed in different time periods using the same set of questions, same assessor, and same assessment methodology.

METHODOLOGY

The experiment was designed to evaluate the reliability of probability measures derived from the three response modes — numerical probabilities, pie diagrams, and odds. Four measures were used to evaluate the reliability of assessed probabilities: linear correlation, mean absolute deviation, mean signed deviation, and the least squares linear regression line. Linear correlation and mean absolute deviation were used to detect random deviation, while mean signed deviation and least squares regression were used for detecting systematic differences between two sets of probability assessments. Probabilities assessed using pie diagrams and odds were converted to numerical probabilities prior to computing reliability measures.

Whereas previous reliability results were obtained from experiments limited to two assessment sessions, the current study employed three. Since deviations in probability assessments due to the effects of practice are expected to diminish over time, extending the number of sessions to three allowed us to differentiate the effects of practice from effects due to the reliability of a particular method of assessment. Additional features and details of the experiment are described in the following sections.

Subjects

Forty-two subjects participated in the study. Twenty-four were upperclassmen and students in various masters degree programs in the College of Business Administration at the University of South Carolina. The remaining 18 subjects were upperclassmen and masters degree students in the College of Business Administration at Bilkent University, Ankara, Turkey. An equal number of undergraduate and masters students were used at both universities.

Procedure

Subjects participated individually in three probability assessment sessions scheduled at least one week apart. Each session was divided into halves. During a session-half, subjects were required to respond to 50 general knowledge questions by choosing one of two possible alternatives and assessing the likelihood that their chosen answer was correct. Each subject used all three methods of probability assessment, one per session-half. Thus, each subject used all three methods exactly twice throughout the course of their participation in the experiment. Care was taken to ensure that the questions used in the experiment were culturally neutral. Three examples of these questions are:

1. What Olympic sport finds competitors using equipment made by Anschutz and Remington?
   a. fencing
   b. shooting
2. Whose heart generally beats faster?
   a. an infant's
   b. a teenager's
3. Do people who are born blind experience rapid eye movement during sleep?
   a. yes
   b. no

General knowledge questions were employed in order to facilitate comparison with previous reliability results, and to enable subjects to maintain a constant knowledge level during the course of their participation in the study, a necessary condition for measuring test–retest reliability. While the questions used in this study do not reflect the dynamic nature of real-world forecasting tasks, they do

.AJ92BD145 00 000007446 0515 00187

require complex reasoning. We will return to the issue of generalizability after the experimental results are presented.

Prior to the first session-half, subjects received individualized training in probability assessment and in three performance measures used to evaluate the accuracy of probability assessments — the mean probability score, overconfidence, and slope. These measures evaluate the overall accuracy, calibration, and discrimination, respectively, for a set of probability assessments and can be computed for a relatively small number of assessments. Formulas and definitions for the mean probability score, overconfidence, and slope are given in the Appendix. Training in performance measures was provided so that subjects knew that they were accountable for their responses and to discourage bluffing or hedging. Subjects were given a set of practice questions to complete. The correct answers, and the mean probability, overconfidence, and slope scores, were provided to subjects as feedback. Shorter practice sessions were given to subjects before the start of the second and third sessions in order to review assessment methodologies and scoring rules, and to ensure that subjects remained motivated.

Probability assessment methods

Numerical probabilities and likelihood ratios are the two most commonly used and studied non-verbal forms of uncertainty measures (von Winterfeldt and Edwards, 1986). All the studies cited in the previous section involved numerical probability assessments and/or one or more forms of likelihood ratios. Numerical probabilities tend to be preferred by individuals with technical backgrounds, whereas likelihood ratio methods are often favored by those less quantitatively oriented (von Winterfeldt and Edwards, 1986). The present study employed numerical probabilities, partioning of a visual representation (pie diagrams), and likelihood ratios (odds).

Numerical probabilities method

Using the numerical probabilities method, subjects assessed the probability that their chosen answer was correct by choosing a number between 0.5 and 1.0. Subjects were told that a probability of 0.5 meant that their chosen answer was no more likely to be correct than the alternative that they did not choose, and that assessing a probability of 1.0 meant that they were certain that their chosen alternative was correct. They were informed that the more strongly they believed their answer to be correct, the closer their assessed probability should be to 1.0.

Pie diagram method

Using the pie diagram method, subjects assessed the probability that their chosen answer was correct by designating an angle between 180° and 360° in a pre-drawn circle. Subjects were informed that an angle of 180° meant that the answer they chose was no more likely to be correct than the alternative, and that an angle of 360° indicated that they were certain that their chosen answer was correct. They were advised that the more strongly they believed their answer to be correct, the closer their assessed angle should be to 360°.

Odds method

Using the odds method, subjects assessed the probability that their chosen answer was correct by stating odds in favor of their chosen answer. They were required to assign odds of x:1, where x could be any whole number or decimal. Subjects were instructed that odds of 1:1 indicated that they did not believe their answer was any more likely to be correct than the alternative, and that

assessing odds where the first number was very much larger than the second indicated that they were very sure that their answer was correct. Subjects were not specifically told that the certainty equivalent for odds was infinity:1, but were told that the more strongly they believed their chosen alternative to be correct, the larger the ratio should be.

Since the odds method is open-ended on one side of the scale, particular care was taken to stress that odds of, say, 10:1 meant that the answer they chose was ten times as likely to be correct, 100:1 meant that their answer was 100 times as likely to be correct, etc. By placing emphasis on the relative likelihood interpretation of odds, it was hoped that subjects would be discouraged from, for example, thinking of odds of 20:1 as being 'moderate' because they had used odds of 100:1 for answers for which they were very certain.

Experimental design

The design of the experiment accounted for the effects of session and method on the reliability of probability assessments. A balanced incomplete block design was employed, where the blocking factor was subjects. In this design, each subject used all three methods of probability assessment and participated in all three sessions, but used only two methods per session. The order of assessment methods was randomized across sessions. The design also included a blocking variable to account for possible effects due to nationality (US and Turkish). Within each nationality, subjects were randomly assigned to one of the six assessment orderings.

One hundred general knowledge questions were used in each assessment session. Fifty questions were used for each assessment method in a session-half. Common blocks of 25 questions were distributed across the session-halves so that reliability measures could be computed for assessments in the first and second, first and third, and second and third sessions. These common blocks are hereafter referred to as B_{12}, B_{13}, and B_{23}, respectively. For each session half, the order in which the common block of questions was presented was randomized to reduce the likelihood that subjects could recall their responses from the previous session.

Reliability measures were computed for each subject for each of the common blocks of questions, B_{12}, B_{13}, and B_{23}. Changes in probability assessments due to the effects of experience could thereby be detected by comparing reliability measures between the common blocks of questions in different pairs of sessions. For example, better reliability scores for B_{23} relative to B_{12} would suggest a lack of experience with probability assessment or with a particular method of probability assessment in the initial session. If, in addition, the reliability scores for B_{13} were worse than those for B_{12}, this would indicate that the effects of experience were still present, but had lessened by the final assessment session. This pattern of results could be realized in the case where the subject had an initial tendency to assess high probabilities for the event of interest, and then adjusted his or her probabilities downward in the second question, and somewhat less downward in the third session.

In order to limit the effects of increased knowledge on probability assessments, subjects were cautioned not to find the answers to the general knowledge questions between sessions. Subjects were requested to report instances when they inadvertently received new information regarding any of the questions they had previously seen. In those situations, the question was eliminated from further analyses.

Response variables

Four measures were used to evaluate the reliability of probabilities assessed for the common blocks of questions: linear correlation, mean absolute deviation, mean signed deviation, and the least squares regression line. In all cases, individual, rather than group, reliability measures were computed.

Linear correlation

The Pearson product-moment coefficient of correlation, r, was used to measure the strength of the linear relationship between two sets of probability assessments. Perfect reliability requires that $r = 1.0$. While high linear correlation is a necessary condition for high reliability, it is not a sufficient condition since r is unaffected by linear transformations and relatively robust to departures from linearity (Wallsten and Budescu, 1983).

Mean absolute deviation

The mean absolute deviation (MAD) between two sets of probability assessments was used to investigate the differences between two sets. The minimum (best) and maximum (worst) values for MAD are zero and one, respectively. A small value for the mean absolute deviation is both a necessary and sufficient condition for high reliability. However, the mean absolute deviation does not provide information for determining whether the deviations are systematic or random in nature.

Mean signed deviation

The mean signed deviation (MSD) was used to reveal systematic differences between probabilities assessed in earlier versus later sessions. Specifically, MSDs are able to detect cases where probabilities assessed in later sessions are consistently smaller or consistently larger than probabilities assessed in earlier ones.

Least squares regression line

The probabilities from the later session were regressed on corresponding probabilities from the previous session using least squares regression. Perfect reliability requires that the slope of the regression line equals one and the y-intercept equals zero. Significant deviations from the identity line can indicate systematic changes in adjudged probabilities from one session to the next. For example, a regression line that consistently fell below the identity line would indicate that probabilities assessed in the second session were consistently smaller than those assessed in the first.

RESULTS

Separate ANOVAs were conducted for each of the four reliability measures. In each case, the response vector consisted of 126 measures (42 subjects × three pairs of sessions). Since there were an unequal number of subjects nested within each culture (24 US versus 18 Turkish subjects), the analysis of variance was conducted using the regression approach. The F-tests were based on the appropriate full and reduced regression models. In cases where the F-tests indicated that the factor level means differed, Tukey's pairwise comparison procedure was used to examine the nature of the differences (Neter *et al.*, 1990).

The first three data columns of Exhibit 1 show the mean values for each of the reliability measures corresponding to the between-session blocks B_{12}, B_{13}, and B_{23}. Associated p-values are in column 4. Columns 5 through 8 report the mean values of the reliability measures for the three probability assessment methods — numerical probabilities (N), pie diagram (PD), and odds — and their corresponding p-values. No significant two- or three-way interactions were detected between session, method, and culture for any of the four ANOVAs. Consequently, the means and p-values corresponding to interaction effects are not reported.

Exhibit 1. Mean values of reliability measures for between-session blocks and N, PD, and Odds probability assessment methods

	B_{12}	B_{13}	B_{23}	*p*-value ($F_{2,68}$)	N	PD	Odds	*p*-value ($F_{2,68}$)
r	0.741	0.694	0.737	0.464	0.712	0.752	0.707	0.323
MAD	0.080	0.085	0.067	0.013	0.075	0.078	0.077	0.903
MSD	−0.029	−0.035	−0.019	0.209	−0.026	−0.028	−0.029	0.912
$\hat{\beta}_0$	0.086	0.090	0.091	0.995	0.092	0.080	0.096	0.908
$\hat{\beta}_1$	0.828	0.790	0.849	0.476	0.807	0.840	0.820	0.795

Linear correlation

No significant differences in linear correlation (*r*) were detected for B_{12}, B_{13}, and B_{23}. Similarly, the linear correlations associated with the three assessment methods were not significantly different. While the magnitudes of the mean correlations seem somewhat low, these values must be interpreted with caution. The assessment procedure used in the present study was a half-range task, which constrained the range of the probabilities from 0.5 to 1.0. This restriction tends to decrease the absolute value of the linear correlation, *r*, relative to the absolute value that would have obtained using a full-range task (Weisberg, 1980).

To illustrate this effect on the results of our study, we randomly selected six subjects and artificially increased the range of their assessed probabilities by replacing 12 of the 25 probabilities for the common blocks of questions with their complements and then recalculating the correlations. Whereas the original mean correlations for B_{12}, B_{23}, and B_{13}, were 0.739, 0.744, and 0.743, respectively, the corresponding adjusted mean correlations were 0.870, 0.843, and 0.838. This general effect would be expected to extend to all subjects in the study. Consequently, in addition to concluding that neither session nor method effects were detected, we conclude that subjects' probability assessments were reliable, in terms of correlation, for all sessions and all methods.

Mean absolute deviation

The MADs corresponding to the three between-session blocks B_{12}, B_{13}, and B_{23} differed significantly. Tukey's pairwise comparison procedure indicated that the mean MAD for B_{23} was significantly lower than that of B_{13}. While this suggests that subjects' ability to assess probabilities reliably improved over time, this result may be more a function of the power of the test than any practical difference in the means. The magnitude of the difference was estimated to be between −0.034 and −0.002 using a family confidence coefficient of 95%. No significant differences were detected in the three methods of probability assessment.

The mean values for MADs for all three between-session blocks and all three method effects indicate a reasonably small amount of difference, or variation, between probabilities assessed for common blocks of questions in different assessment sessions. Accordingly, we conclude that the subjects in this study were able to assess probabilities reliably, as measured by MAD.

An additional analysis using mean absolute deviations was conducted to evaluate the reliability of the three assessment methods at the extreme end of the probability range (numerical probabilities and numerical equivalents for odds and pie diagrams greater than 0.95). Since large differences in extreme odds translate to relatively small differences in probabilities, it seems plausible that extreme odds that have been converted to numerical probabilities may be more stable, and hence more reliable, than extreme values obtained using numerical probabilities or pie diagrams. However, the MADs for extreme responses were comparable to those for all responses, as reported in Exhibit 1. The mean MADs for the N, Odds, and PD methods were 0.078, 0.057, and 0.071, respectively ($F_{2,68}$

p-value = 0.419). Interestingly, the numbers of extreme and certainty responses assessed by subjects using odds[1] (170 and 92) were substantially less than those assessed using either numerical probabilities (298 and 215) or pie diagrams (368 and 257). These findings are contrary to previous studies that reported that subjects assessed more extreme responses using odds than using numerical probabilities (e.g. Phillips and Edwards, 1966). Wright *et al.* (1988) found that subjects used three times as many certainty responses with odds than with numerical probabilities. Perhaps our result was due to the emphasis placed on the likelihood ratio interpretation of odds and the training subjects received with performance measures such as overconfidence.

Mean signed deviation

With respect to MSDs, no significant differences were detected among the three between-session blocks. Nor were any significant differences in MSDs found between the three different methods of probability assessment. However, two observations can be made about the magnitude and sign of the six means. First, the MSDs for all three between-session blocks and all three probability assessment methods were well within the range of MSDs for the reliability data for the series of 13 experiments reported by Goodman (1974). Thus, like Goodman, we conclude that subjects assessed probabilities very reliably with respect to the MSD criterion. Second, the fact that all MSDs were negative suggests systematic deviations across different sessions. Specifically, the probability assessments were slightly smaller in later sessions. To further substantiate the existence of this effect, six simultaneous single-degree-of-freedom tests were conducted to examine the null hypothesis that the mean MSD was equal to zero versus the null hypothesis that it was less than zero (three tests for the mean MSDs corresponding to between-session effects, and three tests for the mean MSDs corresponding to assessment method effects). The family confidence level was controlled at $\alpha = 0.05$. All these tests indicated that the mean MSDs were significantly less than zero.

Least squares regression line

As indicated in Exhibit 1, no significant differences were detected in either the mean slopes or the mean *Y*-intercepts for the estimated regression equations corresponding to the three between-session blocks.[2] Similarly, no significant differences were found in the parameter estimates for the least squares regression equations corresponding to the three probability assessment methods. However, all six regression lines were significantly different from the identity line. Specifically, simultaneous single-degree-of-freedom tests indicated that the *Y*-intercepts for all six equations were significantly greater than zero and the estimated slopes for all six equations were significantly less than one. These results signal the existence of a systematic difference between probabilities for common blocks of questions assessed in later versus earlier sessions.

The nature of this difference can be identified by plotting the regression lines and comparing them to the identity line. Exhibit 2 shows the regression lines for the three between-session blocks, B_{12}, B_{13}, and B_{23}, relative to the identity line. Exhibit 3 compares the regression lines of the three assessment methods to the identity line. All six of the regression lines exhibit the same pattern: lying below the identity line over most of the available range. Since the regression lines were computed

---

[1] For purposes of the analysis, odds of 200:1 or greater were coded as certainty responses. This approach is consistent with the procedure used by Wright *et al.* (1988).

[2] Linearity of the model was checked by conducting two-tailed hypothesis tests for each of the 126 regression analyses (42 subjects × 3 regression equations/subject): 120 tested as having a significantly positive slope (*p*-value < 0.05). Failure of the remaining six cases to attain statistical significance appeared to be due to the relatively narrow ranges of values for assessed probabilities.

by regressing probabilities in the later session on probabilities in the earlier session, this pattern indicates that probabilities assessed in later sessions were generally smaller than those assessed in earlier ones, as previously observed. Exhibits 2 and 3 also support the earlier assertion that subjects participating in this experiment assessed probabilities reliably. The deviations from the identity line are relatively small in magnitude for all six regression lines.

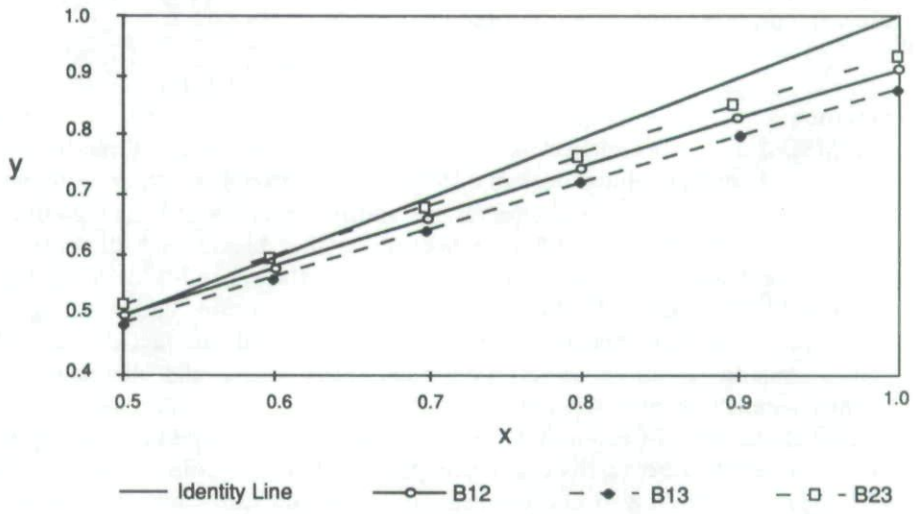

Exhibit 2. Least squares regression lines for Blocks B_{12}, B_{13}, and B_{23}

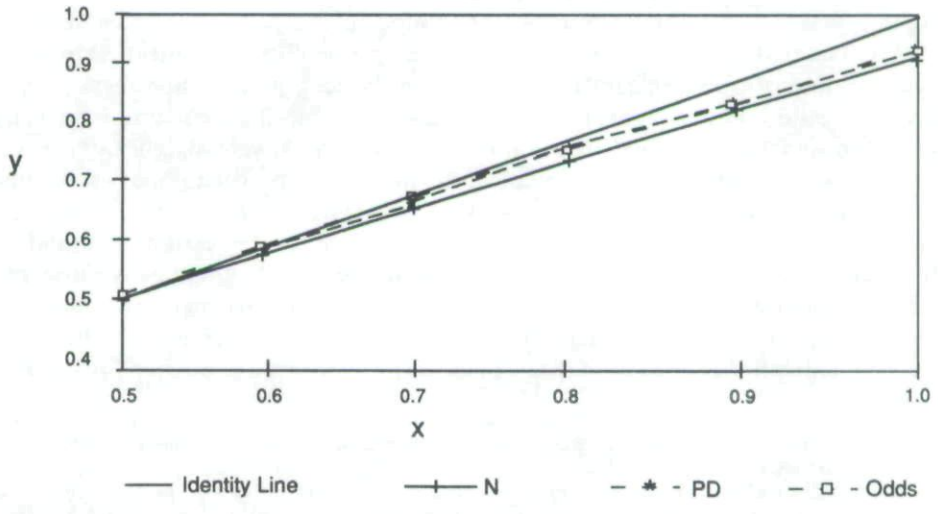

Exhibit 3. Least squares regression lines for methods N, PD and odds

DISCUSSION

Overall, subjects were able to assess probabilities reliably in terms of each of the four measures of reliability employed in this study. Further, none of the three methods of assessment was found to be any more reliable than any other. With respect to session effects, three reliability measures — correlation, MSD, and the least squares regression line — indicated no significant differences. However, MAD revealed a small, statistically significant difference between the reliability using Blocks 1 and 3 (B_{13}) versus Blocks 2 and 3 (B_{23}). Specifically, the MAD for B_{23} was significantly smaller than that for B_{13}, suggesting that the effects of practice on the reliability of probability judgments may diminish somewhat over time.

The two measures that reflect systematic differences in probability assessments — mean signed deviation (MSD) and the least squares regression line (LSR) — indicated that probabilities assessed in later sessions were somewhat smaller than those assessed in previous ones. We investigated the possibility that the systematic decrease in the magnitude of probabilities was related to changes in subjects' probability judgment accuracy over time. It was expected that subjects' accuracy would improve over time as a result of increased practice and training with probability assessment methods and accuracy measures.

Interestingly, increased training and practice did not result in demonstrably improved accuracy, either in terms of overall accuracy or underlying dimensions such as overconfidence, slope, calibration, and resolution. The overconfidence measure did decrease from one session to the next, but the decrease was not significant (0.045, 0.034, 0.001, p-value = 0.303). Still, since overconfidence reflects an assessor's tendency to assess probabilities that are inappropriately high, the decrease in overconfidence could help explain a corresponding decrease in the magnitude of assessed probabilities.

While the remaining accuracy measures were not useful in explaining systematic session to session changes in probability assessments, they did indicate that subjects possessed a reasonable degree of probability judgment accuracy. For example, the mean probability scores were 0.229, 0.214, and 0.220 for assessment sessions one, two, and three, respectively. They were 0.216, 0.222, and 0.225 for the numerical probabilities, pie diagram, and odds methods, respectively. These scores are all better than the best score that could be achieved with no knowledge of the questions, a score of 0.25, and are comparable to the accuracy observed in past studies using similar subject populations (e.g. Yates *et al.*, 1989; Benson and Önkal, 1992).

The design of the experiment also made it possible to compute inter-response mode correlational validity. The experiment was constructed so that common blocks of 25 questions occurred in each session-half within a session, making it possible to compute correlations between probabilities assessed using two different assessment methods within each session. The mean inter-response mode correlations for the N and PD, N and odds, and PD and odds methods were 0.822, 0.823, and 0.768, respectively. These values are indicative of reasonably high construct validity, particularly given that the probabilities were assessed using a half-range task. These results, along with those for the probability accuracy measures, indirectly support our assertion that subjects assessed probabilities reliably, since reliability is a necessary condition for both probability judgment accuracy and construct validity.

Moving beyond the current study, one area of concern is the generalizability of these results to forecasting situations. The applicability of results derived from studies using general knowledge questions to judgmental forecasting has been argued pro and con (e.g. Fischhoff and MacGregor, 1982; Wright and Ayton, 1986; Ronis and Yates, 1987; and Benson and Önkal, 1992). Arguments against generalizability center on findings that probability judgments for general knowledge questions are more overconfident and contain more certainty responses than those for future events. Our results, however, run counter to these findings. The subjects in the present study responded with relatively

fewer extreme and certainty responses than has been observed in comparable studies where subjects assessed probabilities for future events (cf. Wright and Ayton, 1988). This result is probably linked to the extensive training our subjects received. Further, if probability judgments for forecasting tasks do exhibit less overconfidence, then they should also be more reliable, since declining overconfidence appears to lessen session to session decreases in the magnitude of assessed probabilities.

Thus, our evidence suggests that probability forecasters would assess probabilities at least as reliably as individuals assessing probabilities for general knowledge questions. However, studies employing probability forecasting tasks are needed to substantiate this inference. Perhaps the most challenging aspect of conducting studies that evaluate the test–retest reliability of probability forecasts will be the construction of tasks for which subjects' knowledge levels remain constant from one forecasting session to the next.

Another area for future reliability research concerns the use of expert subjects assessing probabilities within their fields of expertise. Again, we see no reason why the reliability of probabilities assessed under these conditions should not equal or exceed reliabilities obtained in the present study. As noted by Wallsten and Budescu (1983):

> There is no reason to think that experts should be worse than nonexperts in this regard [reliability]; on the contrary, when they are evaluating events with which they are highly familiar and which to them are quite concrete they will probably exceed nonexperts in reliability (p.166).

Finally, we note that our results provide useful validation for research in probability assessment and for applications of decision analysis. From a research perspective, reliability is a prerequisite for validity. Thus, our findings substantiate numerous studies in probability assessment that have used general knowledge questions, such as those investigating the external correspondence of probability judgments by means of scoring rules. The results also help to validate probability assessment as it is applied in decision analysis. Decision analysts commonly use multiple response modes to elicit probabilities, reconciling differences among the procedures in consultation with the domain expert (von Winterfeldt and Edwards, 1986). We found that subjects were able to assess probabilities reliably for multiple response modes and that these different response modes demonstrated high inter-response correlational reliability. These results support the practice in decision analysis of treating different response modes as tending to be reliable and consistent.

APPENDIX

The mean probability score

For the two-alternative general knowledge task, we define a target event E as 'My chosen answer is correct'. The assessor's probability judgment for event E is labelled f, an outcome index d is defined. The index d takes on the value 1 if event E occurs (i.e. the chosen answer is, in fact, correct) and takes on the value 0 if event E does not occur (i.e. the chosen answer is not correct). The assessor's mean probability score is then computed as:

$$\bar{PS} = (1/N) \sum_{i=1}^{N} (f_i - d_i)^2$$

Over a set of such questions indexed by i, $\bar{PS}$ is a measure of overall probability judgment accuracy. It ranges between 0 (when all the chosen answers are assigned probabilities of 1 and they are correct) and 1.0 (when all the chosen answers are assigned probabilities of 1 and they are all incorrect). Lower $\bar{PS}$s are indicative of better probability judgment accuracy.

Overconfidence

To measure the assessor's over/underconfidence, the overall difference between the probability assessments and the proportion of correct responses is used.

$$\text{Overconfidence score} = \bar{f} - \bar{d}$$

where $\bar{f}$ is the mean of all probability assessments and $\bar{d}$ is the overall proportion of correct answers. A positive score indicates overconfidence and a negative score indicates underconfidence.

Slope

Slope reflects the assessor's ability to discriminate when a particular event will and will not occur. It is sensitive to the assessor's use of cues that are predictive of the target event versus those that have no predictive validity. It is computed as the difference between the mean of probability assessments for the target event on occasions when it occurs ($\bar{f}_1$), and the mean of such probability assessments on occasions when the target event does not occur ($\bar{f}_0$). That is,

$$\text{Slope score} = \bar{f}_1 - \bar{f}_0$$

Hence, higher slope scores reflect better discrimination.

REFERENCES

Beach, L. R. 'Accuracy and consistency in the revision of subjective probabilities', *IEEE Transactions on Human Factors in Electronics*, **7** (1966), 29–37.

Benson, P. G. and Önkal, D. 'The effects of feedback and training on the performance of probability forecasters', *International Journal of Forecasting*, **8** (1992), 559–73.

Branthwaite, A. 'A note comparing three measures of subjective probability', *Acta Psychologica*, **38** (1974), 337–42.

Carroll, J. B. 'Measurement properties of subjective magnitude estimates of word frequency', *Journal of Verbal Learning and Verbal Behavior*, **10** (1971), 722–9.

Carroll, J. B. and Lamendella, J. T. 'Subjective estimates of consonant phoneme frequencies', *Language and Speech*, **17** (1974), 47–59.

Ebel, R. L. 'Estimation of the reliabilities of ratings', *Psychometrika*, **16** (1951), 407–24.

Fischhoff, B. and MacGregor, D. 'Subjective confidence in forecasts', *Journal of Forecasting*, **1** (1982), 155–72.

Goodman, B. C. *Direct Estimation Procedures for Eliciting Judgments about Uncertain Events*, Engineering Psychology Technical Report 011313-5-T, University of Michigan, 1973.

Neter, J., Wasserman, W. and Kutner, M. H. *Applied Linear Statistical Models*, 3rd edn, Homewood, ILL: Irwin, 1990.

Peterson, C. R., Ulehla, Z. J., Miller, A. J., Bourne, L. E. Jr and Stilson, D. W. 'Internal consistency of subjective probabilities', *Journal of Experimental Psychology*, **70** (1965), 526–33.

Phillips, L. D. and Edwards, W. 'Conservatism in a simple probability influence task', *Journal of Experimental Psychology*, **72** (1966) 346–54.

Ronis, D. L. and Yates, J. F. '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.

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

Wallsten, T. S. and Budescu, D. V. 'Encoding subjective probabilities: a psychological and psychometric review', *Management Science*, **29** (1983), 151–73.

Wallsten, T. S. and Budescu, D. V. 'Subjective estimation of vague and precise uncertainty', in Wright, G. and Ayton, P. (eds), *Judgmental Forecasting*, New York: John Wiley, 1987.

Wallsten, T. S., Forsyth, B. H. and Budescu, D. V. 'Stability and coherence of health experts' upper and lower subjective probabilities about dose–response functions', *Organizational Behavior and Human Performance*, **31** (1983), 277–302.

Weisberg, S. *Applied Linear Regression*, New York: John Wiley, 1980.

Whitcomb, K. M., Curley, S. P., Önkal, D. and Benson, P. G. 'External correspondence of subjective probabilities with respect to assessment method and cross-national differences', Working paper, University of South Carolina, Columbia, South Carolina, June, 1992.

Wright, G. and Ayton, P. 'Subjective confidence in forecasts: a response to Fischhoff and MacGregor', *Journal of Forecasting*, **5** (1986), 117–23.

Wright, G., Saunders C. and Ayton, P. 'The consistency, coherence and calibration of holistic, decomposed and recomposed judgmental probability forecasts', *Journal of Forecasting*, **7** (1988), 185–99.

Yates, J. F. 'External correspondence: decompositions of the mean probability score', *Organizational Behavior and Human Performance*, **30** (1982), 132–56.

Yates, J. F., Zhu, Y., Ronis, D. L., Wang, D.-F., Shinotsuka, H. and Toda, M. 'Probability judgment accuracy: China, Japan, and the United States', *Organizational Behavior and Human Decision Processes*, **43** (1989), 145–71.

Authors' biographies:

Kathleen M. Whitcomb is Assistant Professor of Management Science at the University of South Carolina. She received her PhD in Management Science from the University of Minnesota (1989). Her research interests include imprecise probability assessment and evaluation, probability forecasting, and discriminant analysis.

Dilek Önkal is Assistant Professor of Decision Sciences at Bilkent University, Turkey. She received her PhD in Management Science from the University of Minnesota (1988). Her research interests include probability forecasting and reliability of subjective probabilities.

P. George Benson is Professor and Dean of the Faculty of Management at Rutgers University. He received his PhD in Decision Sciences from the University of Florida. His research interests include subjective forecasting, belief processing under uncertainty, and quality management.

Shawn P. Curley is Associate Professor of Information and Decision Sciences at the University of Minnesota. He received his PhD in psychology from the University of Michigan (1986). His research interests include behavioral decision making, subjective forecasting, and belief processing under uncertainty.

Authors' addresses:

Kathleen M. Whitcombe, College of Business Administration, University of South Carolina, Columbia, SC 29208 USA.

Dilek Önkal, College of Administrative Sciences, Bilkent University, 06533 Ankara, Turkey.

P. George Benson, Faculty of Management, Rutgers University, 81 New Street, Newark, NJ 07102, USA.

Shawn P. Curley, Carlson School of Management, University of Minnesota, 271 19th Ave., S., Minneapolis, MN 55455 USA.