ORGANIZATIONAL BEHAVIOR AND HUMAN DECISION PROCESSES 43, 145–171 (1989)

Probability Judgment Accuracy: China, Japan, and the United States

J. FRANK YATES

The University of Michigan

YING ZHU

Peking University

DAVID L. RONIS

The University of Michigan

DENG-FENG WANG

Peking University

AND

HIROMI SHINOTSUKA AND MASANAO TODA

Hokkaido University

Subjects in China, Japan, and the United States reported probability judgments. In Study 1, Chinese and American subjects indicated degrees of certainty about their answers to general-knowledge questions with discrete alternatives, e.g., whether potatoes grow better in warm or in cool climates. In Study 2, Japanese subjects made similar discrete-alternative assessments. In Study 3, subjects in China and the United States reported probability distribution judgments for various quantities, e.g., the maximum temperature on a specified day. Judgment accuracy was evaluated overall and with respect to several underlying accuracy dimensions. The overall quality of discrete-alternative judgments was indistinguishable among the subjects from the three countries. The accuracy component patterns of the Japanese and American subjects were essentially the same. However, the Chinese subjects achieved the common overall accuracy level very differently. On some accuracy dimensions, e.g., calibration, the American and Japanese subjects' judgments were superior. On others, e.g., discrimination, the assessments of the Chinese subjects excelled. Results for quantity judgments were similar to those for discrete-alternative judgments, although there were notable differences. Potential explanations and implications are discussed.

David L. Ronis is now at the Michigan Health Care Education and Research Foundation, Detroit, MI. Masanao Toda is presently in the Psychology Department at Chukyo University, Nagoya, Japan. We are indebted to Sarah Lichtenstein and Lawrence Phillips for the general-knowledge questions used in Studies 1 and 2. It is our pleasure to acknowledge Beth

0749-5978/89 $3.00

The present article is about likelihood judgment accuracy. From a practical standpoint, this issue is important because decisions can be no better than the accuracy of the judgments on which they rest. If an individual's decisions tend to turn out poorly, the quality of the judgments supporting those decisions is a plausible potential culprit. Simply discovering whether poor judgment *is* the cause of bad decisions is useful in itself. But the accuracy issue is significant theoretically, too. Suppose we seek to understand how likelihood judgments are formed. Indications of how accuracy responds to experimental manipulations provide a window on the underlying mechanisms.

Probability judgments are only one of several commonly used forms of likelihood judgment. However, they are attractive for many reasons. One advantage is that they permit explicit and careful tradeoffs between the certainty of events and the seriousness of their consequences, e.g., in expected utility computations in modern technologies such as decision analysis (cf. Raiffa, 1968; von Winterfeldt & Edwards, 1986; Winkler, 1972). They are also appealing because their precision facilitates the study of likelihood judgment quality.

Calibration is the aspect of probability judgment accuracy that has received more attention than any other (cf. Lichtenstein, Fischhoff, & Phillips, 1982). Probability judgments are "well-calibrated" to the extent that the judgments attached to various events match the relative frequencies with which those events actually occur. Consider 1,000 days, on each of which a weather forecaster says that there is a 70% chance of rain. Then, if that forecaster's judgments are perfectly calibrated, rain will be observed on exactly 700 of those days.

General-knowledge questions are a popular tool for studying calibration. Typically, the subject is asked to pick one of two alternatives, and then to report a probability judgment that he or she has selected the correct answer. An example of such items, sometimes called "almanac questions," is the following:

Which is farther north? (Check one):

Hoetger, Halimah Hassan, and Ju-Whei Lee for their computer programming assistance. We also appreciate the critical comments and suggestions made by George Wright, Lawrence Phillips, Shawn Curley, Ju-Whei Lee, Sarah Lichtenstein, and an anonymous reviewer on earlier versions of the manuscript. The reported work was supported in part by U.S. National Institute of Mental Health Grant MH16892 and by a grant from the Rackham School of Graduate Studies at the University of Michigan. Some of the present research was described previously in Chinese (Yates, Zhu, Ronis, & Wang, 1987). Requests for reprints should be addressed to J. Frank Yates, 136 Perry Building, Department of Psychology, University of Michigan, 330 Packard Road, Ann Arbor, MI 48104-2994; Ying Zhu, Department of Psychology, Peking University, Beijing, China; or Hiromi Shinotsuka, Department of Behavioral Sciences, Hokkaido University, Bungakubu, N.10 W.7, Kita-ku, Sapporo 060, Japan.

_____ (a) London
_____ (b) New York.

Circle the probability that most closely describes how certain you are that your chosen alternative is indeed correct:

50% 60% 70% 80% 90% 100%.

A robust finding in almanac question studies is that people's judgments are miscalibrated in a specific way: they tend to be too high. For instance, on average, Lichtenstein and Fischhoff's (1977) subjects indicated 72.4% certainty in the correctness of their answers to a series of general-knowledge questions. However, only 63.8% of those answers were actually correct. This phenomenon is often interpreted as "overconfidence" (e.g., Fischhoff & MacGregor, 1982). This is not unreasonable, since the probability judgment stated in response to an almanac question actually applies to the event "The answer I chose is correct."

In the 1960s, Lawrence Phillips (personal communication, August 7, 1987) was studying probabilistic judgment in the United States. He noticed that foreign student subjects from Asia seemed to report judgments characteristically different from those of others. Some years later, Phillips had the opportunity to pursue the issue. In collaboration with George Wright at Brunel University in England, Phillips was able to document differences of the sort he suspected. In a fascinating series of studies, Phillips and Wright consistently found that southeast Asians' almanac question judgments were even more strongly positively biased than were British subjects' assessments (Phillips & Wright, 1977; Wright & Phillips, 1980; Wright, Phillips, Whalley, Choo, Ng, Tan, & Wisudha, 1978). The effect was found whether questions were posed in English or in subjects' native languages. Moreover, it was observed for managerial and clerical workers as well as for students.

STUDY 1

Study 1 sought to determine if the East vs West differences found by Wright and Phillips would generalize in a comparison of judgments made by subjects in China and the United States. One reason for doubt about the generalization is that in some judgment situations Americans and Europeans have been found to differ in their confidence. Svenson (1981), for example, asked Swedish and American car drivers to compare their personal driving skills to those of their fellow drivers in the same room. Sixty-nine percent (69%) of the Swedish drivers felt that they were more skillful than the median driver. Remarkably, 93% of the American drivers considered themselves to have better than average skills. A comparison of Chinese and American judgments is also of interest because there are

significant differences between the cultures and social systems of China and the societies of the Asian groups that participated in the Wright and Phillips studies, e.g., Malays, Indonesians, Malaysian Indians, and Hong Kong Chinese. It is easy to offer reasons why some of those societal distinctions might lead to differences in almanac question probability judgments.

Unfortunately, studies contrasting the accuracy of probability judgments made by British and Asian subjects have limited their attention to calibration. Accordingly, we know nothing about how such judgments compare in terms of overall accuracy or other important accuracy dimensions besides calibration. The present study of Chinese and American general-knowledge probability judgment accuracy was designed to evaluate overall accuracy as well as several components of such accuracy, including but not limited to calibration.

Method

Subjects

The Chinese subjects in the study included 62 persons. They were mainly psychology students at Peking University and the Institute for Psychology in Beijing. A few, however, were faculty members in the Peking University Psychology Department. The subjects participated in the study as an exercise in a judgment course being taught at the university.

The group of American subjects included 44 individuals. They were students at the University of Michigan. They were recruited from the University's Human Performance Center subject pool and were paid for their services.

Materials

The American subjects responded to a collection of 51 general-knowledge questions. Twenty-nine of the 31 items presented to the Chinese subjects were selected from the 51 considered by the American subjects. These were items the investigators expected would be equally difficult for Chinese and American subject populations. An illustrative item was, "Potatoes grow best in (a) cool climates or (b) warm climates?"

Procedure

The American subjects participated in the study in groups of varying sizes. All the items were presented to each subject in a booklet. After receiving general instructions, practicing the procedure, and having their questions answered, the subjects responded to the items at their own pace. On each item, the subject indicated the probability that his or her

chosen answer was correct by placing a slash through a probability line. Those responses subsequently were rounded to 50%, 60%, 70%, 80%, 90%, or 100%.

All the Chinese subjects took part in a single group session. General instructions were given, after which the subjects considered a practice item, and then had their procedural questions answered. Each item, translated into Chinese, was displayed on an overhead transparency projector. Subjects reported their chosen answers and probability judgments on a one-page response sheet that, for each item, listed the alternative answers (a) and (b), and the six allowable probability judgments, i.e., 50%, 60%, . . . , 100%.

Results and Discussion

The reported analyses were applied to subjects' responses to the 29 items that were presented to both the Chinese subjects and the American subjects. The first two columns of Table 1 show the medians of the various accuracy measures that were computed for each subject in the Chinese and American samples. The fourth column shows the significance levels of statistical tests applied to those measures.

Overall Accuracy

Proportion correct. Let the target event in a judgment task be labeled A, e.g., A = "My chosen answer is correct" in an almanac question situation. An "outcome index" function d is defined as

$$\begin{aligned} d &= 1, \quad \text{if event A occurs} \\ &= 0, \quad \text{if event A does not occur.} \end{aligned} \tag{1}$$

The mean of the outcome index, $\bar{d}$, is the proportion of times the target event occurs. In the present case, it is simply the proportion of correct almanac question answers selected by the subject.

Table 1 indicates that, although the Chinese subjects were slightly more successful than the Americans at selecting the correct alternatives, this difference was not statistically reliable. This is important for the remaining analyses of probability judgments. Lichtenstein and Fischhoff (1977) found that the extent of overconfidence manifested in subjects' general-knowledge question probability judgments depended on the difficulty of the questions. They assumed that an item's difficulty is indicated by the overall proportion of times the item is answered correctly. Lichtenstein and Fischhoff observed that overconfidence is most evident for judgments concerning hard items. There was evidence that, for easy items, i.e., those that many people answer correctly, *under*confidence sometimes occurs. Thus, item difficulty would have been a potential confound-

TABLE 1
MEDIAN ACCURACY MEASURES AND COMPARISONS

Component/Measure[a]	Group[b]			Comparison significance (p)[c]			
	PRC (1)	USA (2)	JPN (3)	1 vs 2	1 vs 3	2 vs 3	All
Overall							
$\bar{d}$ ↑ (proportion correct)	.690	.655	.655	n.s.	<.01	n.s.	<.04
$\overline{PS}$ ↓	.2214	.2121	.2093	n.s.	n.s.	n.s.	n.s.
$\overline{PS}$ < .25 ↑	71.0%	63.6%	77.2%	n.s.	n.s.	n.s.	n.s.
Calibration							
CIS ↓	.0683	.0555	.0446	<.07	<.001	<.001	<.01
Bias 0 (overconfidence)	.134	.072	.084	<.005	<.001	n.s.	<.001
CIL ↓	.0181	.0063	.0108	<.02	<.05	n.s.	<.03
Discrimination							
MR ↑	.0522	.0402	.0510	<.05	n.s.	n.s.	n.s.
Slope ↑	.117	.089	.099	<.03	<.08	n.s.	<.06
Noisiness							
Scat (f) ↓	.0323	.0252	.0236	<.001	<.001	n.s.	<.001

[a] Definitions in text; ↑—larger values better; ↓—smaller values better; 0—zero the best value.
[b] PRC—China (N = 62); USA—United States (N = 44); JPN—Japan (N = 92).
[c] Pairwise comparisons via Mann–Whitney test; multiple group comparisons via Kruskal–Wallis test; n.s.—$p \geq .10$.

ing consideration if the items selected for the present study were ones the American subjects found easy, but the Chinese subjects found hard. It is unclear that such a confound was not present in some of the original studies comparing British and southeast Asian subject groups.

Probability scores. There are several common procedures for indexing probability judgment accuracy (Shapiro, 1977; Winkler & Murphy, 1968; Yates, 1981, 1982). However, the most widely used measure is attributed to Brier (1950) and, in various forms, is known as the "Brier score," the "quadratic score," and the "probability score." This was the measure employed in the present study. For a given judgment occasion, the "probability score" (PS) is defined by

$$PS(f,d) = (f - d)^2, \tag{2}$$

in which f is the person's probability judgment for target event A.

The outcome index d can be seen as the probability judgment reported by a clairvoyant with perfectly accurate opinions about event A. PS is a squared error loss function. It measures the degree of closeness between the given individual's judgments and those of the clairvoyant. PS is bounded by 0 and 1. The more accurate the person's judgments are, the smaller PS will be. Normally, we seek a sense of a person's *general* skill at making probability judgments for event A, not simply the skill demonstrated in a single instance. This is provided by the mean of PS over many different occasions on which event A might occur, $\overline{PS}$. The sampling distributions for $\overline{PS}$ and most of the other statistics discussed below have not been carefully studied. Accordingly, tests of the reliability of group differences with respect to those statistics have been made via nonparametric procedures (Siegel, 1956).

Over all items and subjects, $\overline{PS} = .2258$ for the Chinese subjects, and $\overline{PS} = .2204$ for the American subjects. That is, the overall accuracy of probability judgments was virtually the same for both subject groups. $\overline{PS}$ was computed for each subject individually, too. Table 1 contains the median values of $\overline{PS}$ for the Chinese and the American subjects. As indicated in the table, the distributions of $\overline{PS}$ were indeed not significantly different from each other.

Suppose a person thinks he or she cannot anticipate the target event's occurrence. A reasonable strategy in this situation is to assign the same probability to all the possibilities. An individual who follows this strategy is called a "uniform judge." In the present study, a uniform judge would select answers at random and report 50% probability judgments that each of the chosen alternatives is correct. It can be shown that this approach yields $\overline{PS} = .25$. Interestingly, it is quite possible to do worse than the uniform judge. In fact, an often revealing statistic is the percentage of

subjects who achieve a better accuracy level than the uniform judge. Table 1 shows that the percentage of subjects surpassing that standard was better for the Chinese than for the American sample, but not significantly so.

Accuracy Component Analyses

Calibration. Calibration has most often been studied with the aid of calibration diagrams.[1] Figure 1a is the calibration diagram summarizing the judgments of both the Chinese and the American subjects. Note that it includes the responses of all the subjects in each group, those of a "megasubject," as it were. The horizontal axis is defined by the subjects' judgments. The vertical axis is identified with the corresponding relative frequencies for the target event, i.e., proportions of correct answers in the present study. The number adjacent to each point indicates the percentage of cases represented by that point. The points are constructed such that the area of each is proportional to the given percentage. The open points are for the Chinese subjects' judgments, the filled points for the Americans' assessments. Thus, for instance, the top point on the lower calibration curve indicates that 47.7% of the Chinese subjects' judgments were for 100% certainty, and that approximately 83% of their answers to the relevant questions were in fact correct.

Two types of calibration can be distinguished (Yates, 1982, 1984). The kind most easily recognized in a calibration diagram refers to the match of individual judgment categories to the corresponding mean outcome indexes, e.g., the closeness to 60% of the proportion of correct answers for which the person indicated 60% certainty, and so forth. This variety of calibration is called "calibration-in-the-small." If a person's probability judgments were perfectly calibrated-in-the-small, the calibration curve for those assessments would lie along the 1:1 diagonal of the calibration diagram. In terms of calibration-in-the-small, overconfidence in answers to general-knowledge questions is best evidenced by a calibration curve that lies to the right of the 1:1 diagonal. Figure 1a indicates that both the Chinese and the American subjects' judgments were overconfident over most of the probability range. It also suggests that the Chinese subjects' overconfidence was somewhat stronger than that of the American subjects.

Calibration-in-the-small is measured by the calibration-in-the-small index (CIS):

[1] In meteorology, calibration is known as "reliability." Because the latter term already has a special meaning in psychology, the term "calibration" is used here.

$$\text{CIS} = (1/N) \sum_{j=1}^{J} N_j (f_j - \bar{d}_j)^2. \quad (3)$$

In Eq. 3, j indexes the various probability judgments the person could report, e.g., $f_1 = 50\%$, $f_2 = 60\%$, etc., in the present study. N_j is the number of times the judgment f_j was offered, e.g., $N_3 = 143 \simeq (.080)(1775)$ corresponding to $f_3 = .70$, as indicated for the Chinese subjects in Fig. 1a. Of course, the total number of judgments is simply $N = \Sigma_{j=1}^{J} N_j \cdot \bar{d}_j$ is the mean outcome index for the given judgment category, e.g., $\bar{d}_3 = .545$ for the Chinese subjects in the current illustration. The formula makes it apparent that the smaller the CIS, the better the calibration-in-the small. As shown in Fig. 1a, CIS was larger for the Chinese subjects than for the American subjects, indicating that the judgments of the latter were better calibrated. Table 1 shows that the difference in the distributions of CIS values for the individual subjects in the two groups approached statistical significance.

An alternative view of judgment accuracy is afforded by covariance graphs (Yates & Curley, 1985). Figures 2a and 2b are the covariance graphs for the judgments made by the Chinese and the American subjects. The outcome index defines the horizontal axis of each graph. For convenience, the events identified with the alternative values of the outcome index are indicated, too, i.e., answers being either "Correct" ($d = 1$) or "Incorrect" ($d = 0$). The number in parentheses adjacent to each value of the outcome index indicates how often the corresponding event actually happened. For instance, in Fig. 2a it is shown that the Chinese subjects selected the correct alternative 1,208 times, but were wrong on 567 occasions. The vertical axis of each graph describes the various probability judgments reported by the subjects.

The distributions shown in Fig. 2 are, in effect, proportion histograms; the sum of the proportions represented in the left- and right-hand histograms taken together in each graph is 100%. The percentage represented by the longest bar in each histogram indicates the scale. Consider, for example, the histogram on the right-hand side in the Chinese subjects' covariance graph. There it is shown that 39.4% of the *total* of 1,775 judgments made by those individuals were reports of 100% certainty in the correctness of answers that were indeed correct. On the "Incorrect" side of the graph it is indicated that 9.6% of the subjects' judgments were indications of 50% certainty when the selected answers were in fact wrong.

"Calibration-in-the-large" is the second type of calibration implied in the previous discussion. If calibration were as good as possible, then over all occasions, the average probability judgment ($\bar{f}$) would be the same as

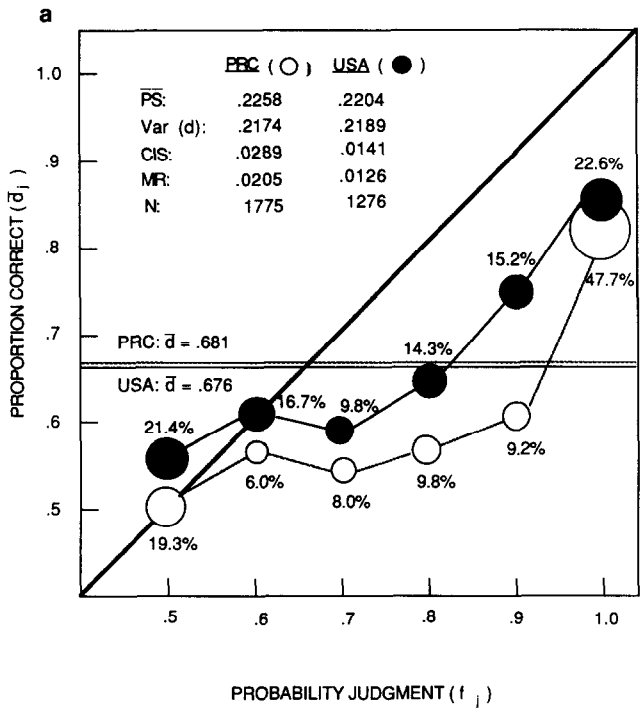


FIG. 1. Calibration diagrams for general-knowledge question probability judgments made by the Chinese (PRC), American (USA), and Japanese (JPN) subjects. The statistics listed are defined in the text. The number adjacent to each point is the percentage of occasions on which subjects used the given judgment category.

the proportion of times the target event actually occurs ($\bar{d}$). Calibration-in-the-large is normally indexed by the "bias" statistic,

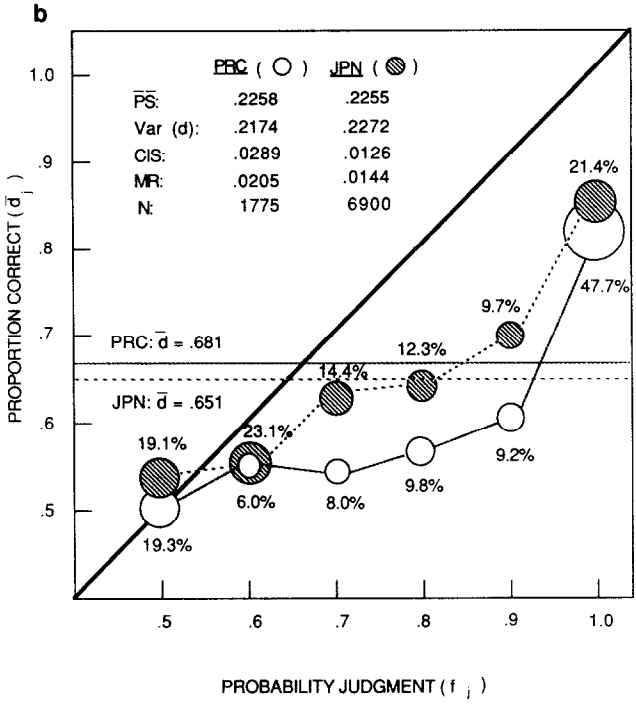
$$\text{Bias} = \bar{f} - \bar{d}, \tag{4}$$

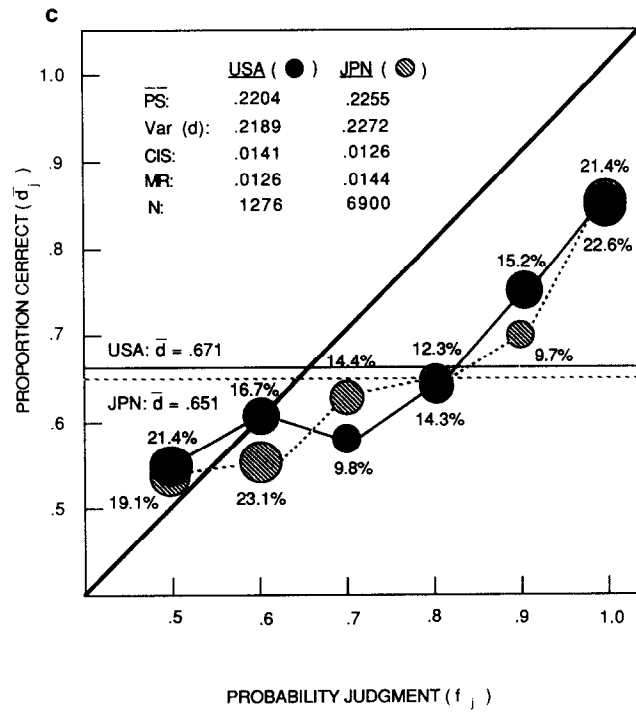
or its square, the "calibration-in-the-large" index (CIL):

$$\text{CIL} = \text{Bias}^2. \tag{5}$$

The larger the absolute value of the bias is, the worse is the calibration-in-the-large.

In the context of general-knowledge questions, calibration-in-the-large is a better indicator of overconfidence than is calibration-in-the-small. In fact, in such situations the bias is sometimes called the "over/underconfidence" statistic (Lichtenstein & Fischhoff, 1977). A person's judgments are overconfident to the extent that the bias is positive and large. Bias is indicated in a covariance graph by the intersection of horizontal and vertical dotted lines. The horizontal line passes through the


b
PRC (○) JPN (▨)
PS: .2258 .2255
Var (d): .2174 .2272
CIS: .0289 .0126
MR: .0205 .0144
N: 1775 6900
PRC: d̄ = .681
JPN: d̄ = .651
21.4%
47.7%
9.7%
12.3%
14.4%
23.1%
19.1%
9.2%
9.8%
6.0%
8.0%
19.3%
PROPORTION CORRECT (d̄ j)
PROBABILITY JUDGMENT (f j)
1.0
.9
.8
.7
.6
.5


c
USA (●) JPN (▨)
PS: .2204 .2255
Var (d): .2189 .2272
CIS: .0141 .0126
MR: .0126 .0144
N: 1276 6900
USA: d̄ = .671
JPN: d̄ = .651
21.4%
22.6%
15.2%
9.7%
12.3%
14.4%
16.7%
14.3%
21.4%
9.8%
19.1%
23.1%
PROPORTION CERRECT (d̄ j)
PROBABILITY JUDGMENT (f j)
1.0
.9
.8
.7
.6
.5

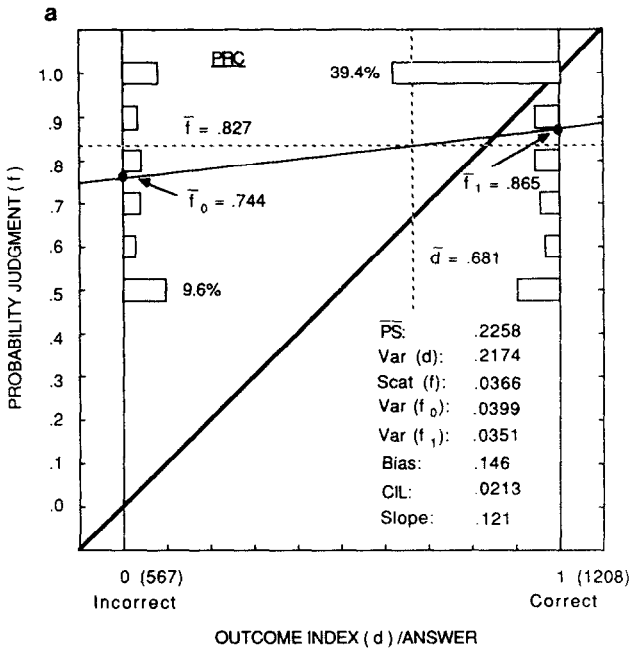
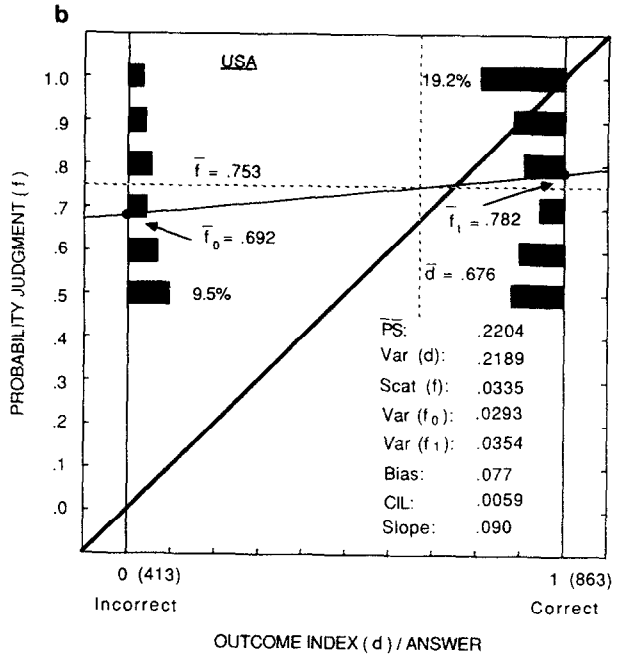


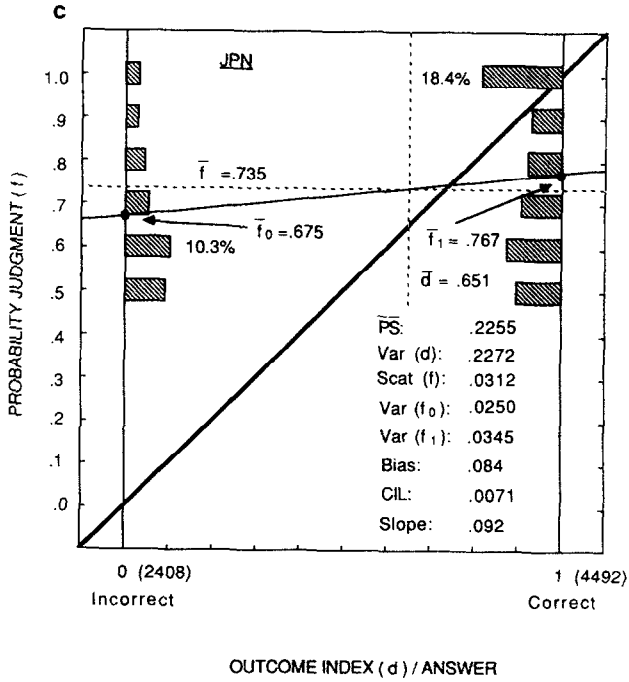
FIG. 2. Covariance graphs for general-knowledge question probability judgments made by the (a) Chinese (PRC), (b) American (USA), and (c) Japanese (JPN) subjects. The statistics listed are defined in the text. The percentage adjacent to each long bar is the percentage of occasions on which subjects reported the given judgment and were either correct or incorrect in their choice of answer.

mean probability judgment ($\bar{f}$). The vertical lines goes through the mean outcome index or "base rate" ($\bar{d}$), which is also the proportion correct in the present study. The bias is positive if the intersection is above the 1:1 diagonal, negative if below, and nil if it lies right on the diagonal.

Figures 2a and 2b show that the biases of both the Chinese and the American subjects were positive. However, the bias of the former group was almost twice as large as that of the latter, indicating much greater overconfidence. The comparison of biases for individual subjects in the two groups was statistically significant, as indicated in Table 1. Median values and tests of CIL are shown in Table 1, too. On an individual subject basis, bias and CIL are redundant. This is not the case overall, however. For example, suppose half the subjects in a group had a bias of $+.15$ and half a bias of $-.15$. The average bias would be 0. But clearly the individual subjects' judgments would be poorly calibrated-in-the-large.

Discrimination. Calibration entails the ability to properly indicate degrees of certainty. In contrast, discrimination refers to the judge's tendency to say something *different*—in any way, numerically or other-


b
USA
19.2%
$\bar{f}$ = .753
$\bar{f}_1$ = .782
$\bar{f}_0$ = .692
$\bar{d}$ = .676
9.5%
PS: .2204
Var (d): .2189
Scat (f): .0335
Var (f_0): .0293
Var (f_1): .0354
Bias: .077
CIL: .0059
Slope: .090
1.0
.9
.8
.7
.6
.5
.4
.3
.2
.1
.0
PROBABILITY JUDGMENT (f)
0 (413)
Incorrect
1 (863)
Correct
OUTCOME INDEX (d) / ANSWER


c
JPN
18.4%
$\bar{f}$ =.735
$\bar{f}_0$ =.675
$\bar{f}_1$ = .767
10.3%
$\bar{d}$ = .651
PS: .2255
Var (d): .2272
Scat (f): .0312
Var (f_0): .0250
Var (f_1): .0345
Bias: .084
CIL: .0071
Slope: .092
1.0
.9
.8
.7
.6
.5
.4
.3
.2
.1
.0
PROBABILITY JUDGMENT (f)
0 (2408)
Incorrect
1 (4492)
Correct
OUTCOME INDEX (d) / ANSWER

wise—on those occasions when the target event is going to happen than on those when it is not. A person's probability judgments are discriminative or "resolved" to the extent that there is any contingency between those judgments and the occurrence or nonoccurrence of the target event. It is irrelevant what the character of that contingency is, e.g., whether high probability judgments are associated with the event's occurrence, and low ones are not.

The extent to which a collection of judgments approaches the ideal of perfect resolution is perhaps most cleanly represented by the "Murphy resolution" (MR) statistic:

$$\mathrm{MR} = (1/N) \sum_{j=1}^{J} N_j (\bar{d}_j - \bar{d})^2. \tag{6}$$

MR increases with the degree to which judgments are well-resolved (cf. Murphy, 1972a, 1972b; Sanders, 1963; Yates, 1982, 1984).

The legitimacy of MR as an index of discrimination is suggested by this argument: Suppose there is indeed a contingency between an individual's judgments and the target event A. Then $P(A|f_j)$ will tend to differ from $P(A|\text{Not } f_j)$. That is, the "true" probability of the target event will differ according to whether the person does or does not offer any particular judgment f_j, e.g., 30%, 80%, etc. And the stronger is the contingency, the greater will be the difference $P(A|f_j) - P(A|\text{Not } f_j)$. Now, if the contingency in question does exist, it will also be the case that $P(A|f_j) \neq P(A)$. In other words, the probability of the target event, conditional on the selection of any given judgment category, will differ from the marginal probability of the target event. Moreover, the strength of the contingency will correspond to the magnitude of the difference $P(A|f_j) - P(A)$. In the notation of Eq. 6, $\bar{d}_j$ is an estimate of $P(A|f_j)$, while $\bar{d}$ approximates P(A). The appropriateness of MR as a discrimination measure follows.

Resolution reveals itself in a calibration diagram through the vertical coordinates of the points. Resolution is good to the extent that the points are far away from the overall relative frequency of the target event, $\bar{d}$. In Fig. 1a, the overall proportions of correct answers by the Chinese and American subjects are identified by the horizontal dotted and solid lines, respectively. As is perhaps consistent with the visual impression conveyed by the arrays of points, the figure also indicates that MR was higher for the Chinese than for the American subjects' judgments, implying better resolution for the Chinese. As indicated in Table 1, the distribution comparison of MR values for individual subjects in the two groups was statistically reliable.

If a person's probability judgments for event A have good accuracy,

then judgments reported when event A occurs generally should be larger than those offered when event A does not occur. In covariance graphs, the mean of the former judgments is denoted by $\bar{f}_1$, the mean of the latter by $\bar{f}_0$. The "slope" statistic is then defined by

$$\text{Slope} = \bar{f}_1 - \bar{f}_0. \tag{7}$$

This measure is literally the slope of the regression line for probability judgments regressed on outcome indexes, and passes through the points $(0, \bar{f}_0)$ and $(1, \bar{f}_1)$. The slope is another reflection of a person's ability to discriminate occasions when the target event will and will not happen.

Figures 2a and 2b show that, over all judgments, the slope for the Chinese subjects was better than that for the American subjects. As indicated in Table 1, the difference in typical individual-subject slopes was reliable. Thus, the Chinese subjects once again demonstrated better skill at discriminating when their chosen answers to general-knowledge questions were correct rather than incorrect.

Noisiness. The final aspect of judgment accuracy we discuss refers to the "noisiness" of the person's judgments. Assessments are noisy to the degree that they vary in ways that are independent of the occurrence or nonoccurrence of the target event. This can happen for at least two reasons. One source of noise is the inherent inconsistency of the judge. Noise will also occur even if the judge is perfectly consistent, but relies on cues that are themselves only weakly associated with the target event.

Judgment noisiness is best appreciated in covariance graphs. Specifically, it is implicated in a special form of judgment variability. The dispersion of a person's probability judgments about the respective conditional means $\bar{f}_1$ and $\bar{f}_0$ is useless, as far as anticipating the target event is concerned; it is analogous to error variance in the analysis of variance. The amount of random variation, or "scatter," in a collection of judgments is indexed by the variances of those judgments about the conditional means, $\text{Var}(f_1)$ when the target event occurs, and $\text{Var}(f_0)$ when it does not. An overall measure of such variability is provided by the "Scat(f)" statistic, which is a weighted mean of the conditional variances:

$$\text{Scat}(f) = [N_1\text{Var}(f_1) + N_0\text{Var}(f_0)]/[N_1 + N_0], \tag{8}$$

where N_1 is the number of occasions on which the target event occurs, N_0 the number of instances when it does not.

Figures 2a and 2b show that the overall random variability in the judgments of the Chinese subjects was greater than that in the American subjects' judgments. Table 1 indicates that the difference in Scat(f) values for individual subjects in the two groups was statistically significant.

Tradeoffs among accuracy components. The various accuracy dimensions we have described all contribute to overall accuracy. In fact, they compensate for one another. The results indicate that, in terms of overall accuracy, the Chinese and American subjects' probability judgments were equivalent. Nevertheless, on some accuracy dimensions the Chinese judgments were stronger. It was thus necessarily the case that, as we indeed observed, the American subjects' assessments were stronger with respect to other aspects of judgment quality.

The implied tradeoffs are made precise in various decompositions of $\overline{PS}$. The decomposition due to Murphy (1973) shows how calibration and resolution contribute to overall probability judgment accuracy:

$$\overline{PS}(f,d) = \text{Var}(d) - \text{MR} + \text{CIS}. \quad (9)$$

In this equation, $\text{Var}(d) = \bar{d}(1 - \bar{d})$ is the variance of the outcome index.

The roles of bias, slope, and noisiness in determining overall judgment accuracy are clarified in the following "covariance decomposition" of the mean probability score (Yates, 1982):

$$\overline{PS}(f,d) = \text{Var}(d) + \text{MinVar}(f) + \text{Scat}(f) + \text{Bias}^2 - 2(\text{Slope})(\text{Var}(d)). \quad (10)$$

The only new statistic indicated in Eq. 10 is $\text{MinVar}(f) = (\text{Slope})^2(\text{Var}(d))$. Conditional upon achieving a given slope, $\text{MinVar}(f)$ is the total variance that would be present in the person's judgments, even if those judgments contained no random variability.

STUDY 2

Japan has notable similarities to as well as differences from both China and the United States. On the one hand, Japan shares certain cultural and historical traditions with China, e.g., aspects of written language, exposure to common philosophical and religious ideas, and commercial ties spanning hundreds of years. On the other hand, in terms of current broad-based technological development and participation in international economic activity, Japan is more like the United States. Thus, it was not immediately obvious whether probability judgments made by Japanese subjects should be more similar to those of Chinese or American subjects. Study 2 was undertaken with a view toward characterizing the accuracy of Japanese probability judgments.

Method

Subjects

Ninety-two students at Hokkaido University served as subjects. They provided their responses as a course requirement.

Materials and Procedure

The subjects were presented with 75 general-knowledge questions. Forty-nine of these items were drawn from a pool used in calibration studies conducted in Britain and the United States. Similar to what was done in Study 1, these items were expected to be equally familiar to Japanese and Western populations. The remaining 26 items were constructed anew, and involved content of special interest in Japan. An example was the item, "Which prefecture has a larger population, (a) Kanagawa, or (b) Aichi?" The procedure was essentially the same as that used in Study 1, with the exception that the subject indicated his or her 50–100% probability judgment in a blank rather than by placing a slash through a scale or by circling one of several prescribed possibilities. For analytical purposes, each of these responses was rounded to the percentage nearest 50%, 60%, 70%, 80%, 90%, or 100%.

Results and Discussion

The calibration diagrams in Fig. 1b and 1c display the calibration curve for the Japanese subjects' judgments. That curve is shown alongside those for the Chinese and the American subjects, respectively, to facilitate comparisons. Figure 2c shows the Japanese subjects' covariance graph.

As indicated above, the sampling distributions of various accuracy component statistics are not well understood. However, Sarah Lichtenstein (personal communication, October 1987) has observed that both CIS and MR seem to decrease systematically as the number of judgments made by the subject increases. Accordingly, we sought to equate the effective numbers of items considered by subjects in all three groups before making comparisons. Disjoint samples of 29 items were randomly sampled from the entire set of 75 seen by the Japanese subjects. One pair of these samples, i.e., 58 items total, was selected to be representative of the entire item pool with respect to difficulty. Specifically, samples were chosen whose mean proportions correct were not statistically different from each other, from the overall Japanese item pool mean, or from the means of the Chinese and American subjects' judgments. Accuracy statistics were then computed for each subject on each of the 29-item subsets. The average of the two resulting statistics was taken as the pertinent measure for that subject. For instance, a given subject's CIS measure was the average of CIS computed for that person on each of the two 29-item subsets. The medians of various statistics computed for individual Japanese subjects, as well as comparisons to those of Chinese and American subjects, are shown in Table 1.

Table 1 indicates that there were no statistically reliable differences

between the Japanese and American subjects in terms of overall accuracy or any accuracy component. Also, on the whole, the differences between the Japanese subjects' judgments and those of the Chinese subjects paralleled the previously observed Chinese vs American differences. The one noteworthy exception was that, whereas the resolution measure MR was significantly better for the Chinese than for the American subjects, the corresponding comparison between the Chinese and Japanese subjects was not significant.

STUDY 3

There are many situations in which probability judgments about quantities rather than about inherently discrete events are useful. For instance, agricultural officials have a need to know how much rain will fall during a given time period. Production managers need to anticipate what the prices of various supplies will be. Study 3 addressed the accuracy of such judgments.

Probability judgments about quantities are commonly conceptualized in terms of "judged probability distributions." These distributions are typically characterized by their fractiles and the implied credible intervals. A popular procedure for eliciting such distributions is the "fractile method." In this technique, the subject reports several fractiles for the given quantity. In the present application, fractiles associated with the following seven probabilities were requested for each quantity: .01, .10, .25, .50, .75, .90, and .99. As a concrete illustration, the following question was used to obtain the American subjects' .01 fractiles for the average U.S. lifespan: "What number of years is such that there is a 1% chance that the actual average lifespan in the U.S. is that number of years or fewer?"

More generally, suppose that Q is the quantity of interest. The fractile associated with probability p is denoted q_p, and is interpreted to mean that, in the given individual's opinion, $P'(Q \leq q_p) = p$, where P' indicates a probability judgment. For instance, if a weather forecaster believes there is a 10% chance that there will be no more than 1.5 cm of rain during a certain period, this would imply that $q_{.10} = 1.5$. A "credible interval" is a range of potential values for a quantity, along with an indication of the chance that the actual value of the quantity will be contained in that interval. For example, the .10 and .50 fractiles define a 40% credible interval, $(q_{.10}, q_{.50}]$.[2] That is, $P'(q_{.10} < Q \leq q_{.50}) = .40$; the person feels that there is a 40% probability that the actual value of the quantity will fall within the interval.

[2] In this notation, a parenthesis means that the interval excludes the associated endpoint, while a bracket implies that the interval includes the endpoint.

An individual's probability judgments about quantities exhibit perfect "distribution calibration" if all of his or her credible intervals are perfectly calibrated in the previous sense of the term. That is, for any probability $X\%$, the actual values assumed by various quantities fall within exactly $X\%$ of that individual's $X\%$ credible intervals. Suppose a weather forecaster reported judged probability distributions for the amount of rainfall on 1,000 different occasions. For each of those distributions a 40% credible interval could be identified. If the forecaster's judgments have perfect distribution calibration, then the actual amounts of observed rainfall will lie within the specified intervals in 400 of those cases; they will be outside those intervals in the remaining 600.

Suppose that p is a small probability, e.g., 1%. The observation of an actual quantity value smaller than q_p or larger than q_{1-p} is referred to as a "surprise" (cf. Alpert & Raiffa, 1982). For example, if the actual amount of rain that falls in an area is smaller than $q_{.01}$ or larger than $q_{.99}$ in a forecaster's judged probability distribution, then we would say that a "2% surprise" has occurred. The terminology is appropriate; the forecaster indeed should be surprised upon observing an amount of rain in what he or she thought was a highly improbable range of values. Over many occasions, the number or proportion of surprises is called the "surprise index."

A special type of distribution miscalibration is commonly interpreted as overconfidence. A person's opinions are said to be overconfident if the surprise indexes for extreme credible intervals in his or her judged probability distributions are larger than they would be under perfect calibration. This implies that those distributions are too narrow. Such a distribution indicates that the person is overly certain that the actual magnitude of the given quantity will fall close to some specific value, e.g., the median of the distribution.

Judged probability distributions of professionals, e.g., meterologists (Murphy & Winkler, 1974) and accountants (Tomassini, Solomon, Romney, & Krogstad, 1982), have been found to be slightly overconfident. The overconfidence of distribution judgments by laypersons is often marked. For example, Alpert and Raiffa (1982) had students make probability distribution judgments concerning various quantities about which the average person could be expected to have reasonable awareness, e.g., the number of students enrolled in a local doctoral program. The 2% surprise index for the students' judgments was an enormous 42.6%.

The main purpose of Study 3 was to test whether the calibration difference between Chinese and American subjects' probability judgments for discrete events observed in Study 1 would generalize to distribution judgments. Wright and Wisudha (1982) reported results suggesting that the typically observed difference in the calibration of British and Asian

subjects' general-knowledge question probability judgments does not extend to opinions about events that might happen in the future, e.g., that " 'At least one national leader (president or prime minister, etc.) (a) will (b) will not die during the next 30 days' " (p. 221). So a secondary aim of the present study was to determine if such a similar interaction of subject group and temporal focus, i.e., past vs future, would characterize distribution judgment calibration.

Method

Subjects

The Chinese subjects who participated in the study were 60 students at Peking University. They were compensated for their services. The 54 American subjects were students in a psychology course at the University of Michigan. They took part in the study as part of a course project.

Materials

Each group of subjects made judgments about 20 different quantities. Half the items concerned quantities whose values existed at the time the subjects responded to the items, e.g., the length of the Yangtze River, for the Chinese subjects. The remaining items were about quantities that would only exist in the future, e.g., the minimum temperature recorded in Guangzhou a few days hence.

The item pools seen by the two groups of subjects were constructed to be parallel to each other. Some target quantities were identical for both groups, e.g., the area of Australia. Other items were designed to be different, but comparable. For instance, instead of being asked about the minimum temperature recorded in Guangzhou, the American subjects were requested to forecast the high temperature in Miami.

Procedure

The Chinese subjects participated in group sessions. The procedure was explained and practiced, after which the subjects' questions were answered. The items were presented one at a time. As indicated above, subjects' judgments were elicited via the fractile method.[3] They reported the fractiles of their judged probability distributions on individual response sheets.

[3] It is known that overconfidence is more strongly evidenced in judgments obtained through the fractile method than by other procedures (e.g., Seaver, von Winterfeldt, & Edwards, 1978). The technique was used here nevertheless because it lends itself to judgments for quantities on different scales and because the primary concern was with relative rather than absolute amounts of overconfidence.

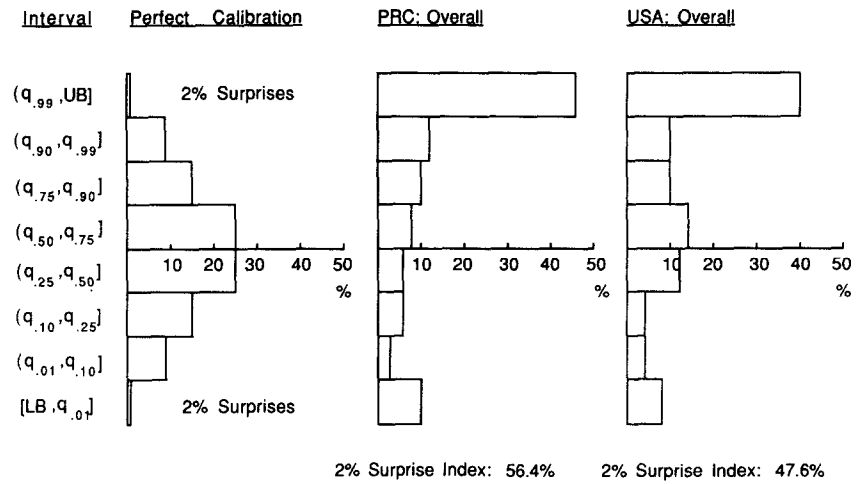


FIG. 3. Proportions of actual quantity values falling within credible intervals of Chinese (PRC) and American (USA) subjects' judged probability distributions.

The items considered by the American subjects were presented to them in a booklet. The procedure was explained and practiced in a large group session. The subjects were then allowed to respond to all the items in the booklet at home, returning the completed booklets at the next class meeting. The subjects were asked to perform the task alone and without consulting any other source.

Results and Discussion

Figure 3 shows the proportions of actual quantity values, over all items, falling within the various consecutive credible intervals defined by the fractiles reported by the Chinese and American subjects. These proportions are indicated by the second and third histograms. The first histogram shows what the observed proportions would have been if distribution calibration had been perfect.

As Fig. 3 indicates, the distribution judgments of both subject groups yielded 2% surprise indexes that were much too high, implying marked overconfidence. Tests of individual surprise indexes against their expectations according to perfect calibration were highly significant statistically for both the Chinese and the American subjects ($t(59) = 35.08, p < .0001$, for the Chinese subjects; $t(53) = 21.75, p < .0001$, for the American subjects).[4] The fact that, in each instance, most surprises were in the ($q_{.99}$,

[4] To accommodate potential variance stability problems, tests were performed on surprise indexes (P) transformed as follows: $\arcsin(P^{1/2})$.

UB] rather than the [LB, $q_{.01}$] intervals ("UB" stands for "upper bound," "LB" for "lower bound") implies that the subjects tended to underestimate the various quantities. Nevertheless, for both subject groups, the surprises in the [LB, $q_{.01}$] intervals were far more numerous than perfect calibration would allow ($t(59) = 6.31$, $p < .0001$, for the Chinese subjects; $t(53) = 5.13$, $p < .0001$, for the American subjects). Thus, subjects' overconfidence was not simply a result of their underestimating the quantities. Consistent with what has been found for judgments concerning discrete events, the surprise indexes for the Chinese subjects were higher than those for the American subjects ($t(112) = 2.65$, $p < .01$).

Figure 4a illustrates the calibration comparison of current vs future judgments made by the Chinese subjects. Figure 4b does the same for the American subjects' responses. As the graphs and surprise indexes indicate, both groups exhibited less overconfidence in their judgments about future rather than current quantities ($t(59) = 2.36$, $p < .03$, for the Chinese subjects; $t(53) = 3.74$, $p < .001$, for the American subjects). These results are consistent with the main effect of temporal focus observed by Wright and Wisudha (1982) in the context of discrete event judgments. However, contrary to the interaction hypothesis suggested by their findings, there was no evidence that, for future quantities, the distribution calibration of the Chinese subjects was better than that of the American subjects.

GENERAL DISCUSSION

Major conclusions concerning the Chinese vs American comparisons are summarized as follows: Previous research has shown that the calibration of general-knowledge question probability judgments differs for British subjects as compared to various groups of Asian subjects. The present work revealed a parallel difference between the judgments of American and Chinese subjects. Moreover, this discrete-alternative calibration difference was found to generalize to probability distribution judgments for various quantities. Both groups' distribution judgments were overconfident, and this effect was especially pronounced for the Chinese subjects. The most important new finding was that the calibration difference was complemented by a discrimination difference. Specifically, the Chinese subjects' judgments about their answers to general-knowledge questions were especially good with respect to their ability to distinguish occasions when those answers were correct from occasions when they were not.

It is particularly noteworthy that the overall accuracy levels of the Chinese, American, and Japanese subject groups were virtually indistinguishable. This result is all the more arresting because the routes to this

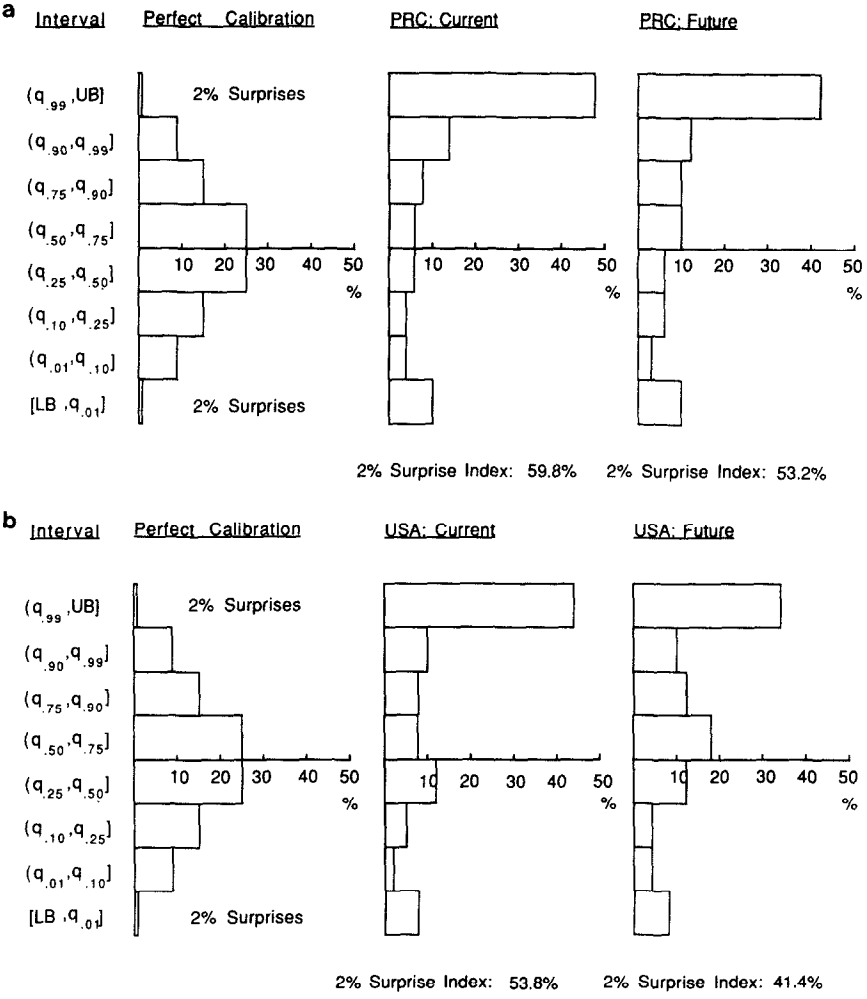


FIG. 4. Proportions of actual values falling within credible intervals of (a) Chinese (PRC) and (b) American (USA) subjects' judged probability distributions for current vs future quantities.

common achievement were so different for the Chinese as compared to the Americans and Japanese. On the whole, the judgments of the latter two groups were essentially the same in character.

The challenge of future studies is to discover why the observed differences in the various *components* of probability judgment accuracy occur, while overall accuracy remains the same. It is conceivable that the differences reflect mere response biases. That is, in terms of true opinions, the distinctions between the national groups might not exist. They only

appear to exist because psychologists' methods accurately assess some groups' actual beliefs but not others'. There are various ways this response bias hypothesis can and should be tested. Assuming that the response bias explanation can be dismissed, more substantive hypotheses should be examined.

Wright and his colleagues (e.g., Wright & Phillips, 1980) suggest that British–Asian calibration distinctions reflect fundamental cultural differences in how people think about uncertainty. According to this view, whereas Westerners tend to think in terms of degrees of certainty, Asians are more apt to view things in absolutes. This is an interesting thesis that should be thoroughly investigated. The results for the Japanese subjects indicate that a simple cultural explanation might be untenable. Instead, the differences perhaps rest equally, if not more heavily, on the subjects' current socioeconomic situations. For instance, technologically oriented societies, such as those in Britain, Japan, and the United Staes, might demand more attention to the kind of precision represented by good calibration.

But why should the Chinese subjects' judgments have such good discrimination, e.g., resolution and slope? Once again, a plausible and testable hypothesis is that this is a reflection of the pervading demands of the society. Good discrimination is possible only if the person reporting judgments fundamentally "knows" what is or is not going to happen. It is conceivable that the reward structure in Chinese society is more generous for outstanding discrimination than it is for proper numerical labeling, e.g., calibration.

Future studies must pursue more than explanations for the specific effects observed here. For one thing, discrete event judgments other than those concerning general-knowledge questions must be collected and analyzed: Do the present conclusions apply to judgments about the kinds of externally determined future events that underlie important practical decisions, e.g., the outcomes of medical procedures and changes in business conditions? Rightly so, some authors (e.g., Ronis & Yates, 1987; Wright & Ayton, 1986) have noted that conclusions derived from studies of responses to general-knowledge questions might not necessarily apply to true forecasting situations.[5] There should also be serious study of judgments by experts. Would results parallel to the present ones be found for assessments made by professionals in their areas of practice, e.g., among meteorologists, physicians, and economic forecasters?

[5] This does not mean that, if the observed differences do not extend to future-event judgments, they are unimportant. The existing data strongly implicate cross-national differences in metacognition about factual knowledge. These effects are inherently interesting. They could also apply to other kinds of knowledge and might have relevance for educational practice. See Newman (1984) for discussion of related issues.

The practical significance of national differences in probability judgment accuracy depends in part on the foundations of those distinctions. Some implications seem independent of the underpinnings, however. At a minimum, the differences suggest that cross-national miscommunication about uncertainty is virtually guaranteed. In many Western countries there is a growing consensus that discussion about uncertainty would be less ambiguous if expression were in the form of probabilities (cf. Beyth-Marom, 1982; Bryant & Norman, 1980; Kong, Barnett, Mosteller, & Youtz, 1986). The present results imply that this approach alone would not necessarily improve international communications. Unless a concerted effort were made to assure a common interpretation of what probability judgments should mean with respect to concepts such as calibration, probability expression might actually *exacerbate* misunderstanding.

Decision analysis often involves applying tools like expected utility maximization to important practical problems. Wright *et al.* (1978) have suggested that, because Asians seem to think about uncertainty differently than Westerners do, decision analysis might be less useful in Asia than in the West. Independently, Pollock and Chen (1986) have also expressed pessimism about the suitability of decision analysis in China, for reasons that seem similar to those of Wright *et al.*: "What we found in China, . . . , was a decision-making environment that was almost completely devoid of a formal concern for uncertainty" (p. 35). Phenomena such as probability judgment miscalibration undoubtedly detract from the value of decision analysis. However, it might be premature to conclude that technologies such as decision analysis are inherently unworkable in China because of such effects. One reason is that miscalibration sometimes can be improved by mere mathematical transformations of individuals' judgments. A more important reason is suggested by the good discrimination evident in the Chinese subjects' judgments in Study 1.

In their attempt to apply decision analysis in China, Pollock and Chen (1986) appear to have experienced the manifestations as well as the possible cause of a Chinese emphasis on discrimination. They noted that the prevailing mode of decision making made it "particularly uncomfortable for our Chinese colleagues to accept uncertainty" (p. 36). Thus, while relatively weak discrimination might be acceptable in Japan and the West, it is too costly to be tolerated in China. As implied by Pollock and Chen's remarks, and in previous suggestions here, the good discrimination exhibited by the Chinese could be simply a matter of current social incentives. That is, individuals might be motivated to devote more effort to achieving good discrimination, possibly at the expense of good calibration. In addition, however, a tradition emphasizing discrimination might have led to cognitive strategies that permit better discrimination than is afforded by the judgment procedures found elsewhere. Future studies

should be directed toward determining whether this is the case. If it is, then the prospects for techniques like decision analysis in China are bright indeed. This is because good discrimination provides a more solid foundation for accurate judgments than does good calibration. Moreover, the judgment procedures which lead to good Chinese discrimination would be worthy of imitation by others.

REFERENCES

Alpert, M., & Raiffa, H. (1982). A progress report on the training of probability assessors. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 294–305). New York: Cambridge Univ. Press.

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

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78**(1), 1–3.

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

Fischhoff, B., & MacGregor, D. (1982). Subjective confidence in forecasts. *Journal of Forecasting*, **1**, 155–172.

Kong, A., Barnett, G. O., Mosteller, F., & Youtz, C. (1986). How medical professionals evaluate expressions of probability. *New England Journal of Medicine*, **315**, 740–744.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know?: The calibration of probability judgments. *Organizational Behavior and Human Performance*, **20**, 159–183.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). New York: Cambridge Univ. Press.

Murphy, A. H. (1972a). Scalar and vector partitions of the probability score: Part I. Two-state situation. *Journal of Applied Meteorology*, **11**, 273–282.

Murphy, A. H. (1972b). Scalar and vector partitions of the probability score: Part II. N-state situation. *Journal of Applied Meteorology*, **11**, 1183–1192.

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

Murphy, A. H., & Winkler, R. L. (1974). Subjective probability forecasting experiments in meteorology: Some preliminary results. *Bulletin of the American Meteorological Society*, **55**, 1206–1216.

Newman, R. S. (1984). Children's numerical skill and judgments of confidence in estimation. *Journal of Experimental Child Psychology*, **37**, 107–123.

Phillips, L. D., & Wright, G. N. (1977). Cultural differences in viewing uncertainty and assessing probabilities. In H. Jungermann & G. de Zeeuw (Eds.), *Decision making and change in human affairs* (pp. 507–519). Dordrecht, Holland: Reidel.

Pollock, S. M., & Chen, K. (1986). Strive to conquer the Black Stink: Decision analysis in the People's Republic of China. *Interfaces*, **16**(2), 31–37.

Raiffa, H. (1968). *Decision analysis*. Reading, MA: Addison-Wesley.

Ronis, D. L., & Yates, J. F. (1987). Components of probability judgment accuracy: Individual consistency and effects of subject matter and assessment method. *Organizational Behavior and Human Decision Processes*, **40**, 193–218.

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

Seaver, D. A., von Winterfeldt, D., & Edwards, W. (1978). Eliciting subjective probability distributions on continuous variables. *Organizational Behavior and Human Performance*, **21**, 379–391.

Shapiro, A. R. (1977). The evaluation of clinical prediction. *New England Journal of Medicine*, **296**, 1509–1514.

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw–Hill.

Svenson, O. (1981). Are we all less risky and more skillful than our fellow drivers? *Acta Psychologica*, **47**, 143–148.

Tomassini, L. A., Solomon, I., Romney, M. B., & Krogstad, J. L. (1982). Calibration of auditors' probabilistic judgments: Some empirical evidence. *Organizational Behavior and Human Performance*, **30**, 391–406.

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

Winkler, R. L. (1972). *Introduction to Bayesian inference and decision*. New York: Holt, Rinehart & Winston.

Winkler, R. L., & Murphy, A. H. (1968). "Good" probability assessors. *Journal of Applied Meteorology*, **7**, 751–758.

Wright, G., & Ayton, P. (1986). Subjective confidence in forecasts: A response to Fischhoff and MacGregor. *Journal of Forecasting*, **5**, 117–123.

Wright, G. N., & Phillips, L. D. (1980). Cultural variation in probabilistic thinking: Alternative ways of dealing with uncertainty. *International Journal of Psychology*, **15**, 239–257.

Wright, G. N., Phillips, L. D., Whalley, P. C., Choo, G. T., Ng, K. O., Tan, I., & Wisudha, A. (1978). Cultural differences in probabilistic thinking. *Journal of Cross-Cultural Psychology*, **9**, 285–299.

Wright, G., & Wisudha, A. (1982). Distribution of probability assessments for almanac and future event questions. *Scandinavian Journal of Psychology*, **23**, 219–224.

Yates, J. F. (1981). *Scoring rules for forecasts* (Tech. Rep. No. 16). Ann Arbor: University of Michigan, Michigan-Chicago Cognitive Science Program.

Yates, J. F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance*, **30**, 132–156.

Yates, J. F. (1984). Evaluating and analyzing probabilistic forecasts. *The UMAP Journal*, **5**, 75–118.

Yates, J. F., & Curley, S. P. (1985). Conditional distribution analyses of probabilistic forecasts. *Journal of Forecasting*, **4**, 61–73.

Yates, J. F., Zhu, Y., Ronis, D. L., & Wang, D.-F. (1987). Probability judgment accuracy in China and the United States (in Chinese). *Information on Psychological Sciences*, No. 2, 5–11.

RECEIVED: June 29, 1987