Journal of Behavioral Decision Making, Vol. 8, 51–67 (1995)

Probability Judgment Accuracy for General Knowledge

Cross-national Differences and Assessment Methods

KATHLEEN M. WHITCOMB
University of South Carolina, USA

DILEK ÖNKAL
Bilkent University, Turkey

SHAWN P. CURLEY
University of Minnesota, USA

P. GEORGE BENSON
Rutgers University, USA

ABSTRACT

In this study we compare the probability judgment accuracy of subjects from the United States and Turkey. Three different response modes were employed — numerical probabilities, pie diagrams, and odds. The questions employed in the study were restricted to two-alternative, general-knowledge items. The observed pattern of differences in the components of probability judgment accuracy paralleled those of studies that have compared Western and Asian subjects. In particular, Turkish subjects exhibited better discrimination but worse calibration than their US counterparts. This result persisted across all three response modes. These findings lend support to previous assertions that observed cross-national differences arise from socioeconomic rather than Asian versus Western cultural differences. However, the consistency of the observed differences across response modes refutes a previous assertion that observed cultural differences are merely the result of response bias.

KEY WORDS Probability assessment Subjective probabilities Cross-cultural comparison

Subjective probabilities are critical components of decision processes. They have been studied extensively within the contexts of decision making under uncertainty, decision theory, decision analysis, and statistical decision theory (see Hogarth, 1980; Fishburn, 1981; Keeney, 1982; Winkler, 1982; and Abelson and Levi, 1985, for discussions of these areas). Although both qualitative and quantitative forms of likelihood judgment exist, probabilities have an advantage in that measures exist for formally evaluating their quality, or accuracy. The most extensively used measure of evaluation is known as the mean probability score, $\overline{PS}$. It indexes the overall accuracy of probability judgments and

CCC 0894–3257/95/010051–17

Received 17 April 1994
Accepted 30 August 1994

can be partitioned into component scores that facilitate the identification of various aspects of probability assessment performance (Murphy, 1973; Yates, 1982).

This study uses $\overline{PS}$ and its components to investigate and compare the quality of subjective probabilities assessed by individuals in the United States and Turkey. To facilitate comparisons with previous cross-cultural studies, we employ two-alternative general-knowledge questions such as the following (Ronis and Yates, 1987):

Voltaire was a French
(a) Writer and philosopher
(b) Physicist
The correct answer is _____.
The likelihood that your chosen answer is correct is _____.

Studies that have used such questions to examine cultural differences in probability assessment performance include Phillips and Wright (1977), Wright *et al.* (1978), Wright and Phillips (1980), Yates *et al.* (1989), and Lee *et al.* (1990). While these studies focused on differences between 'Western' and 'Eastern' cultures, the current study compares the assessments of individuals from the United States and Turkey, arguably the most Westernized of Near-Eastern countries.

Three methods of probability assessment were examined — numerical probabilities, partitioning of visual representations (i.e. pie diagrams), and odds. Consequently, we were also able to investigate whether different elicitation techniques result in differences in culture-specific tendencies that might impact various dimensions of probability judgment accuracy.

The next section presents a review of the pertinent literature on cross-cultural differences. The third section describes the methodology used in this study. Our results and a discussion of the results are in the fourth and fifth sections, respectively. A brief summary of our results and their implications for future research are presented in the sixth section. Formulas and extensive definitions for the probability judgment accuracy measures used in the present paper are contained in the Appendix.

RELEVANT CROSS-CULTURAL STUDIES

Cultural differences in probabilistic thinking were first investigated by Phillips and Wright (1977). Comparisons were made of English students attending school in Uxbridge, England, Chinese students attending school in Hong Kong, Chinese businessmen living in Hong Kong, Chinese nurses living in Hong Kong, and Chinese nurses living in London. It was found that, even though there were no differences in the number of distinct verbal expressions of uncertainty used, the English and Chinese subjects significantly differed in terms of the number of probability categories employed. The authors concluded that the Chinese were less likely than the English to ascribe to probabilistic thinking. The Chinese subjects' coarser use of the probability scale was attributed to the belief that the Chinese culture is relatively more fate-oriented than the English culture.

In a related study, Wright *et al.* (1978) compared students in Britain, Hong Kong, Malaysia, and Indonesia with respect to their probabilistic thinking. They found the British subjects to have better calibration-in-the-small, in particular to be less *overconfident*, than all the Asian subjects combined. An assessor is well calibrated (in-the-small) to the extent that his or her assessed probabilities correspond to the proportions of outcomes that occur for each of the probability judgment categories used. In this context, an assessor is said to be 'overconfident' to the extent that her probability judgments exceed her corresponding proportions correct in the judgment categories used. British subjects also used finer numerical and verbal expressions of uncertainty in comparison to the Asian subjects. The influence of fate-orientation on probabilistic thinking was again stressed. It was con-

cluded that (1) differences in viewing and using subjective probabilities may have implications for the communication of uncertainty between the countries (at the governmental, institutional, and individual levels), and that (2) use of decision analysis techniques in such (fate-oriented) cultures may be problematic.

Similar results were obtained in an extended study where Wright and Phillips (1980) compared Malay, Chinese, and Indian Malaysians, Moslem and Christian Indonesians, British civil servants, Hong Kong managers, and Indonesian managers. Again, British subjects were found to be better calibrated, and to use more expressions (both numerical and verbal) to express their uncertainty.

In contrast to studies focusing mainly on the calibration of British and Asian subjects, Yates *et al.* (1989) compared various probability assessment skills of Chinese, American, and Japanese subjects. While the three groups displayed essentially equivalent mean probability scores, significant differences in the components of the mean probability score were revealed. In particular, it was found that American subjects exhibited better overall calibration in their probability responses to general-knowledge questions than the Chinese subjects. Overall calibration is also referred to as calibration-in-the-large, and is often expressed as bias. In terms of overall calibration, a subject is considered to be *overconfident* (positively biased) to the extent that the mean of her probability judgments exceeds her overall proportion of correct answers. The Americans also showed less excess variation (scatter) in their probability responses to general-knowledge questions than the Chinese. However, the Chinese subjects were able to attain higher (better) slope and resolution scores than the American subjects, indicating a better ability to discriminate between occasions when they were correct and incorrect.

Comparisons of Chinese and Japanese subjects also demonstrated the superior discrimination ability of the Chinese subjects as measured by slope. In contrast, the Japanese subjects' probability judgments exhibited better (lower) overconfidence with less scatter, thus resembling those of the American subjects. The authors speculated that the results were reflective of fundamental cultural differences in how individuals think about uncertainty. It was suggested that the superior discrimination achieved by the Chinese subjects could be viewed as a product of the social incentive structure of the Chinese society. Similarly, the superior calibration of Western subjects may be a product of a social incentive structure that places more emphasis on correct numerical labeling, i.e. calibration.

Focusing only on calibration measures, Lee *et al.* (1990) compared students in the USA, Japan, Taiwan, Singapore, and India. Subjects from India and Taiwan were found to display high positive bias (around 13%), when compared to Singapore and US subjects (around 7%). Japanese subjects were the least overconfident (around 4%). The authors concluded that overconfidence seems to be a pervasive phenomenon, and it can be especially prominent in some Asian cultures.

In summary, previous research suggests:

(1) Asian subjects used the probability scale more coarsely than British subjects, possibly due to a fate orientation in the Asian culture.
(2) Asian subjects exhibited greater miscalibration, in terms of calibration-in-the-small, than American and Japanese subjects.
(3) Asian subjects exhibited greater positive bias than American subjects.
(4) Chinese subjects had better discrimination than American and Japanese subjects.
(5) Japanese subjects' responses more closely resembled American responses than Chinese responses.

A general limitation of much of the prior research is the tendency to focus only on one of the characteristics of subjects' responses, e.g. calibration or overconfidence. This difficulty is highlighted by Yates *et al.* (1989), who noted that studies contrasting the probabilistic judgments of Western

and Eastern subjects had not been evaluated in terms of overall accuracy or component scores other than calibration measures.

Our research extends the previous studies in several important ways. First, we broaden the investigation by considering a variety of differences using multiple measures. Second, we employ several different elicitation techniques. This corresponds with the recommended practice in decision analysis of using multiple encoding methods, since subjects' responses have been shown to vary across elicitation techniques (von Winterfeldt and Edwards, 1986). Sensitivity to elicitation techniques across cultures has not been investigated in terms of probability judgment performance measures. Third, we compare subjects from the United States and Turkey, the most Westernized Near-Eastern country (Cindoglu, 1991). This allows us to test the generalizability of the American–Japanese correspondence observed by Yates *et al.* (1989).

METHODOLOGY

The experiment was designed to evaluate and compare probability assessments made by US and Turkish subjects using three different methods of probability assessment — numerical probabilities, pie diagrams, and odds. Assessments made using pie diagrams were converted to probabilities by measuring the designated angle, x, in degrees and dividing by 360. Odds ratios, x:1, were converted to numerical probabilities by dividing x by $(x + 1)$. The mean probability score and various component scores were used to measure the quality of the subjects' adjudged probabilities. The details of the experiment are delineated in the following subsections.

Subjects

Sixty-two subjects participated in the study. Thirty-two were upperclassmen and students in various master's degree programs in the College of Business Administration at the University of South Carolina. Thirty of the subjects were upperclassmen and master's degree students at the College of Business Administration at Bilkent University, Ankara, Turkey. Within each university, an equal number of undergraduate and master's students were used.

Procedure

Subjects were individually scheduled for three probability assessment sessions. The sessions were arranged so that they were separated by at least one week. During each of the assessment sessions, subjects were required to respond to a hundred general-knowledge questions by choosing one of two possible alternatives and then assessing the likelihood that their chosen answer was correct. In each session, subjects used two methods for assessing these likelihoods (fifty questions each). Thus, over three sessions, each subject used each of three methods of probability assessment twice.

Prior to the first assessment session, subjects received general training in probability assessment and the three assessment methods employed in the study. They also received information about the measures used to evaluate the accuracy of probability assessments. Specifically, subjects assessed probabilities for a practice set of twenty questions and their mean probability, slope, and overconfidence scores for these questions were immediately calculated. Subjects were then told how each of the three scores was computed, the best and worst possible scores for each measure, and their own scores. Similar training sessions using ten practice questions were conducted prior to the second and third assessment sessions. Training with scoring rules was provided to maintain subjects' motivation levels throughout the experiment and to encourage them to report their true opinions, that

is, to discourage bluffing or hedging. Monetary incentives based on probability judgment accuracy scores could not be used to motivate subjects since questions were repeated in the different assessment sessions. Such an incentive scheme could have encouraged researching of answers. All subjects were paid a fixed amount for their participation.

Probability assessment methods

Numerical probabilities and likelihood ratios are the most commonly used and studied non-verbal forms of uncertainty expression. Numerical probabilities tend to be preferred by those with more technical backgrounds, whereas there is some evidence that likelihood ratio methods are favored by those less quantitatively sophisticated (von Winterfeldt and Edwards, 1986). In this study, we used numerical probabilities, partitioning of visual representations (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 designating 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 closer their assessed probability was to 1.0, the more strongly they believed their answer to be correct.

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 designating an angle of 360° indicated that they were certain that their chosen answer was correct. They were advised that the closer their angle to 360°, the more certain they were that their chosen answer was correct.

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 that their answer was any more likely to be correct than the alternative not chosen, and that assessing odds where the first number was very much larger than one indicated that they were very sure that their answer was correct. They were told that the larger the ratio, the more certain they were that their chosen alternative was correct.

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 that they chose was ten times more likely to be correct than the answer not chosen, 100:1 meant that their answer was a hundred times as likely to be correct than the answer not chosen, 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 and analysis

The design was constructed to account for the effects of culture and probability assessment methodology on the accuracy of probability judgments. The orders in which probability assessment methods were assigned to the subjects in each of the three sessions were randomized across subjects and sessions in order to avert a potential bias due to the effects of practice. For the analysis of the various accuracy measures, the design was a 3 (assessment method) × 2 (culture) design. The first factor was manipulated within-subjects, the second was between-subjects.

One hundred general-knowledge questions were used in each session, fifty questions for each of two assessment methods. The same set of one hundred questions, presented in different random orders, was used in each of the three assessment sessions. In computing calibration-in-the-small and resolution, probabilities were first categorized into one of five probability intervals: [0.5, 0.6) [0.6, 0.7), [0.7, 0.8), [0.8, 0.9], and [0.9, 1.0]. The mean probability response for each of the k categories, $\bar{f}_k$, and the proportion of correct answers in each of the k categories, $\bar{d}_k$, were used to calculate calibration-in-the-small and resolution.

Separate ANOVAs were conducted for each of the eight accuracy measures — the mean probability score, proportion correct, overconfidence, calibration, resolution, ANDI, slope, and scatter. In each case the response vector consisted of 186 responses (62 subjects times three assessment sessions). Since there was an unequal number of assessors nested in each culture (32 US versus 30 Turkish subjects), the analyses were conducted using the regression approach. The F-tests were based on the appropriate full and reduced regression models.[1] In cases where the F-test for the effects of probability assessment method was significant, Tukey's pairwise comparison procedure was used to determine which means differed.

RESULTS

Exhibit 1 contains the mean values of the accuracy measures based on the individual scores for US and Turkish subjects, averaged across subjects and probability assessment methods. The associated probability values for the F-tests are also reported. Similarly, Exhibit 2 contains the means of the eight scores for the three probability assessment methods averaged across subjects and culture along with the corresponding probability values of the F-tests. No significant two-way interactions between method and culture were detected for any of the eight measures.

Aggregate calibration diagrams and covariance graphs for the US and Turkish subjects are presented in Exhibits 3 and 4. Calibration diagrams are used to illustrate both calibration and resolution. Covariance graphs are often used to depict the measures of overconfidence, slope, and scatter. Unlike Exhibits 1 and 2, which summarize probability judgment accuracy measures for individual assessors, the calibration diagrams and covariance graphs of Exhibits 3 and 4 were constructed by pooling probability assessments across subjects and assessment methods within each group. Some of the component scores reported in Exhibits 3 and 4 will not agree with those reported in Exhibit 1 due to different methods of calculation (pooling versus averaging of subject scores). However, the patterns of differences for the pooled data in Exhibits 3 and 4 are consistent with, and illustrative of, the patterns found in Exhibit 1.

The calibration diagrams for US and Turkish subjects are shown in Exhibits 3(a) and 3(b), respectively, while the covariance graphs for the US and Turkish subjects are depicted in Exhibits 4(a) and 4(b), respectively. Since there were few pronounced differences in the probability accuracy mea-

[1] Although F-tests are robust against departures of the dependent variable from normality, we also conducted non-parametric, one-way ANOVAs using the Kruskal–Wallis procedure. Overall, the results of the Kruskal–Wallis tests were consistent with those for the F-tests. Two cases where they were inconsistent are reported in the following section.

Exhibit 1. Mean accuracy measures — US and Turkish subjects

Component/measure[a]		United States	Turkish	*p*-value
Overall				
$\bar{d}$ (proportion correct)	↑	0.644	0.633	0.082
$\overline{PS}$	↓	0.218	0.226	0.065
Calibration				
Cal-in-small	↓	0.013	0.020	0.000
Bias	0	0.009	0.082	0.000
Discrimination				
Resolution	↑	0.023	0.026	0.039
ANDI	↑	0.060	0.077	0.049
Slope	↑	0.079	0.097	0.000
Noise				
Scatter	↓	0.022	0.028	0.000

[a] Here and in Exhibit 2 the symbol ↑ indicates that higher scores are better, ↓ that lower scores are better. 0 indicates that the best score is zero.

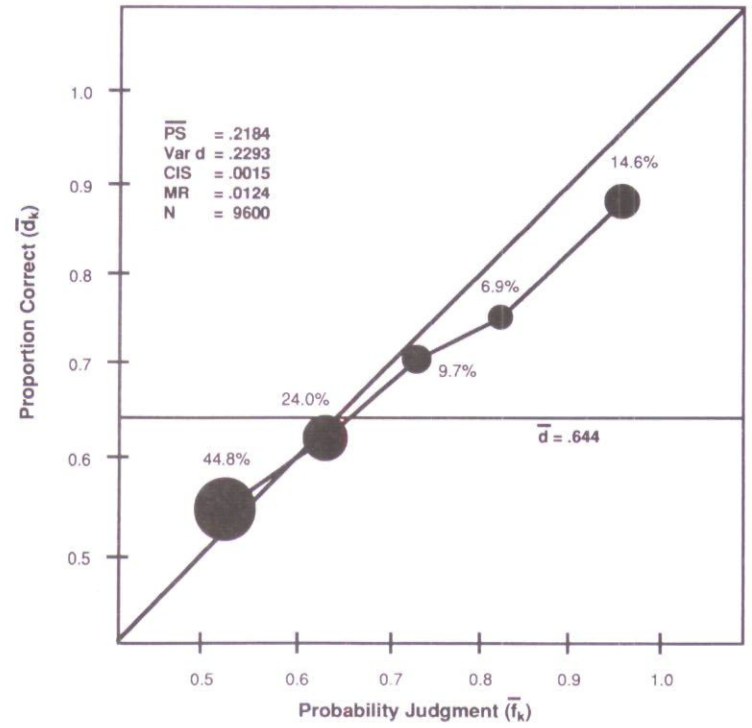
Exhibit 2. Mean accuracy measures — numerical probabilities, pie diagram, and odds methods

Component/measure		Numbers	Pie diagram	Odds	*p*-value
Overall					
$\bar{d}$ (proportion correct)	↑	0.636	0.646	0.634	0.371
$\overline{PS}$	↓	0.221	0.223	0.223	0.280
Calibration					
Cal-in-small	↓	0.017	0.019	0.016	0.231
Bias	0	0.042	0.038	0.051	0.148
Discrimination					
Resolution	↑	0.028	0.024	0.027	0.020
ANDI	↑	0.084	0.064	0.079	0.024
Slope	↑	0.094	0.090	0.080	0.000
Noise					
Scatter	↓	0.026	0.029	0.022	0.000

sures between the three assessment methods, their calibration diagrams and covariance graphs were not very informative and thus are not presented.

Overall accuracy

Exhibit 1 reveals that the proportion of questions answered correctly by US subjects was slightly greater than for Turkish subjects. Using the *F*-test, this difference was somewhat significant (*p*-value = 0.082). However, the Kruskal–Wallis test found no significance difference (*p*-value = 0.296). Since there was no obvious departure from normality in the error terms for the parametric ANOVA, it is likely that this discrepancy was due to the greater power of the *F*-test versus that of the Kruskal–Wallis test. In particular, the multi-factor ANOVA accounted for substantial within-subjects variation that the one-way ANOVA could not. A difference in the proportions correct between the two groups is of interest because it is indicative of question difficulty, which has been found


PS = .2184
Var d = .2293
CIS = .0015
MR = .0124
N = 9600
14.6%
6.9%
9.7%
24.0%
44.8%
d = .644
Proportion Correct (dk)
Probability Judgment (fk)
1.0
0.9
0.8
0.7
0.6
0.5

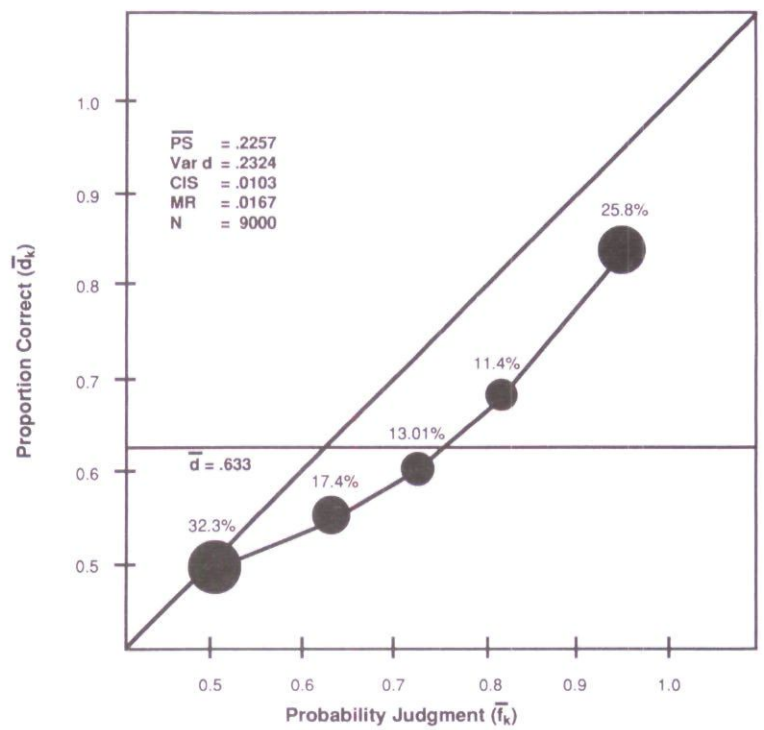
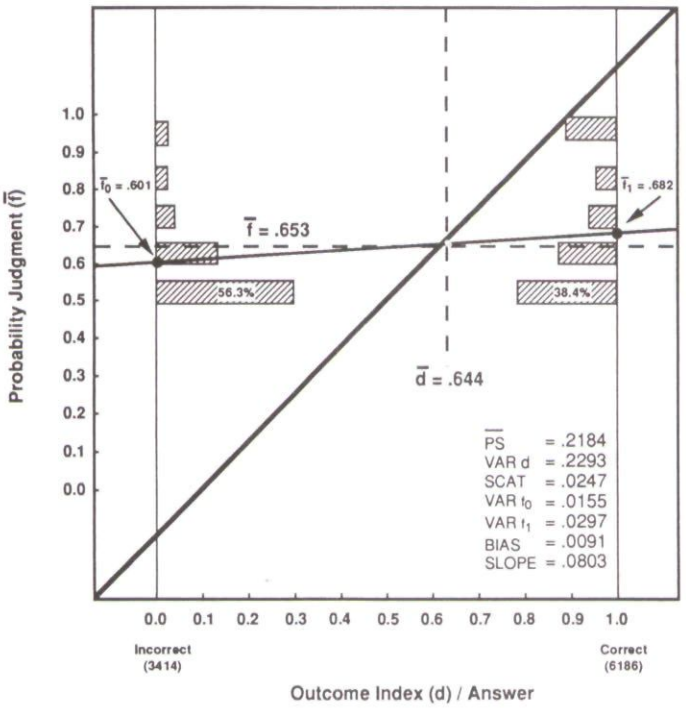

PS = .2257
Var d = .2324
CIS = .0103
MR = .0167
N = 9000
25.8%
11.4%
13.01%
17.4%
32.3%
d = .633
Proportion Correct (dk)
Probability Judgment (fk)
1.0
0.9
0.8
0.7
0.6
0.5

Exhibit 3.


Probability Judgment (f̄)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
f̄₀ = .601
f̄₁ = .682
f̄ = .653
56.3%
38.4%
d̄ = .644
PS = .2184
VAR d = .2293
SCAT = .0247
VAR f₀ = .0155
VAR f₁ = .0297
BIAS = .0091
SLOPE = .0803
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Incorrect (3414)
Correct (6186)
Outcome Index (d) / Answer

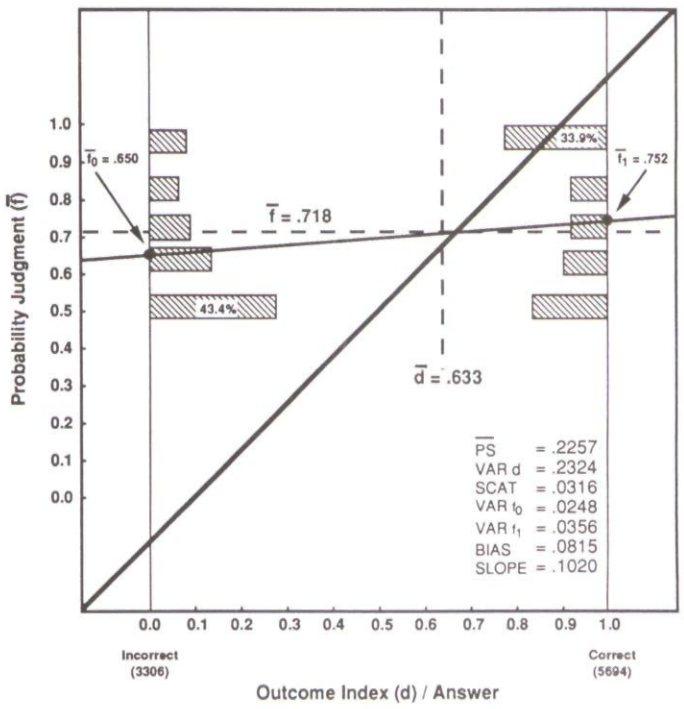

Probability Judgment (f̄)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
f̄₀ = .650
f̄₁ = .752
f̄ = .718
33.9%
43.4%
d̄ = .633
PS = .2257
VAR d = .2324
SCAT = .0316
VAR f₀ = .0248
VAR f₁ = .0356
BIAS = .0815
SLOPE = .1020
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Incorrect (3306)
Correct (5694)
Outcome Index (d) / Answer

Exhibit 4.

to affect subjects' over/underconfidence. In particular, Lichtenstein and Fischhoff (1980) found evidence indicating that subjects tend to be overconfident for hard items and underconfident for easy items. With respect to the present study, the difference detected by the *F*-test in the proportions of questions answered correctly is slight and probably more reflective of the power of the test than of any practical difference. As would be expected, no significant differences were found in the proportions correct among the three methods of probability assessment.

The mean $\overline{PS}$ for US subjects was slightly lower than for Turkish subjects, 0.2184 versus 0.2256, respectively. There was no significant difference among the three methods and only a marginally significant difference between cultures. These scores are indicative of at least a modest amount of probability assessment skill on the part of subjects. They are better than 0.25, which is the expected score for a constant probability assessor with no knowledge who always assigns a probability of 0.50 to the chosen alternative. These scores are also comparable to those obtained from studies using similar subject populations and similar stimuli. For example, Yates *et al.* (1989) reported mean $\overline{PS}$s of 0.2258, 0.2204, and 0.2255 for Chinese, US, and Japanese students, respectively, making probability judgments for general-knowledge questions.

Calibration measures

US subjects were better calibrated-in-the-small than Turkish subjects (0.0132 versus 0.0204). They also exhibited less overconfidence (0.0090 versus 0.0820). These differences are clearly visible on the calibration diagrams of Exhibit 3. The diagrams plot the proportion correct in each category, $\bar{d}_k$, against the corresponding mean probability response, $\bar{f}_k$. The area of each point $(\bar{f}_k,\bar{d}_k)$ is in direct relation to the percentage of probability responses made for that judgment category. The number placed adjacent to each point refers to this percentage. The calibration diagrams reveal that both groups were overconfident. The calibration curves generally fall below the identity line, indicating that the proportion correct was generally less than the mean probability assessed within each of the k probability categories. However, the calibration curve derived from Turkish subjects' assessments is everywhere below the curve for US subjects, indicating that these subjects exhibited more overconfidence at every level of probability usage.

The difference in overconfidence between US and Turkish subjects is also evident on the covariance graphs in Exhibit 4. Overconfidence is indicated in a covariance graph by the distance that the intersection of the dashed lines for $\bar{f}$ and $\bar{d}$ falls above the identity line. This distance is undetectable for the US subjects (Exhibit 4(a)), whereas it is quite marked for the Turkish ones (Exhibit 4(b)). Turkish subjects exhibited noticeable overconfidence in their adjudged probabilities, whereas US subjects did not, regardless of which assessment method was employed. No significant differences in calibration or overconfidence were detected among the three methods of probability assessment.

Discrimination measures

Exhibit 1 indicates that Turkish subjects' mean resolution score was moderately better (higher) than the mean score of US subjects. This continued to be true when the resolution score was transformed to the adjusted, normalized discrimination index (ANDI). Resolution is depicted in a calibration diagram by the vertical distances of the points $(\bar{f}_k,\bar{d}_k)$ from the horizontal line $\bar{d}$ weighted by the proportion of points in each of the k judgment categories. Typically, resolution is good to the extent that the points are far away from the overall proportion correct, $\bar{d}$. Using only these vertical distances as a gauge, Exhibit 3 does not reflect the better resolution of Turkish subjects relative to US subjects. Rather, the higher (better) resolution scores achieved by Turkish subjects are a consequence of the greater percentage of probability judgments assessed in probability categories for which the vertical distance between $(\bar{f}_k,\bar{d}_k)$ and $\bar{d}$ are greatest. In particular, Turkish subjects made far more probability

judgments in the $0.9 \leq 1.0$ category compared to US subjects. Since the $0.9 \leq 1.0$ probability category happened to be the farthest from $\bar{d}$ for both groups of subjects, Turkish subjects achieved better resolution.

The difference in slope scores between Turkish and US subjects was more pronounced. Specifically, Turkish subjects had noticeably better (higher) slope scores than US subjects. Slope is quite literally depicted in the covariance graph as the slope of the line connecting the points $(0,\bar{f}_0)$ and $(1,\bar{f}_1)$. Exhibit 4 demonstrates the steeper (better) slope for Turkish assessors compared to US assessors. Thus, the Turkish assessors were better able to differentiate, on average, between instances when they were correct and instances when they were incorrect.

Exhibit 2 shows that there is a significant difference in the mean resolution and ANDI with respect to probability assessment method. Tukey's pairwise comparison procedure revealed that the mean resolution and ANDI associated with the numerical probabilities method was significantly better than the pie diagram method. The slope scores also differed significantly with respect to the *F*-tests (p-value ≈ 0). Tukey's procedure indicated that the numerical probabilities method was significantly better (higher) in terms of slope than the odds method. However, the Kruskal–Wallis test detected no significant differences in slope among the three methods (p-value $= 0.274$). Unlike the previously discussed disparity between the *F*- and Kruskal–Wallis tests, this discrepancy is in part due to non-normality of the error terms. In this case, the Kruskal–Wallis test may be the more appropriate indicator of statistical significance. Consequently, we would conclude that there are no significant differences in slope with respect to assessment method.

Noise measure

US subjects had significantly less excess variation or 'noise' associated with their probability judgments than Turkish subjects, as indicated by their better (lower) scatter scores. Excess variation in probability judgments is generally attributed to one or two sources: (1) the innate inconsistency of the probability assessor, and (2) the assessor's reliance on cues that are not especially indicative of the target event (Yates, 1990). Scatter is denoted in Exhibit 4 by the variability in the distributions of probability assessments for cases where the target event did not occur and for where the target event did occur. These distributions are illustrated in the covariance graphs by the relative frequency histograms. The percentage listed adjacent to the longest bar within each distribution refers to the percentage of occasions for which subjects assessed probabilities in a particular judgment category given that they were either correct or incorrect in their choice of answers. Comparison of the histograms illustrates that the scatter scores are better (lower) for US subjects than for Turkish subjects. This result is consistent with the results for slope and a tendency for a slope/scatter 'trade-off' (Yates and Curley, 1985).

Even though the conditional probability judgments for the Turkish subjects were more uniformly distributed than for US subjects, as indicated in Exhibit 4, Turkish subjects still used fewer probability categories, i.e. made coarser probability distinctions, than the US subjects. This finding was consistent for all three methods of probability assessment and also coincides with results reported in Phillips and Wright (1977), Wright *et al.* (1978), and Wright and Phillips (1980). Also, as indicated in Exhibit 2, both groups of subjects did better (had lower scatter scores) using the odds method of probability assessment than with either the numerical probabilities or pie diagram methods.

DISCUSSION

In terms of overall accuracy (i.e. the proportion correct and mean probability score), US subjects achieved better scores compared to their Turkish counterparts. However, the magnitudes of the

differences were quite small and comparable to differences observed in previous cross-cultural studies where it was concluded that the differences in overall accuracy were virtually imperceptible (Yates *et al.*, 1989). Interestingly, however, the US and Turkish subjects achieved this comparable overall accuracy in quite different ways, as revealed by the multiple components of assessment performance.

Analogous to prior research, the US subjects in this study were better calibrated than the Turkish subjects. This result applied to both calibration-in-the-small and bias. In their discrimination, Turkish assessors paralleled the performance of Chinese assessors by exhibiting better discrimination scores than did the American assessors. While this difference was most discernible for the slope, the differences in the resolution and ANDI scores were consistent and statistically significant.

Overall, these results bear upon explanations for the differences in the quality of probability assessments across cultures. Wright and Phillips (1980) suggested that individuals from Western cultures are taught to think in terms of degrees of certainty, whereas people in Eastern cultures think in terms of absolutes. However, a study conducted by Yates *et al.* (1989) found no discernible differences between the overall accuracy or the component scores for American and Japanese subjects. Their results indicate that a simple (Far) Eastern versus Western cultural explanation does not provide a complete accounting for dissimilarities in probability judgment accuracy. The current study identified differences between Near Eastern and Western cultures that were similar to the Eastern/Western differences described by Wright and Phillips (1980). The result is consistent with a suggestion by Yates *et al.* (1989) that the observed differences in the quality of probability assessments might be more related to socioeconomic than cultural factors. Specifically, they suggest that technologically oriented societies, such as Japan, the United States, and Britain, emphasize the ability to assign proper numerical labels (calibration) rather than discrimination. In contrast, less technologically oriented societies, such as China and Turkey, have incentive structures that place more significance on 'knowing' or discrimination. Since Turkish culture is not strictly Asian, this explanation seems more reasonable than that of cultural differences.

One difference between the current study and previous ones comparing Western and Eastern probability assessors is in the magnitudes of the component scores. For example, our US subjects exhibited almost no overconfidence on average, whereas the average overconfidence level for Turkish subjects was similar to that previously reported for Western subjects. This decrease in overconfidence relative to previously reported values plausibly is attributable to two features of the current study. First, our study included practice sessions in which subjects received training in probability assessment, the overconfidence measure, mean probability, and slope scores. Past research has found that training with calibration measures is effective in reducing overconfidence and improving calibration, whereas training with discrimination measures has mixed results (Lichtenstein and Fischhoff, 1977; Sharp *et al.*, 1988; Benson and Önkal, 1992). The current study is consistent with those results. Our subjects exhibited better calibration, but not necessarily better discrimination (particularly as measured by the slope). Second, our study included more assessments. Lower calibration and resolution scores have been conjectured to result from an increase in the number of probability judgments from which these component scores are computed.[2]

Another important difference between the current and past studies is our use of multiple assessment techniques. Their inclusion leads to a contradiction of the hypothesis that the inferred cultural differences have simply been the result of a response bias. According to this hypothesis, no underlying differences exist between cultures. Instead, the observed differences reflect disparities in the suitability of a particular method of assessment among diverse groups of individuals. If this were so, the differ-

---

[2] In a reported communication between Sarah Lichtenstein and J. Frank Yates, Lichtenstein noted that she had observed a systematic decrease in both calibration (in-the-small) and resolution as the number of probability assessments on which they were based increased (Yates *et al.*, 1989).

ences should vary with changes in the response mode. However, we found no such variation. The same pattern of differences in the component scores persisted across all three assessment methods.

It is important to recognize that the probability assessment tasks employed in the current study were restricted to general-knowledge questions. While this constraint facilitates the comparison of the results of the present study with those from previous cross-cultural studies, it also raises the issue of generalizability to probability judgments for future events. This issue has been debated in numerous studies. In particular, Yates (1982) and Ronis and Yates (1987) discuss conceptual difficulties that arise when discrimination measures are used to evaluate probability judgments for general-knowledge items. These difficulties are most apparent by considering an assessor who demonstrates perfect discrimination in his probability judgments for two-alternative general knowledge questions. While one could imagine how a probability forecaster might achieve perfect discrimination without perfect accuracy, it is difficult to conceptualize how an assessor could discriminate perfectly between occasions when she is correct versus incorrect without also having perfect knowledge of the correct answer.

Other studies questioning generalizability have focused primarily on the 'overconfidence phenomenon' (Wright and Ayton, 1986). The overconfidence phenomenon refers to the observation that probability judgments for general-knowledge questions tend to display more positive bias and calibration curves that fall farther below the identity line than do probability judgments for future events. Recently, researchers have presented evidence indicating that the overconfidence phenomenon may be a consequence of biased selection of general-knowledge questions on the part of experimenters, rather than any cognitive bias on the part of the assessors (Gigerenzer *et al.*, 1991; Juslin, 1994). Clearly, future studies investigating national differences in probability judgment accuracy should include probability judgment tasks for future events in order to test the generalizability of results derived using general-knowledge questions.

CONCLUSION

This study usefully extends our knowledge of cross-national differences in the quality of probability judgment accuracy. Using multiple accuracy measures, we found that the differences between Turkish and American subjects parallel those previously observed between Eastern and Western subjects. Using multiple response modes, we found that the differences generalize across probability assessment methods. The former results lends support to the assertion by Yates *et al.* (1989) that the observed differences arise from socioeconomic rather than Eastern versus Western cultural differences. The latter result contradicts the hypothesis that the observed cross-national differences derive from a response bias.

Assuming that these findings translate to probability judgments for externally determined future events, they have serious implications for cross-national applications of decision analysis. Indeed, the potential for miscommunicating uncertainty is almost assured in certain instances. This suggests the need for elicitation aids for bridging the cultural gap.

A recent study by Lee *et al.* (1990) provided evidence that cross-national differences in probability judgment accuracy among individuals from different socioeconomic backgrounds were linked to differences in cognitive rather than motivational processes. If true, differences in the expression of uncertainty among different cultures may only be resolved by focusing attention on the cognitive processes from which subjective probabilities are derived. Accordingly, we would suggest that future cross-national studies investigate the cognitive, belief processing differences (Smith *et al.*, 1991; Benson *et al.*, in press) that are manifested in the observed accuracy differences.

APPENDIX

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. The outcome index is labelled d; it takes on the value 1 if event E occurs (i.e. the chosen answer is, in fact, the correct one) 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:

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

Over a set of N such questions indexed by i, $\overline{PS}$ 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 all correct) and 1.0 (when all the chosen answers are assigned probabilities of 1 and they are all incorrect). Lower $\overline{PS}$ is indicative of better probability judgment accuracy.

Partitioning $\overline{PS}$ had yielded components that address different aspects of probability judgment performance (for details of these decompositions, see Sanders, 1963; Murphy, 1973; Yates, 1982). The performance measures utilized in this study represent the most widely adopted evaluation criteria emanating from these decompositions (Yates *et al.*, 1991). The specific components are briefly described below.

Proportion correct

An assessor's categorical judgment accuracy is reflected in the proportion of questions answered correctly by an assessor. The *proportion correct*, also called $\bar{d}$, is an indicator of the assessor's general knowledge. In addition, when $\bar{d}$ is computed over a set of assessors, it is viewed as an indication of task difficulty (Fischhoff and MacGregor, 1982; Lichtenstein and Fischhoff, 1980; Wright, 1982). Aside from its inherent interest, task difficulty affects the scores attained on other components, and hence, should be taken into consideration in the analysis of results.

Calibration-in-the-small

An assessor is perfectly *calibrated-in-the-small* if for all chosen outcomes assigned a given probability, the proportion of those outcomes that occur (i.e. the proportion of those answers that are correct) equals the probability assigned. That is, an assessor is perfectly calibrated if for all the questions for which she assessed a 0.8 probability, 80% were actually correct; for all the questions given a 0.6 probability, 60% were actually correct, etc. The following measure of calibration-in-the-small requires that the responses be divided into K categories.

$$\text{Calibration score} = (1/N) \sum_{k} N_k (\bar{f}_k - \bar{d}_k)^2$$

where $\bar{f}_k$ is the mean probability response in category k; $\bar{d}_k$ is the proportion of correct answers in category k; N_k is the number of responses in category k, and N is the total number of responses. Lower scores are reflective of better calibration-in-the-small. (Note that the present calibration score differs from more traditional indexes of calibration in which $\bar{f},k$ is replaced by a predetermined, fixed value f_k) Calibration-in-the-small is often illustrated by a 'calibration diagram'. In this diagram the proportion correct is plotted against the corresponding mean probability assessment for each

of the k assessment categories. An assessor is perfectly calibrated if each of the plotted points falls on the identity line. For judgment tasks using general-knowledge questions, over/underconfidence is evidenced by the extent to which the plotted points fall below/above the identity line.

Bias

Since the calibration score is a squared measure, it does not indicate if the probability assessment ($\bar{f}_k$) systematically exceed (or are systematically exceeded by) the proportion of correct answers ($\bar{d}_k$). To measure the assessor's *bias*, or *over/underconfidence*, the overall difference between the probability assessments and the proportion of correct responses is used:

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

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 underconfidence.

Resolution

Resolution reflects the assessor's ability to discriminate in the use of different response categories. It is computed as the weighted average of the squared differences between the proportion correct in each assessment category and the overall proportion correct. That is,

$$\text{Resolution score} = (1/N) \sum_k N_k (\bar{d}_k - \bar{d})^2$$

Larger scores indicate better resolution.

ANDI

Resolution is recognized to have certain limitations (Sharp *et al.*, 1988; Yaniv *et al.*, 1991). Specifically, the measure is influenced by the variability in outcomes, var(d), the number of judgment categories, k, and the total number of judgments, N. Yaniv *et al.* (1991) proposed an alternative, the 'adjusted, normalized discrimination index', *ANDI*. ANDI is expressed as follows:

$$(k - 1)/N \times \text{resolution score} \times 1/\text{var}(d)$$

Slope

Slope reflects the extent to which an assessor assigns larger probabilities to the target event on occasions when the target event occurs than on occasions when the target event does not occur. As such, slope reflects the assessor's ability to discriminate when the target 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 ($\bar{f}_0$). That is,

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

Hence, higher scores reflect better discrimination.

Scatter

Scatter reflects probability judgment variability that is independent of the occurrence or nonoccurrence of the target event. Accordingly, scatter reveals the overall 'noise' or excess variability in the assessor's probabilities. It is computed as follows:

$$\text{Scatter scores} = (1/N)\,(N_1\,\text{var}(f_1) + N_0\,\text{var}\,(f_0))$$

where $\text{var}(f_1)$ is the variance of probability assessments for the target event on the N_1 occasions when it occurs; $\text{var}(f_0)$ is the variance of such assessments on the N_0 occasions when the target event does not occur; and $N = N_1 + N_0$. Lower scatter scores reflect less 'noisiness' in the assessments, and hence, are more desirable.

REFERENCES

Abelson, R. P. and Levi, A. 'Decision making and decision theory', in Lindzey, G. and Aronson, E. (eds), *Handbook of Social Psychology*, 3rd edn, Vol.1 (pp. 231–309), New York: Random House, 1985.

Benson, P. G., Curley, S. P. and Smith, G. F. 'Belief assessment: An underdeveloped phase of probability elicitation', *Management Science*, in press.

Benson, P. G. and Önkal, D. 'The effects of feedback and training on the performance of probability forecasting', *International Journal of Forecasting*, **8** (1992), 559–74.

Cindoglu, D. *Re-viewing Women: Images of Patriarchy and Power in Modern Turkish Film*, unpublished PhD dissertation, SUNY at Buffalo: New York, 1991.

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

Fishburn, P. C. 'Subjective expected utility: A review of normative theories', *Theory and Decision*, **13** (1981), 139–99.

Gigerenzer, G., Hoffrage, U. and Kleinbölting, H. 'Probabilistic mental models: A Brunswikian theory of confidence', *Psychological Review*, **98** (1991), 506–28.

Hogarth, R. M. *Judgment and Choice: The Psychology of Decision*, New York: John Wiley, 1980.

Juslin, P. 'The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items', *Organizational Behavior and Human Decision Processes*, **57** (1994), 226–46.

Keeney, R. L. 'Decision analysis: An overview', *Operations Research*, **30** (1982), 803–38.

Lee, J.-W., Yates, J. F., Shinotsuka, H., Yen, N.-S., Singh, R., Onglatco, M. L. U., Gupta, M. and Bhatnagar, D. 'Generality of cross-national differences in probability judgment calibration', Paper presented at the Joint Meeting of ORSA/TIMS, Philadelphia, 1990.

Lichtenstein, S. and Fischhoff, B. 'Do those who know more also know more about how much they know? The calibration of probability judgments', *Organizational Behavior and Human Performance*, **20** (1977), 159–83.

Lichtenstein, S. and Fischhoff, B. 'Training for calibration', *Organizational Behavior and Human Performance*, **26** (1980), 149–71.

Murphy, A. H. 'A new vector partition of the probability score', *Journal of Applied Meteorology*, **12** (1973), 595–600.

Phillips, L. D. and Wright, G. N. 'Cultural differences in viewing uncertainty and assessing probabilities', in Jungermann, H. and de Zeeuw, G. (eds), *Decision Making and Change in Human Affairs*, (pp. 507–19), Amsterdam: D. Reidel.

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.

Sanders, F. 'On subjective probability forecasting', *Journal of Applied Meteorology*, **2** (1963), 191–201.

Sharp, G. L., Cutler, B. L. and Penrod, S. D. 'Performance feedback improves the resolution of confidence judgments', *Organizational Behavior and Human Decision Processes*, **42** (1988), 271–83.

Smith, G. F., Benson, P. G. and Curley, S. P. 'Belief, knowledge, and uncertainty: A cognitive perspective on subjective probability', *Organizational Behavior and Human Decision Processes*, **48** (1991), 291–321.

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

Winkler, R. L. 'Research directions in decision making under uncertainty', *Decision Sciences*, **13** (1982), 517–33.

Wright, G. 'Changes in the realism and distribution of probability assessments as a function of question type', *Acta Psychologica*, **52** (1982), 165–74.

Wright, G. N., Phillips, L. D., Whalley, P. C., Choo, G. T., Ng, K. O., Tan, I. and Wisudha, A. 'Cultural differences in probabilistic thinking', *Journal of Cross-Cultural Psychology*, **9** (1978), 285–99.

Wright, G. N. and Phillips, L. D. 'Cultural variation in probabilistic thinking: Alternative ways of dealing with uncertainty', *International Journal of Psychology*, **15** (1980), 239–57.

Wright, G. N. and Ayton, P. 'Subjective confidence in forecasts: a response to Fischhoff and MacGregor', *Journal of Forecasting*, **5** (1982), 117–23.

Yaniv, I., Yates, J. F. and Smith, J. E. K. 'Measures of discrimination skill in probabilistic judgment', *Psychological Bulletin*, **110** (1991), 611–17.
Yates, J. F. 'External correspondence: Decompositions of the mean probability score', *Organizational Behavior and Human Performance*, **30** (1982), 132–56.
Yates, J. F. *Judgement and Decision Making*, Englewood Cliffs, NJ: Prentice Hall, 1990.
Yates, J. F. and Curley, S. P. 'Conditional distribution analyses of probabilistic forecasts', *Journal of Forecasting*, **4** (1986), 61–73.
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.
Yates, J. F., McDaniel, L. S. and Brown, E. S. 'Probabilistic forecasts of stock prices and earnings: The hazards of nascent expertise', *Organizational Behavior and Human Decision Processes*, **49** (1991), 60–79.

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 probability forecasting, imprecise probability assessment and evaluation, and Bayesian statistical 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 are in decision analysis and probability forecasting.

Shawn P. Curley, PhD, University of Michigan, is Associate Professor of Information and Decision Sciences at the University of Minnesota. His research interests include subjective forecasting, belief assessment, and practical reasoning.

P. George Benson is Professor and Dean of the Graduate School 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.

Authors' addresses:

Kathleen M. Whitcomb, Department of Management Science, College of Business Administration, University of South Carolina, Columbia, SC 24208, USA.

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

Shawn P. Curley, Department of Information and Decision Sciences, Carlson School of Management, University of Minnesota, 271 19th Avenue S., Minneapolis, MN 55455, USA.

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