ORGANIZATIONAL BEHAVIOR AND HUMAN DECISION PROCESSES 40, 193–218 (1987)

Components of Probability Judgment Accuracy: Individual Consistency and Effects of Subject Matter and Assessment Method

DAVID L. RONIS AND J. FRANK YATES

University of Michigan

An experiment is reported in which subjects assigned probabilities to the outcomes of basketball games and to the truth of general-knowledge items. Three different methods were used for eliciting subjects' probability judgments. Subjects were more successful in selecting answers to the general-knowledge questions than they were in picking basketball game winners. The overall accuracy of their probability judgments for general-knowledge items was superior, too. On the other hand, subjects' judgments about general-knowledge questions were more overconfident, more poorly calibrated, and included greater scatter. One method of probability assessment gave subjects an irrelevant cue. This was found to increase confidence and overconfidence and to hurt calibration. Correlations between measures of performance on general-knowledge questions and basketball predictions showed substantial individual consistency in confidence, but only weak consistency in other components of judgment quality. Theoretical and practical implications are discussed. © 1987 Academic Press, Inc.

When forecasters and other judges of unknown events are not *sure* what outcome will occur, it is important that their judgments be appropriately qualified. If such judges can accurately assess the probabilities of outcomes, decision makers can make trade-offs between the values of various outcomes and the chances of their occurrence. A substantial amount of research has been conducted to evaluate the quality of probabilistic judgments, almost all of it focusing on calibration. Probabilistic judgments are said to be well calibrated when the relative frequencies of events match the corresponding judged probabilities. Precipitation forecasts would be perfectly calibrated, for example, if it rains on 10% of days when the forecast is a 10% probability of rain, on 20% of the days when the forecast is for a 20% probability of rain, etc. Professional weather forecasters and professional oddsmakers provide well-calibrated judgments, but almost all other probability judgments that have been ex-

The authors thank Sarah Lichtenstein for providing stimulus materials and Beth Hoetger for assistance in data analysis. This research was supported by NIMH Grant MH 16892. Requests for reprints should be sent to Dr. David L. Ronis, 640 Rolling Green Circle, Rochester Hills, MI 48309.

amined are overconfident and poorly calibrated (Lichtenstein, Fischhoff, & Phillips, 1982).

The probability judgments used in practical decision situations almost always apply to (currently unknowable) future events, such as the weather, the performance of a potential employee, or the helpfulness of a medical treatment for a patient with a given diagnosis. Most psychological research on probability judgments, however, has used general-knowledge questions whose answers are already known to the researchers. These are often an assortment of items about history, vocabulary, geography, literature, and science, sometimes referred to as "almanac questions." There are clear advantages of using such questions. Perhaps the most important is that one does not have to wait to learn the answers. This facilitates the research process and makes it possible to give immediate feedback to the subjects in training studies (cf. Fischer, 1982; Lichtenstein & Fischhoff, 1980).

Unfortunately, two potentially critical differences between the tasks in studies using general-knowledge questions and those in real-world forecasting raise questions about generalizability. The first difference is that real-world forecasters recognize that future events are unknown to anyone and are not definitely answerable at the time predictions are made, while subjects in the laboratory are aware that answers to general-knowledge questions are known to the researchers and hence potentially knowable by the subjects themselves. This difference may create or increase overconfidence in the laboratory. Past research has typically indicated that judgments about general-knowledge questions are indeed more confident than those about future events (Fischhoff & MacGregor, 1982; Wright, 1982; Wright & Wisudha, 1982).

The second important difference between laboratory and real-world probability judgments is procedural. Real-world forecasters typically assign probabilities to predefined target events, e.g., "Rain will fall," "The employee will succeed," "The patient will live." Probabilities for the target event are given on a scale from 0 to 100%. In most laboratory studies, however, judgments are made by a two-step procedure. Subjects are usually given two alternative answer to each question. First they *choose* the answer they believe is more likely to be correct. Then they assign a probability to their preferred answer using a scale from 50 to 100%. It is possible that the act of choosing an answer before assigning a probability may affect the judgment process and the probability assigned.

Self-perception theory (Bem, 1967) and early versions of the theory of cognitive dissonance (Brehm & Cohen, 1962; Festinger, 1957) suggest that freely choosing a course of action will increase the attractiveness of the chosen action and decrease the attractiveness of alternative actions. Thus the two-step procedure (choose then judge) common in laboratory studies of probability judgment may contribute to the observed overconfidence. An exploratory study reported by Fischhoff, Slovic, and Lichtenstein (1977, Experiment 1) provided some support for this hypothesis. In the Fischhoff *et al.* study, subjects were completely sure 22% of the time when they used a two-step procedure and approximately 18% of the time when they used a one-step procedure. However, the results of that study were not definitive because the specific questions differed somewhat for the subjects using the two different response methods.

The current study was conducted in order to examine the effects on probability judgment accuracy of subject matter (predictions of the outcome of basketball games versus general-knowledge questions) and of assessment methods (choice versus no-choice and 50-point scale versus 100-point scale). The study also tested the consistency of individual tendencies in the accuracy of probability judgments about general-knowledge questions and about the outcomes of future basketball games. Determining the extent of such consistency is important for the task of selecting and training forecasters.

There are several separable dimensions of probability judgment accuracy. Judges may differ, for example, in the proportion of times they identify the outcome that occurs. They can also differ in the degree of systematic overconfidence or underconfidence of their judgments and/or in the amount of random variability in their assessments. Thus, the issues described above were examined as they pertained to aspects of judgment accuracy as indexed by seven measures that have appeared in the literature.

METHODS

Subjects

One hundred twenty-eight paid subjects from the local university community participated in the experiment. Subjects filled out the experimental questionnaires in a classroom setting. Completion of the questionnaires took about 40 min.

Design

The factors varied in the experiment were the *topic* of the judgments and the *method* of probability assessment. Each subject answered 51 questions on each of two topics: (a) outcomes of upcoming professional *basketball* games and (b) two-alternative *general-knowledge* questions. The general-knowledge questions covered a wide variety of subjects. The first three questions, for example, were these: (1) A Japan finish is characteristically (a) smooth and glossy or (b) dulled by hand rubbing; (2) Which is the greatest distance from Paris? (a) Honolulu or (b) Cairo; and

(3) The motto of the Boy Scouts of America is (a) be prepared or (b) be helpful.

The basketball questions asked which team would win each of 51 National Basketball Association games to be played in the next few weeks. Each question gave the date of the game and identified the home team and the visiting team. As background information, the questionnaire included the divisional standings of all the teams. Thus all subjects had some information on which to base their predictions. It was anticipated that subjects would be more confident about the general-knowledge questions than about basketball predictions.

Each subject was randomly assigned to use one of three methods of probability assessment. In the standard *Choice-50* method, the subject first circled one of the two possible answers for a question and then assigned a probability from .50 to 1.00 that the chosen answer was correct. In the *No-Choice-100* method, one of the two possible answers was already circled. Subjects were correctly informed that the decision about which answer had been circled was made by flipping a coin. Their task was to assign a probability from 0.00 to 1.00 that the precircled answer was correct. In the *Choice-100* method the subject first circled his or her preferred answer and then assigned a probability from 0.00 to 1.00 that the chosen alternative was correct. (It would be sensible for subjects in the Choice-100 condition to use only the top half of the probability scale.) The Choice-50 and No-Choice-100 methods were included to test the impact of choice on probability judgment accuracy. In comparing only these two methods, however, effects of choice would be confounded with any possible effects of scale length. The Choice-100 method was included to complete the design and eliminate the confounding of scale length and choice.

For comparability with the other methods, data from the No-Choice-100 method were recoded in the following fashion: If a probability greater than .50 was given, the circled answer and probability were taken literally, i.e., the subject was coded as choosing the answer that was precircled and giving the assigned probability. If a probability below .50 was assigned, the subject was coded as choosing the answer that was not precircled and assigning a probability equal to 1.00 minus the marked probability. If the probability of exactly .50 was given, an answer was randomly selected and the probability of .50 was assigned. These recodings were done by computer. Responses from the other response methods were not recoded. It was hypothesized that subjects would be more confident in the choice conditions.

Subjects were randomly assigned to answer either the general-knowledge questions or the basketball questions first. So the experiment consisted of a 2 (topic) × 3 (method) × 2 (order) factorial design. Topic varied within subjects while method and order varied between subjects.

Accuracy Measures

Brier (1950) proposed an overall measure of the accuracy of probabilistic judgments, now known as the "Brier score." In order to provide separate measures of different aspects of judgment accuracy, Sanders (1963), Murphy (1973), and Yates (1982) have proposed decompositions of the Brier score. That is, they have divided the Brier score into several components. The measures used to describe performance in this study included the Brier score, all components of the above mentioned decompositions, and several other descriptive statistics. This report focuses on seven measures that are particularly important and easy to interpret: the Brier score, reliability-in-the-small, proportion correct, mean confidence, mean overconfidence, slope, and scatter.

The Brier score is an overall measure of judgment accuracy. Low Brier scores indicate good judgment. Proportion correct, mean confidence, and mean overconfidence are self-explanatory. Scatter is variability in assigned probabilities that does not correspond to differences in outcome. Low scatter is a desirable characteristic of probabilistic judgments. The slope indicates the extent to which the forecaster assigns higher probabilities to events that occur than to those that do not. High slope indicates good performance on this component of probabilistic judgment. Reliability-in-the-small is a measure of the extent that the probabilistic judgments are well calibrated, i.e., that the proportion correct at each level of confidence equals the stated level of confidence. Low scores on reliability-in-the-small indicate good performance; a score of zero indicates perfect calibration. Formal definitions of these measures follow.

Brier score. Let f denote the probability assigned to the target event the judge is trying to predict. An "outcome index" d for the target event is defined as follows: $d = 1$, if the target event occurs, $d = 0$, if the target event does not occur. The outcome index may be thought of as the probability that would have been assigned by a clairvoyant. The Brier score or "mean probability score" ($\overline{PS}$) is the mean squared difference between the stated probability and the outcome index: $\overline{PS} = \Sigma(f - d)^2/N$, where N is the number of judgments. Clearly, a low Brier score indicates good performance. In the current study and many other studies of probabilistic judgment, there were two alternative answers to each question. So an individual who had no reason to prefer one answer over another should assign a probability of .5 to either answer. An individual answering each question this way would receive a Brier score of .25. Thus, Brier scores over .25 indicate especially poor performance.

Proportion correct, mean confidence, and mean overconfidence. In the present study, and other studies using general-knowledge questions rather than repeated predictions of the same type of event (e.g., rain, worker success, home team victory), the target event was defined as "My preferred answer is correct." Thus, the mean outcome index, $\bar{d}$, is equivalent to the proportion correct. Similarly, with this definition, the mean probability assigned, $\bar{f}$, is equivalent to the judge's "mean confidence" that he or she has selected the correct answer. The difference between mean confidence and proportion correct, $\bar{f} - \bar{d}$, is called the "mean overconfidence" when positive and the "mean underconfidence" when negative. A more general term for $\bar{f} - \bar{d}$ is the "bias." In studies using different definitions of the target event, bias refers to something other than over/underconfidence. In precipitation forecasts, for example, bias may refer to a tendency to judge rain to be more likely than it really is.

Covariance graphs, slope, and scatter. The scatter, slope, and some other aspects of forecasting performance are best illustrated in a graphical display called a "covariance graph" (Yates, 1982; Yates & Curley, 1985). Figure 1 shows the covariance graphs for two sets of probability judgments. The outcome index defines the horizontal axis of a covariance graph. For convenience, the events identified with the alternative values of the outcome index (answer incorrect, answer correct) are also indicated. The vertical axis of a covariance graph indicates the various probability judgments the person makes. In studies like the current one, where the target event is defined as "My preferred answer is correct," covariance graphs are essentially histograms of probability judgments separately for correct and incorrect answers. The mean probability judgment within each histogram is indicated in the graph ($\bar{f}_0$ when incorrect, $\bar{f}_1$ when correct) and connected by a line. The difference between these two means ($\bar{f}_1 - \bar{f}_0$), or the *slope,* is an indication of the judge's success in assigning higher probabilities when his or her answers are correct than when they are incorrect.

Some judges may feel that they cannot distinguish among questions which they are more and less likely to answer correctly, so they assign the same probability for each question—leading to a slope of zero. Most judges, however, believe they have a basis for varying their probability judgments. If they are correct in this belief, at least some of the variability in their assigned probabilities manifests itself in a positive slope. Unless their judgments are perfectly discriminative, some of the variability appears as variability of probability judgments *within* each histogram.

The variability or scatter of probability judgments within each histogram is useless variance as far as anticipating the target event is concerned. It is analogous to error variance in analysis of variance. An overall measure of the amount of *scatter* in a person's judgments is the weighted mean of the variances within the two histograms:

$$\text{Scatter} = (N_1 \operatorname{Var} f_1 + N_0 \operatorname{Var} f_0)/ \\ (N_1 + N_2),$$

where N_1 is the number of instances in which the event occurs, N_0 is the number of instances in which it does not, and Var f_1 and Var f_0 are the conditional variances in those two sets of occasions. Scatter will be particularly high if the probability judgments are influenced by considerations that are not indicative of the outcome.

Reliability diagrams, calibration, and reliability-in-the-small. The tendency of judgments at various levels of confidence to be correct, incorrect, overconfident, underconfident, and well calibrated are best illustrated in figures called reliability diagrams. In such diagrams, the relative frequency or proportion correct ($\bar{d}$) for each level of probability is plotted against the assigned probability. If the judgments are perfectly calibrated, all the points will fall on the diagonal line. Reliability-in-the-small is the mean squared deviation of the points from this line (weighted by the number of observations):

$$\text{Reliability-in-the-Small} = \left(\frac{1}{N}\right) \sum_{j=1}^{J} N_j \, (f_j - \bar{d}_j)^2,$$

where $\bar{d}_j$ is the relative frequency or proportion correct over the N_j occasions when the judge assigns probability f_j and N is the total number of judgments. In calculating reliability-in-the-small, all probability judgments are rounded to the nearest .1. A reliability-in-the-small of zero indicates perfect calibration. Figure 2 presents reliability diagrams for the same subjects shown in Fig. 1. In panel A of the figure we note that the subject was underconfident when assigning the probability of .6, but overconfident when assigning probabilities of .9 and 1.0. The subject whose judgments are shown on panel B of the figure was overconfident when assigning probabilities of .6 and made few assignments more extreme than .6. See Yates (1982) and Yates and Curley (1985) for more complete descriptions of the covariance decomposition of the Brier score and discussion of other decompositions.

It should be noted that the components of decompositions of the Brier score are not independent of each other. A change in judgment that alters the score on one component will usually alter the scores on other components as well. Basing one's probability judgments on an irrelevant cue, for example, tends to increase both miscalibration (reliability-in-the-small) and scatter. A change in judgment that improves one component may even hurt performance on another component. For example, it is possible to eliminate scatter by assigning the same probability on each judgment, but this improvement would prevent the judge from obtaining a positive slope. Given the desirability of probability judgments that are correlated with outcomes, judges are usually willing to accept some scatter in order to increase slope. The complexity of the relations among the components highlights the danger of excluding important measures from a study.

Before presenting the results, we wish to bring attention to some issues which have their origin in how the target event is defined. All schemes for analyzing probability judgment accuracy require a target event that can be repeated over a large number of occasions. Since every event literally occurs only once, the requirement is actually met by examining a set of events that are "essentially similar" to one another. For certain judgment tasks this requirement is easy to satisfy. For instance, in weather forecasting, most observers concede that the event "Precipitation occurs" is substantially the same every time a meteorologist makes a prediction. In other judgment situations, the events of inherent interest are very different from one another. As an example, in the study by Fischhoff and MacGregor (1982), one judgment item concerned the victor in an upcoming mayoral election, while another was about who would win a particular baseball game. The alternatives in general-knowledge questions also constitute essentially unique "events."

Despite the seeming incomparability of the events involved, researchers have still sought conclusions about judgment tendencies indicated by responses to collections of items about one-of-a-kind future events and general knowledge. This is done, as in the present study, by recoding individual responses into a form whereby the designated target event does appear to be essentially the same from one item to the next. Namely, the subject picks one of the alternatives as the best prospect and then indicates a degree of certainty in the selection. The target event subsequently encoded for the subject is "My preferred answer is correct." There is the option to employ this type of "internal coding" with replicable events, too, although this is virtually never done in practice. For example, a weather forecaster could pick either "Precipitation" or "No Precipitation," then state a 50–100% probability that the chosen condition will prevail. That judgment would be interpreted as applying to "My chosen alternative will occur." Internal coding was employed with both types of questions in the present study to permit comparisons between judgments about basketball games and general-knowledge questions.

Internal coding appears innocuous. But it creates surprising problems of interpretation for some standard measures of judgment accuracy, problems that do not exist with externally coded replicable events. For instance, Yates (1982, pp. 153–154) alluded to the ambiguities of resolution measures for internally coded responses. In the present study slope is the statistic whose meaning is complicated by internal coding. If, as in this study, the target event is "My preferred answer is correct," probabilities below .5 are meaningless. Thus, the largest possible slope value is .5. This would be achieved if a probability judgment of 1.0 is given whenever the judge picks the correct alternative, and .5 whenever the judge selects the wrong one. However, the latter condition seems unreasonable, and certainly unlikely to occur. In essence, it says that, whenever the preferred answer is going to be wrong, the judge reports probability .5, which normally indicates complete uncertainty. Because .5 is always reported under these circumstances, in a sense the judge "knows" that the preferred answer is incorrect. And with two alternatives, if one knows that one alternative is wrong, the other must be concluded to be right. So a judge who is capable of attaining the maximum slope should be able to always pick the correct alternative. If a judge in fact consistently made the proper selection, there would be only one conditional mean judgment, $\bar{f}_1$ in the previous notation. Hence, the slope, $\bar{f}_1 - \bar{f}_0$, would be undefined.

It remains an open problem to determine appropriate statistical treatment and interpretation of measures such as resolution and slope for internally coded events when the judge demonstrates "perfect knowledge" of the correct answers, i.e., when he or she assigns only one probability level and always chooses the correct answer, or when he or she assigns exactly two probability levels: one when right and the other when wrong. For the purpose of interpretive clarity, it is fortunate that the ideal of perfect knowledge was never approached in the current study. In the case of slope, when the judge does not achieve perfect knowledge, it is clear that larger slopes are better than smaller ones.

RESULTS

Each of the performance measures was subjected to a topic × method × order repeated-measures analysis of variance. Order did not have significant effects and is not discussed. In general, the results were dominated by main effects of topic and method. Table 1 lists the mean and standard deviation of each of the seven focal measures for the six cells in the topic × method design. Table 2 lists the same kinds of statistics for six supplementary measures.

Topic Effects

The two topics, *basketball* and *general-knowledge,* were significantly different, $p < .001$, on all seven focal measures of the study. These topic

TABLE 1

MEANS[a] OF FOCAL PERFORMANCE MEASURES BY TOPIC AND ASSESSMENT METHOD

	Brier score	Proportion correct	Confidence	Overconfidence	Reliability-in-the-Small	Scatter	Slope
General knowledge							
No-Choice-100	0.255	0.646	0.812	0.165	0.071	0.027	0.070
	(0.076)	(0.104)	(0.068)	(0.107)	(0.049)	(0.009)	(0.057)
Choice-100	0.212	0.698	0.734	0.035	0.048	0.041	0.115
	(0.061)	(0.112)	(0.083)	(0.101)	(0.033)	(0.023)	(0.111)
Choice-50	0.225	0.664	0.762	0.098	0.052	0.027	0.095
	(0.070)	(0.108)	(0.071)	(0.105)	(0.043)	(0.008)	(0.062)
Basketball							
No-Choice-100	0.263	0.589	0.676	0.087	0.048	0.013	0.005
	(0.041)	(0.085)	(0.068)	(0.098)	(0.040)	(0.007)	(0.036)
Choice-100	0.252	0.603	0.625	0.022	0.035	0.013	0.006
	(0.031)	(0.068)	(0.082)	(0.082)	(0.034)	(0.009)	(0.038)
Choice-50	0.259	0.601	0.667	0.066	0.042	0.013	0.001
	(0.025)	(0.070)	(0.073)	(0.085)	(0.025)	(0.007)	(0.028)
Anova results							
Topic main effect $F(1,122)$	19.79***	43.37***	287.73***	15.38***	14.46***	190.48***	139.94***
Method main effect $F(2,122)$	4.73*	2.39 ns	10.31***	16.70***	3.83*	7.22*	2.81 ns
Topic × method interaction $F(2,122)$	2.72 ns	1.31 ns	3.26*	3.59*	1.08 ns	13.07***	2.94 ns

[a] Standard deviations are given in parentheses.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

TABLE 2

MEANS[a] OF SUPPLEMENTARY PERFORMANCE MEASURES BY TOPIC AND ASSESSMENT METHOD

	Proportion completely certain	Proportion completely uncertain	Proportion agreeing with precircled	Probability assigned to precircled	Confidence when agreeing with precircled	Confidence when disagreeing with precircled
General-knowledge						
No-Choice-100	0.298	0.119	0.565	0.551	0.821	0.802
	(0.183)	(0.101)	(0.063)	(0.044)	(0.076)	(0.070)
Choice-100	0.240	0.149	0.561	0.537	0.739	0.728
	(0.129)	(0.127)	(0.060)	(0.033)	(0.078)	(0.099)
Choice-50	0.213	0.107	0.549	0.524	0.763	0.762
	(0.145)	(0.102)	(0.063)	(0.036)	(0.077)	(0.074)
Basketball						
No-Choice-100	0.014	0.091	0.546	0.519	0.681	0.670
	(0.061)	(0.091)	(0.069)	(0.032)	(0.074)	(0.069)
Choice-100	0.009	0.182	0.498	0.492	0.618	0.633
	(0.018)	(0.246)	(0.047)	(0.010)	(0.078)	(0.089)
Choice-50	0.017	0.080	0.497	0.495	0.664	0.671
	(0.053)	(0.169)	(0.047)	(0.017)	(0.077)	(0.073)
Anova results						
Topic main effect *F*(2,122)	296.36***	0.33 ns	39.43***	85.99***	293.05***[b]	209.42***[b]
Method main effect *F*(2,122)	2.76 ns	4.04*	6.97**	14.76***	12.68***[b]	7.18**[b]
Topic × method interaction *F*(2,122)	3.41*	2.07 ns	3.44*	1.62 ns	2.80 ns[b]	3.13*[b]

[a] Standard deviations are given in parentheses.

[b] Order was not included as a factor in analysis of this variable. Hence, the denominator degrees of freedom are 125.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

effects were all stable across the three methods of probability assessment. Most interesting was the finding that, for some aspects of judgment accuracy, subjects' judgments about general-knowledge questions were better than their judgments about future basketball games, but for other aspects, this was reversed.

Performance on general-knowledge questions was significantly better than on basketball questions on the following measures: Brier score, proportion correct, and slope (see Table 1). In fact, the slope for subjects' basketball predictions was not significantly greater than zero, $t(127) = 1.33$, $p > .1$, meaning that the probabilities subjects assigned to basketball outcomes were not reliably higher when they were right about the outcome than when they were wrong about it. In contrast, judgments were more confident, more overconfident, and showed worse reliability-in-the-small, and more scatter for general-knowledge questions than for basketball questions (see Table 1).

Figure 1 shows the covariance graphs for typical performances on general-knowledge questions (1A) and basketball questions (1B), both using the standard Choice-50 response method. Several characteristics identified by the statistical analysis are illustrated in these graphs, notably the higher proportion correct, higher confidence, and higher scatter of judgments about general-knowledge questions. (Overconfidence and slope were not noticeably different for these two subjects.)

Figure 2 shows the reliability diagrams for the same two subjects whose judgments were shown in Fig. 1. This figure also illustrates the higher confidence and proportion correct for the general-knowledge questions. In addition, the better calibration (lower reliability-in-the-small) for basketball questions is apparent.

Figure 3 shows the reliability diagrams for the six combinations of topics and methods in the experiment collapsed over subjects and order. The calibration curves are generally below the diagonal, indicating overconfidence. A higher proportion of correct answers and a greater mean confidence are apparent for general-knowledge questions than for basketball questions. Note that reliability diagrams for group data cannot be counted on to provide information about reliability-in-the-small for the typical subject. If half the subjects were overconfident and half were underconfident, the grouped data might suggest perfect calibration.

In general, subjects were better able to select the correct answer on general-knowledge questions than for basketball predictions, but were less able to assign confidence levels (probability judgments) that corresponded to their success in prediction. It is notable that subjects were completely sure (100% confident) about 25.0% of their general-knowledge answers but expressed such certainty for only 1.3% of their basketball predictions (see Table 2). These sure answers were correct 85% of the time for general-knowledge questions and 69% of the time for basketball predictions (see Fig. 3).

Method Effects and Interactions

The effects of the *method* of probability assessment were more subtle than the topic effects. Brier scores were best (lowest) using the Choice-100 method, worst with the No-Choice-100 method, and intermediate with the Choice-50 method. However, neither proportion correct nor slope differed reliably across the three methods.

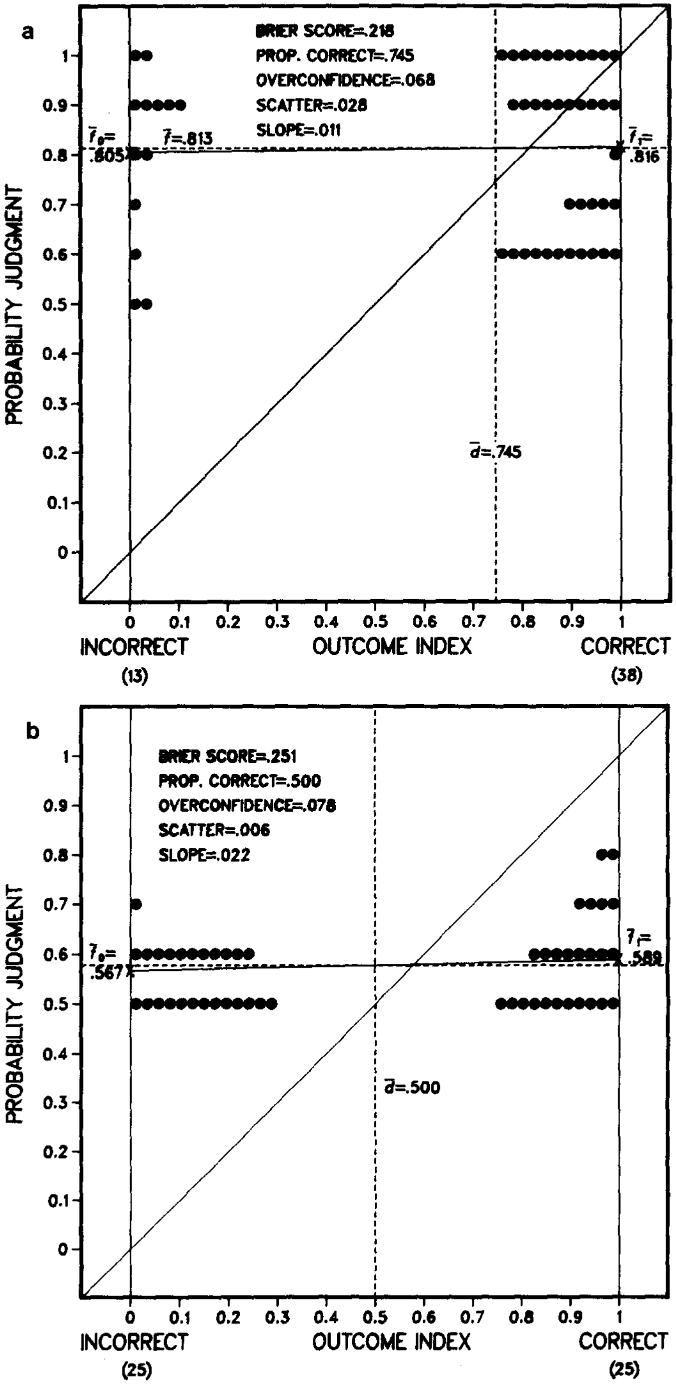
Contrary to our predictions, confidence and overconfidence were highest with the No-Choice-100 method, lowest with the Choice-100 method, and intermediate with the Choice-50 method. Method × topic interactions indicated that these main effects were weaker for basketball predictions than for general-knowledge questions. Reliability-in-the-small showed this same method main effect.

Scatter was higher with the Choice-100 method than with the other methods. A method × topic interaction indicated that this effect held for general-knowledge questions, but was nonexistent for basketball predictions. See Table 1 for descriptive and inferential statistics for all seven measures.

No-Choice-100 method. Why did the No-Choice-100 method increase confidence, overconfidence, and miscalibration (reliability-in-the-small)? With this method, subjects saw one of the two answers circled and judged the probability that the circled answer was correct. So it is possible that subjects (consciously or unconsciously) took the circle as a cue to the correct answer, even though they had been informed that it was irrelevant.

To test this possibility, several additional measures were derived and subjected to analysis of variance: (a) the proportion of times subjects agreed with the precircled answer (whether or not they were in the No-Choice-100 condition), (b) mean probability assigned to the precircled answers, (c) mean confidence when subjects agreed with the precircled answer, and (d) mean confidence when subjects disagreed with the precircled answer.[1] Means and standard deviations for these measures and other supplementary measures are shown in Table 2. These analyses indicated that subjects using the No-Choice-100 method were more likely to agree with the precircled answers than were subjects using the other

[1] In the No-Choice-100 condition, subjects were coded as agreeing with the precircled answer if they assigned it a probability greater than .50. When subjects in this condition assigned probabilities of exactly .50, their preference (for or against the precircled answer) was randomly assigned.


a
BRIER SCORE=.218
PROP. CORRECT=.745
OVERCONFIDENCE=.068
SCATTER=.028
SLOPE=.011
$\bar{f}_0$= .805
$\bar{f}$=.813
$\bar{f}_1$= .816
$\bar{d}$=.745
PROBABILITY JUDGMENT
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
INCORRECT (13)
OUTCOME INDEX
CORRECT (38)
b
BRIER SCORE=.251
PROP. CORRECT=.500
OVERCONFIDENCE=.078
SCATTER=.006
SLOPE=.022
$\bar{f}_0$= .567
$\bar{f}_1$= .589
$\bar{d}$=.500
PROBABILITY JUDGMENT
INCORRECT (25)
OUTCOME INDEX
CORRECT (25)

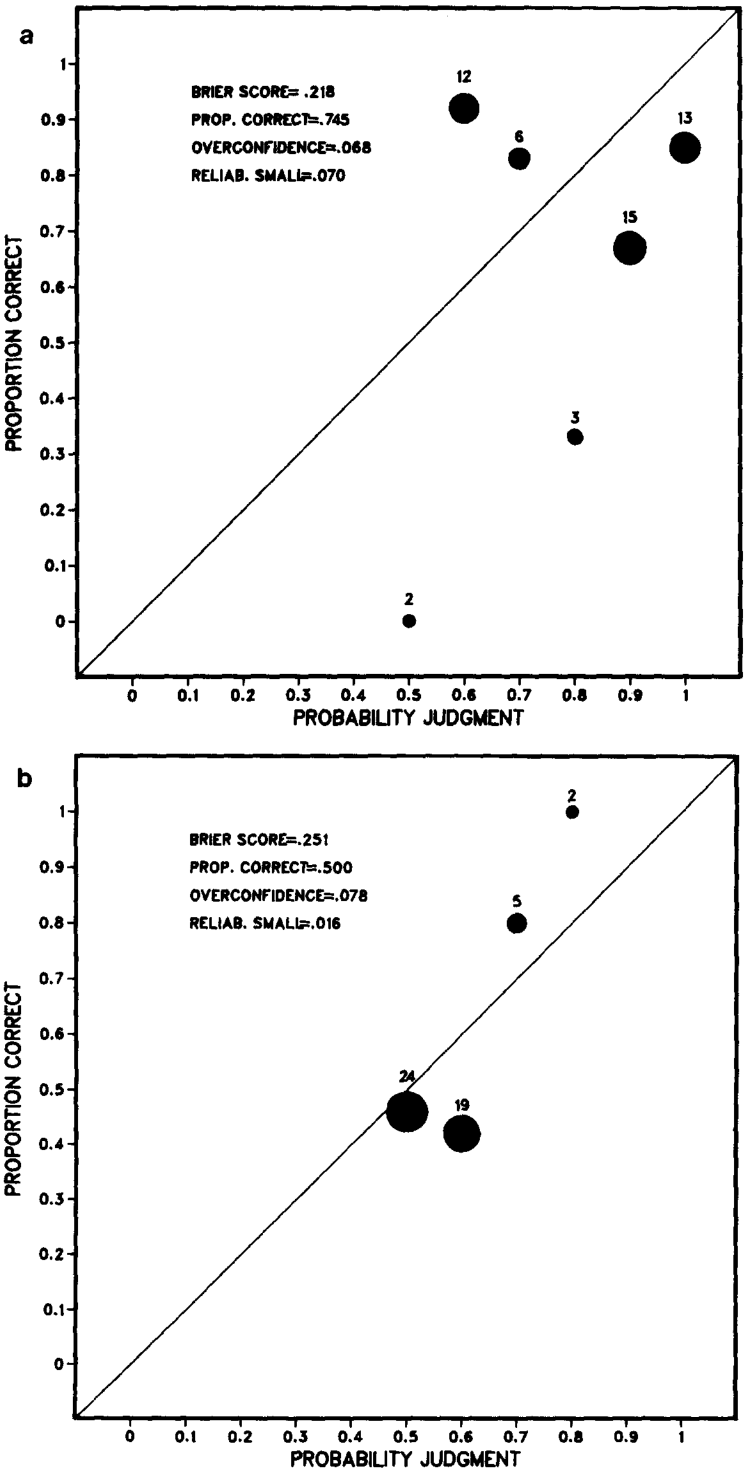
methods. In addition, these subjects assigned higher probabilities to the precircled answers than subjects using the other methods assigned to these answers. Thus, the circle acted as a cue to affect judgment.

However, this cuing effect is only part of the explanation of the high confidence in the No-Choice-100 condition. This was discovered by examining confidence separately for cases in which the subject agreed and disagreed with the precircled answer. Whether subjects agreed or disagreed with the precircled answers, they were more confident when they had used the No-Choice-100 method than when they had used the other techniques (see Table 2). So there must be some other process contributing to the high confidence with No-Choice-100.

Choice-100 method. Why were confidence, overconfidence, and reliability-in-the-small lowest, and scatter highest with the Choice-100 method? This method allowed subjects to choose an answer and then assign a probability lower than .5 to their preferred answer. Twenty of the 42 subjects using the Choice-100 method gave one or more probabilities below .5. Of these 20 subjects, 6 gave only one probability below .5; one subject assigned 56 probabilities below .5 (out of 102 questions), and the remaining 13 subjects gave an average of 9.38 such probabilities. About two-thirds of the probabilities below .5 were assigned to general-knowledge questions. Since there are only two possible answers to each question, assigning a probability below .5 is peculiar and difficult to interpret. It could mean that the subject actually prefers the answer he or she did not circle (a failure to follow the instruction to circle the answer thought to be correct) or it could mean that the subject is misusing the probability scale—assigning probabilities near zero instead of .50 when very unsure of the answer.

An implicit scoring technique can provide information about the relative popularity of these two misuses of the response method. In the implicit technique, whenever a subject reports a probability below .5, the subject's choice is recoded to indicate a *preference for the answer not circled.* Since subjects usually choose the correct answer, implicit scoring will improve the overall proportion correct if that is their meaning. On the other hand, if subjects are misusing the probability scale, implicit scoring will tend to lower the proportion of correct

FIG. 1. Covariance graphs for typical performances on general-knowledge and basketball questions, using the Choice-50 method. Panel A displays the general-knowledge judgments of the subject whose Brier score for general-knowledge questions was closest to the median of subjects using the Choice-50 method. Panel B displays the basketball judgments of the subject whose Brier score for basketball was closest to the median using the same method. Note: In all figures, probability judgments are rounded to the nearest .1.


a
BRIER SCORE= .218
PROP. CORRECT=.745
OVERCONFIDENCE=.068
RELIAB. SMALL=.070
12
6
13
15
3
2
PROPORTION CORRECT
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
PROBABILITY JUDGMENT
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
b
BRIER SCORE=.251
PROP. CORRECT=.500
OVERCONFIDENCE=.078
RELIAB. SMALL=.016
2
5
24
19
PROPORTION CORRECT
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
PROBABILITY JUDGMENT
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

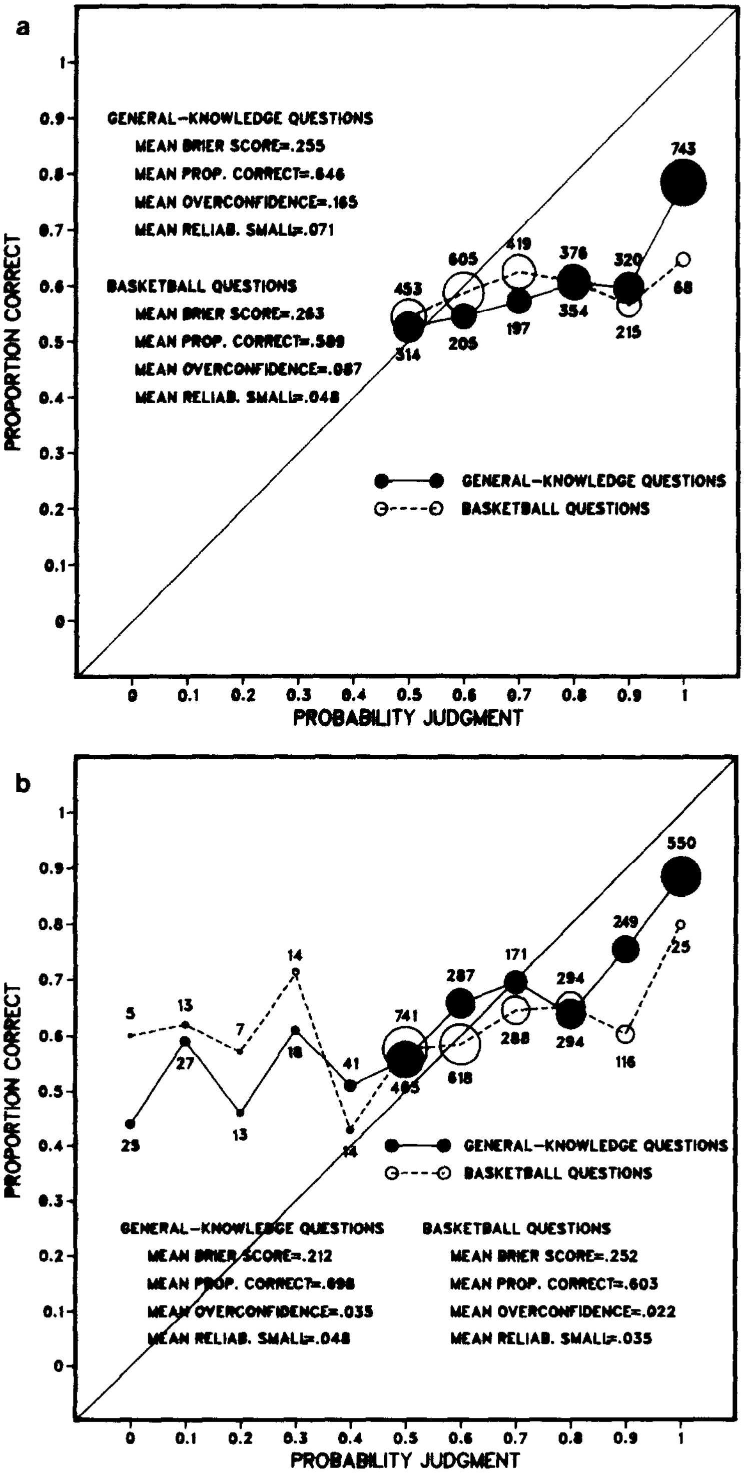
choices. As it turned out, implicit scoring decreased the proportion correct for the Choice-100 condition from .650 to .635. So most of the low probabilities are the result of misuse of the probability scale itself.

Examination of individual subjects' data revealed, however, that low probabilities had different meanings for different subjects. The subject with the best Brier score for general-knowledge questions, for example, assigned nine probabilities below .5. In each case, the circled answer was incorrect (see Figure 4A). Apparently, the subject was using the probability scale correctly, but preferred the answer he/she did not circle. In contrast, the subject with the worst Brier score for basketball questions assigned most probabilities below .5, but most of the circled answers were correct (see Fig. 4B). It seems likely that this subject was misusing the probability scale.

To see whether these probabilities below .5 were responsible for the low confidence, low overconfidence, low reliability-in-the-small, and high scatter for the Choice-100 methods, data from this method were reanalyzed after deleting subjects who gave probabilities below .5. It should be recognized that deleting a nonrandom sample of subjects from one condition while not deleting a comparable group from the other conditions may bias comparisons between conditions. Though we are aware of this potential problem, we have not been able to think of a better way to analyze these data to determine the effects of subjects who assigned probabilities below .5 when using the Choice-100 method. We, therefore, present the results in the following paragraph while urging caution in interpreting the comparisons between conditions that are described.

Deleting the subjects who assigned probabilities below .5 in the Choice-100 condition reduced scatter from .027 to .020 and eliminated the significant effect of method on scatter, indicating that the high scatter for the Choice-100 method was entirely due to subjects who gave probabilities less than .5. Deleting these subjects also increased mean confidence (from .680 to .694) and overconfidence (from .028 to .038). However, confidence and overconfidence were still noticeably lower for the Choice-100 method than for other methods, indicating that these effects were not completely due to failure to follow instructions and misunderstandings of the probability scale. Deleting these subjects actually reduced reliability-in-the-small (from .042 to .032). Thus, the probabilities

FIG. 2. Reliability diagrams for typical performances on general-knowledge and basketball questions, using the Choice-50 method. These reliability diagrams show the proportion correct for judgments made within levels of confidence. The number of judgments at each level of confidence is indicated by the size of the circle and by the associated number. This figure shows the same performances shown in Fig. 1.


a
GENERAL-KNOWLEDGE QUESTIONS
MEAN BRIER SCORE=.255
MEAN PROP. CORRECT=.646
MEAN OVERCONFIDENCE=.165
MEAN RELIAB. SMALL=.071
BASKETBALL QUESTIONS
MEAN BRIER SCORE=.263
MEAN PROP. CORRECT=.589
MEAN OVERCONFIDENCE=.087
MEAN RELIAB. SMALL=.048
743
453
605
419
376
320
68
314
205
197
354
215
GENERAL-KNOWLEDGE QUESTIONS
BASKETBALL QUESTIONS
PROPORTION CORRECT
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
PROBABILITY JUDGMENT
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
b
550
249
25
14
171
287
294
5
13
7
741
27
18
41
288
294
116
23
13
14
485
618
GENERAL-KNOWLEDGE QUESTIONS
BASKETBALL QUESTIONS
GENERAL-KNOWLEDGE QUESTIONS
MEAN BRIER SCORE=.212
MEAN PROP. CORRECT=.698
MEAN OVERCONFIDENCE=.035
MEAN RELIAB. SMALL=.048
BASKETBALL QUESTIONS
MEAN BRIER SCORE=.252
MEAN PROP. CORRECT=.603
MEAN OVERCONFIDENCE=.022
MEAN RELIAB. SMALL=.035
PROPORTION CORRECT
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
PROBABILITY JUDGMENT
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

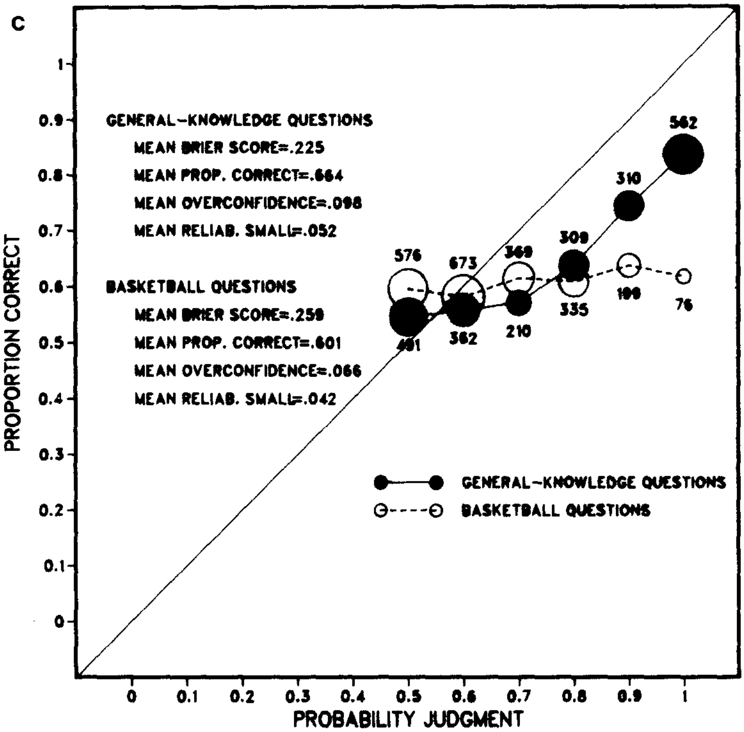


FIG. 3. Reliability diagrams for grouped data from the six combinations of topics (basketball and general-knowledge) and assessment methods used in the study. Data are collapsed over subjects and orders. The number of judgments at each level of confidence is indicated by the size of the circle or dot and by the associated number.

below .5 were not at all responsible for the good calibration for the Choice-100 method. Though the processes (other than failure to follow instructions and misuse of the probability scale) that produced the effects of the Choice-100 method are not known, it is notable that the proportion of .50 (complete uncertainty) responses was higher using the Choice-100 method than for the other methods (see Table 2).

Correlational Analyses

To test for consistency in probability judgment performance, Pearson correlations were computed between the accuracy scores for general-knowledge questions and for basketball predictions. To ensure that these correlations indicated individual consistency, rather than method effects, they were computed while partialing out the effects of response method. Modest, but significant, correlations were found between Brier scores $r(124) = .21$, $p < .05$, overconfidence, $r(124) = .25$, $p < .01$, scatter,

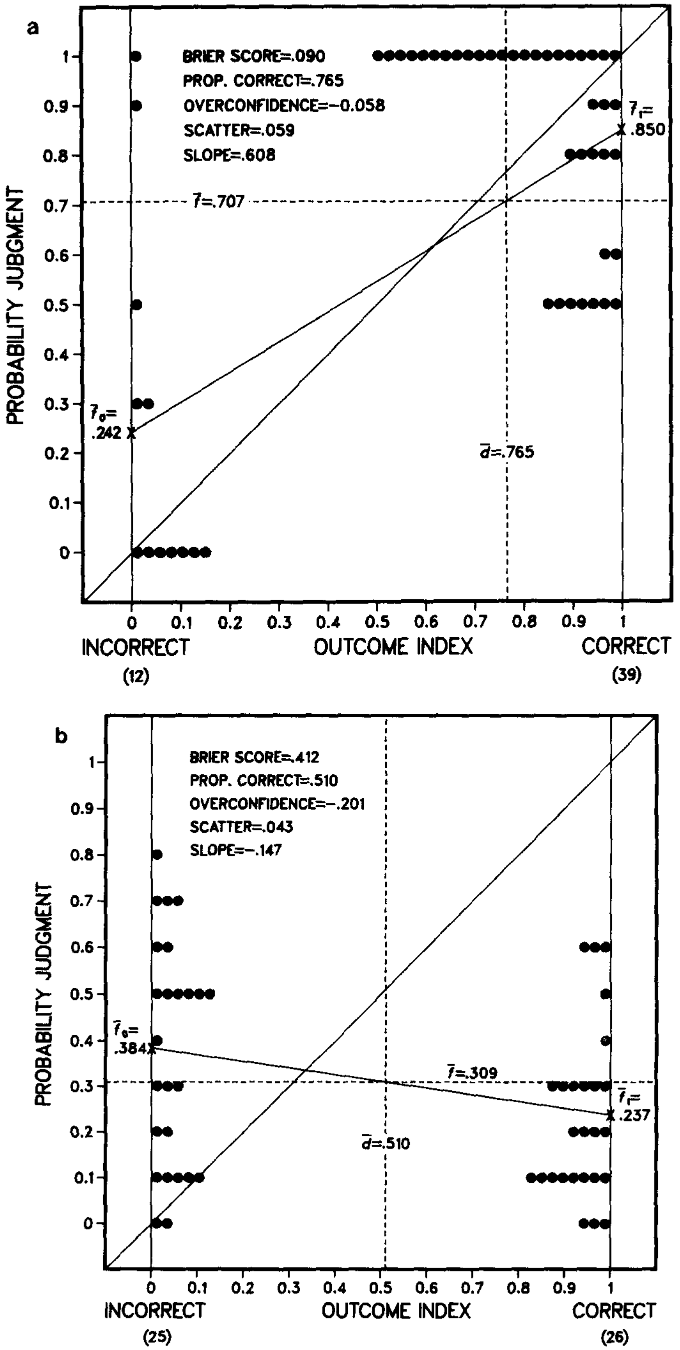


FIG. 4. Covariance graphs for two subjects who assigned probabilities below .50 in the Choice-100 condition.

$r(124) = .23, p < .01$, and reliability-in-the-small, $r(124) = .27, p < .01$ on the two topics. The strongest correlation between scores on the two topics was for mean confidence, $r(124) = .50, p < .001$. There were no reliable correlations between the proportions of correct answers chosen, $r(124) = .12, p > .05$ or between the slopes, $r(124) = .05, p > .05$, on the two topics. Thus, the only substantial consistency was in confidence levels.

DISCUSSION

In this experiment, subjects reported probability judgments for the outcomes of upcoming basketball games and for the answers to two-alternative, general-knowledge questions. Three methods were used for assessing subjects' judgments. Reliable and interpretable differences in various components of judgment accuracy were observed as a function of topic and method of probability assessment. The major findings of the study are briefly summarized and discussed. This is followed by a general discussion of the utility of measuring separable components of probability judgment accuracy.

Topic Effects

Subjects were more accurate in selecting answers to the general-knowledge questions than they were in predicting which basketball team would win each game. However, their judgments about general-knowledge questions were also more confident, more overconfident, more poorly calibrated, and showed more scatter. Subjects said they were absolutely sure about general-knowledge questions more than 15 times as often as they said they were sure about basketball predictions. These differences in confidence are too great to be explained simply as a result of the difference in question difficulty.

There appears to be an important difference between judgments about general-knowledge questions and predictions of basketball outcomes. Subjects were much more likely to recognize that they cannot know for sure what the outcome of a basketball game will be than they are to recognize the limitations of their general knowledge. The lower level of confidence and greater (appropriate) reluctance to express complete certainty about future events has been observed in several studies by other researchers (Fischhoff & MacGregor, 1982; Wright, 1982; Wright & Wisudha, 1982). One implication of this difference is that the results of studies using general-knowledge questions may not generalize to the real-world problem of forecasting future events.

Individual Consistency

Correlations between measures of performance on general-knowledge questions and those same measures applied to basketball predictions showed substantial consistency in mean confidence, but only weak consistency in other components of judgment quality. These findings are consonant with the results of Ramanaiah and Goldberg (1977), who found more individual consistency for confidence than for 11 other components of judgment. Clearly, skill in probability judgment is not a unitary trait. There are probably a multiplicity of skills and knowledge that allow a person to do a good job in making probability judgments. The skills and knowledge required apparently differ from one content domain to another. Like the main effects of topic discussed above, these findings raise the question of the generalizability of findings from one type of judgment to another.

Method Effects

No-Choice-100. In the No-Choice-100 method of measuring subjective probabilities, a randomly selected answer to each question was circled. Subjects were to judge the probability that the precircled answer was correct. This method produced the worst Brier scores, the worst calibration, the highest confidence, and the highest overconfidence. Subjects using this method assigned higher probabilities to the precircled answers than subjects using the other methods assigned to the same answers. Thus, the circling acted as a cue, even though subjects were informed that the decision about which answer was circled was made by flipping a coin.

The cuing effect was at best partially responsible for subjects' high confidence with the No-Choice-100 method. Subjects were most confident with this method even when they disagreed with the precircled answer. At this point we can only speculate about other processes leading to their high confidence. Perhaps seeing an answer circled focused subjects' attention on one alternative so they tended not to consider the pros and cons of the other answer. This might have led to more extreme probability assignments because the smaller sample of possible arguments would be more likely than a large sample to be one-sided (via the law of large numbers). It is plausible that the same mechanisms underlie both the present results and supra-additivity (Wright & Whalley, 1983). The latter is the phenomenon whereby the sum of subjects' probability judgments for all the events in a sample space partition tends to be greater—often much greater—than 100%.

Studies reported by Koriat, Lichtenstein, and Fischhoff (1980) and by Fischhoff and MacGregor (1982) demonstrated that inducing subjects to consider the pros and cons of both alternatives reduced confidence and overconfidence. Attending to arguments against one's preferred answer seemed particularly important. Similarly, studies of social judgment have demonstrated that focusing attention on one possibility, or explaining one potential alternative can increase the perceived likelihood of that possibility. Further, inducing people to consider the opposite possibility can reduce or eliminate the bias (Anderson & Sechler, 1986; Hirst & Sherman, 1985; Lord, Lepper, & Preston, 1984). Studies using more process-sensitive measures (e.g., reaction times, think-aloud protocols) would be needed to further clarify this issue.

The high confidence with the No-Choice-100 method is opposite to the hypothesis that choice would increase confidence. This hypothesis was inspired by self-perception theory (Bem, 1967), by early versions of dissonance theory (Brehm & Cohen, 1962; Festinger, 1957), and by an exploratory study reported by Fischhoff *et al.* (1977, Experiment 1). In the Fischhoff *et al.* study, more subjects were completely sure of their answers when they used a format analogous to our Choice-50 method (their Format 3), than when they used a procedure analogous to our No-Choice-100 method (their Format 4). The reason for the discrepant findings is not clear, but it is notable that the specific items differed somewhat from group to group in the Fischhoff *et al.* study, while the same items were used in all conditions in the current study. Since confidence varies greatly from item to item, this could explain the discrepancy. Though the current findings demonstrate that choice does not always increase confidence, they should not be taken as strong evidence against the hypothesis that (other things being equal) choice increases confidence. It could be that choosing an answer increased subjects' confidence, but that some other feature of the No-Choice-100 method (such as the focusing of attention) increased confidence even more.

Evaluating the No-Choice-100 method is easier than understanding all of its effects. This method leads subjects to be especially overconfident and biased toward the precircled answer. Clearly this is not a very good method of probability assessment for two-alternative general-knowledge questions like those in the current experiment. This negative evaluation of the No-Choice-100 method should not be overgeneralized to all assessment methods using a full (100-point) scale. Full-scale methods are very commonly used and seem quite natural when the same kind of event (e.g., precipitation) is being repeatedly predicted. The results do, however, raise the possibility that the Choice-50 method may be superior to full-scale methods even for repeated forecasts of the same kind of event. For instance, it is not out of the question that precipitation forecasts would be even better than they already are (Murphy & Brown, 1984; Murphy & Winkler, 1984) if meteorologists first indicated whether they thought "Precipitation" or "No precipitation" was more likely, then reported their confidence in the chosen alternative. Further research is required to assess the merits of full-scale and half-scale methods for such topics.

Choice-100. In this method the subject was instructed to circle his or her preferred answer and to assign a probability between 0.00 and 1.00 that the circled answer was correct. This method resulted in the lowest confidence and overconfidence, the best calibration (lowest reliability-in-the-small), and the most scatter. The high excess variance was entirely due to subjects who assigned probabilities below .50 to their preferred answers. The low confidence and overconfidence were only partially due to these subjects. The good calibration (low reliability-in-the-small) was not due to these subjects.

Most of the probabilities below .50 were reported by subjects who misused the probability scale. These subjects used probabilities less than .50 to indicate very low confidence. Some subjects, however, used probabilities below .50 to indicate that they preferred the answer they did not circle.

Although the reduction in overconfidence and improved calibration make the Choice-100 method somewhat attractive, these benefits are countered by the difficulty of interpreting probabilities below .50. In our opinion, this disadvantage is more important than the advantage. So it seems that the standard Choice-50 method is the most appropriate of the three methods tested in this study.

Scientific utility of measuring separate components of forecaster performance. Use of a variety of measures of different aspects of judgment quality proved to be helpful in finding and explaining substantively important differences in forecaster performance. The measures permitted the identification of various problems which differed in severity across topics and methods of probability assessment: (a) failure to select the correct answer (especially for basketball predictions); (b) failure to assign higher probabilities to correct than to incorrect answers (especially for basketball predictions); (c) poor calibration (especially for general-knowledge questions and the No-Choice-100 method); (d) overconfidence (especially for general-knowledge questions and the No-Choice-100 method); (e) high scatter (especially for general-knowledge questions and the Choice-100 method); (f) utilization of an irrelevant cue (the circled answers in the No-Choice-100 method); (g) failure to follow instructions (circling one's *less* preferred answer when using the Choice-100 method); and (h) misuse of the probability scale (i.e., assigning probabilities below .50 to indicate low confidence). These effects were generally interpretable and contributed to our understanding of probability judgment. In addition, examination of measures of several aspects of judgments revealed that the strongest individual consistency was in the overall level of confidence. Thus, looking at the components was scientifically useful. These findings suggest that it will be valuable for future research to use a broad selection of measures to facilitate understanding of influences on probability judgment.

Applications. Several recommendations can be drawn directly from the results of this study. First, since the No-Choice-100 method produced high overconfidence and invited subjects to use an irrelevant cue, its use for judgments of nonrepetitive events should be avoided. Second, since the subjects using the Choice-100 method often gave ambiguous responses (probabilities below .50 for their preferred answers), this method should also be avoided. Of the methods included in this experiment, the standard Choice-50 method seems most suitable. It should be remembered, however, that the experiment did not test the usefulness of full-scale methods for judgments of repetitive events, like precipitation. Such methods may be highly appropriate. Third, since the quality of a subject's basketball predictions was not highly correlated with the quality of his or her probability assessments for general-knowledge questions, one should not use a test of general-knowledge questions to select basketball oddsmakers. More generally, screening tests for judges should consist of questions for the same subject matter and (if possible) time frame as the forecasts that will be made on the job.

The success of using a wide selection of measures in identifying reliable and interpretable effects suggests that measurement of separate components may have many other applications as well. Since different types of training are needed to overcome different types of judgment errors, measures of separate components of probability judgment skill may be particularly useful for selecting appropriate forms of training. Judges who have difficulty choosing the most likely outcome can be helped by theoretical information about what cues are relevant and irrelevant, and practice in identifying and using those cues.

Judges who are overconfident, underconfident, or otherwise poorly calibrated need entirely different information. They can be helped with feedback about the proportion of their judgments (at each level of confidence) that turned out to be correct. For example, a judge could be informed that 78% of his or her judgments assigned 100% confidence turned out to be correct; that 72% of his or her judgments assigned 90% confidence turned out to be correct, etc. This would be expected to help judges bring their subjective probabilities into agreement with their hit rates (cf. Lichtenstein & Fischhoff, 1980). Incentives and direct suggestions to lower (or raise) their probabilities may also be useful (Fischer, 1982). Other kinds of errors in probability judgment (e.g., scatter) may require other kinds of training. The use of the measures of components of skill in probability judgment for the selection of training procedures seems very promising. Clearly there is a need for research on this possibility.

REFERENCES

Anderson, C. A., & Sechler, E. S. (1986). Effects of explanation and counterexplanation on the development and use of social theories. *Journal of Personality and Social Psychology*, **50**, 24–35.

Bem, D. J. (1967). Self-perception: An alternative explanation of cognitive dissonance phenomena. *Psychological Review,* 74, 183–200.

Brehm, J. W., & Cohen, A. R. (1962). *Explorations in cognitive dissonance.* New York: Wiley.

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

Festinger, L. (1957). *A theory of cognitive dissonance.* Stanford, CA: Stanford Univ. Press.

Fischer, G. W. (1982). Scoring-rule feedback and the overconfidence syndrome in subjective probability forecasting. *Organizational Behavior and Human Performance,* 29, 352–369.

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

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance,* 3, 552–564.

Hirt, E. R., & Sherman, S. J. (1985). The role of prior knowledge in explaining hypothetical events. *Journal of Experimental Social Psychology,* 21, 519–543.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory,* 6, 107–118.

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance,* 26, 149–171.

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.

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology,* 47, 1231–1243.

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

Murphy, A. H., & Brown, B. G. (1984). A comparative evaluation of objective and subjective weather forecasts in the United States. *Journal of Forecasting,* 3, 369–393.

Murphy, A. H., & Winkler, R. L. (1984). Probability forecasting in meteorology. *Journal of the American Statistical Association,* 79, 489–500.

Ramanaiah, N. V., & Goldberg, L. R. (1977). Stylistic components of human judgment: The generality of individual differences. *Applied Psychological Measurement,* 1, 23–39.

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

Wright, G. (1982). Changes in realism of probability assessments as a function of question type. *Acta Psychologica,* 52, 165–174.

Wright, G., & Whalley, P. (1983). The supra-additivity of subjective probability. In B. P. Stigum & F. Wenstop (Eds.), *Foundations of utility and risk theory with applications* (pp. 233–244). Dordrecht, Holland: Reidel.

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

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

RECEIVED: January 13, 1986