Bulletin of the Psychonomic Society
1987, 25(1), 61-64

Temporal aspects of probabilistic predictions

GIDEON KEREN
TNO Institute for Perception, Soesterberg, The Netherlands

and

WILLEM A. WAGENAAR
University of Leiden, Leiden, The Netherlands

Subjects who considered themselves to be experts made probability assessments regarding the outcomes of 108 soccer games. The mean number of correct predictions was less than that obtained by a linear model using type of game (home vs. outside games) and teams' rankings as predictors. From a normative viewpoint, other things being equal, the more remote (in the future) an event, the less certain one should be regarding the outcome. Subjects apparently fail to take into account this temporal variable.

In a recent article, Wagenaar and Keren (1986) distinguished between two types of probability assessments, *predictions* and *confidence ratings*. A prediction is always related to *future* events, such as the likelihood of rain tomorrow in the central part of The Netherlands or the probability of hitting a well in a certain oil field within the coming 3 months. Confidence ratings refer to the probability that an answer is correct and relate to one's knowledge and/or memory. A confidence rating, for instance on the question "was President Kennedy murdered in 1962 or in 1963?" is usually related to a *past* event.

The present study is concerned with predictions. From a normative viewpoint, the most important thing regarding predictions is to identify the main variables underlying the phenomenon of interest. As Dawes and Corrigan (1974) concluded, "The whole trick is to decide what variables to look at and then to know how to add" (p. 105). Probabilities attached to predictions reflect the limitations of our knowledge (even of experts' knowledge) regarding the variables underlying the particular phenomenon under interest. The source of the assessor's uncertainty in the case of confidence ratings is usually lack of personal knowledge or expertise.

How confident should one be in predicting future events? That would depend on the extent to which one believes that he or she identified the appropriate variables. In a dynamic world in which the role of different variables may constantly change (Hogarth, 1981), one should take into consideration an additional variable, namely the time factor. Other things being equal, the more remote (in time) a predicted event, the less confident one should be in the prediction of a particular outcome. For instance, consider predicting the Dow Jones Index a month from now and a year from now. Regardless of the variables used in making the predictions, the longer the lead time (defined here as the time between forecasting an outcome and its occurrence), the lower the confidence should be in the prediction. Despite the logical and intuitive appeal of such a guideline, it is questionable whether people follow it in practice. A major goal of the present paper was to test this question empirically.

The study involved the assessments of probabilities for the outcomes of professional soccer games in the national Dutch league. In addition to being an ideal topic for a probability assessment experiment (Winkler, 1971), it had the advantage that the temporal factor was naturally built-in and easy to investigate. An additional advantage is that a mechanical schema for prediction can be easily constructed. Two related questions may then be raised: First, how well do subjects' predictions compare with such mechanical predictions? Second, in what manner do subjects' predictions differ from those made by a mechanical schema? The answer to the latter question may provide insight into the processing underlying subjects' probabilistic predictions.

METHOD

Subjects

Subjects were 29 students from the State University of Leiden. They were recruited in a previous experiment from a population of 200 subjects. We asked the 200 subjects whether they were interested in soccer, and if so whether they would be interested in participating in an additional experiment. Those who responded positively were further asked to answer three factual questions concerning soccer in The Netherlands. Only those subjects who responded correctly to all three questions were included in the final sample (2 subjects were rejected because of one or more wrong answers).

Design and Procedure

Subjects were asked to predict the outcomes of 108 games in the top division of the Dutch soccer league. The games took place during a period of approximately 3 months. The 9 games planned for each weekend were typed on a separate sheet. For each game there were three possible out-

The authors would like to thank C. Lewis and J. Raaijmakers for their helpful comments. Requests for reprints should be addressed to Gideon Keren, TNO Institute for Perception, P.O. Box 23, 3769 ZG, Soesterberg, The Netherlands.

comes: win of home team, win of visiting team, and a tie. There were two different forms for making predictions. Fourteen subjects received Form A, in which they were asked, for each game, to pick the most likely outcome, and then assign it a probability (from 0.34 to 1.0) of being correct. The remaining 15 subjects were asked to assign, for each game, a probability for each of the three possible outcomes. They were explicitly told that the most likely outcome should be assigned the highest probability and that the sum of the three probabilities must be 1.0. Finally, on both forms, subjects were asked to make an overall estimate of the number of games they expected they would predict correctly (global assessments).

The forms were mailed to the subjects, who were asked to return them within a week. Approximately 2 months later, after 43 games had been played, an additional form was sent to all the subjects. They were given the results of 21 games already played and were asked to rate the extent to which they were surprised by each outcome. Each rating was made on a 10-point scale, 1 meaning *not surprised at all* and 10 meaning *extremely surprised.*

RESULTS

For all the analyses conducted, there was no significant difference in the pattern of responses to Form A or Form B. Consequently, all subjects were pooled into one sample, and the results reported below refer to this combined sample.

The mean number of correct predictions, over the 108 games, was 51.3 games (47.5%), which is significantly above chance level ($p < .001$). The range was between 42 and 62 correct predictions. The corresponding mean confidence rating was 59.1, suggesting that subjects were overconfident. This overconfidence was also reflected in the global assessments made by subjects. The mean global assessment (i.e., total number of correct predictions) was 60.8 (56.3%) games, which was significantly higher than the observed mean [$t(28) = 4.45$, $p < .001$].

The three possible outcomes of a game were not equally likely. Out of 108 games, 53 (49.1%) ended with a home team win, 26 (24.1%) ended with a visiting team win, and 29 (26.8%) resulted in a tie. The subjects in our sample were probably aware of these base rates and the relative frequencies of their predictions matched these base rates: 47.8% of all predictions were for a home team win, 26.6% were for a visiting team win, and 25.6% were for a tie.

The three different outcomes were not equally difficult to predict: Percentages of correct predictions were 58.1, 46.0, and 29.0 for home wins, visiting wins, and draws, respectively. The corresponding mean confidence ratings for these three predictions were 64.1, 60.1, and 48.8, respectively, suggesting that subjects took notice of the differential difficulty in predicting these three different outcomes, but did not sufficiently adjust for it.

To test subjects' sensitivity to the temporal variable, we divided the games into two groups: the 52 games played during the first 6 weeks were called *early games*, and those played during the remainder of the period were called *late games*. Calibration curves for early and late games are presented in Figure 1 and show the percentage of correct responses associated with the mean confidence for each level of confidence expressed by subjects

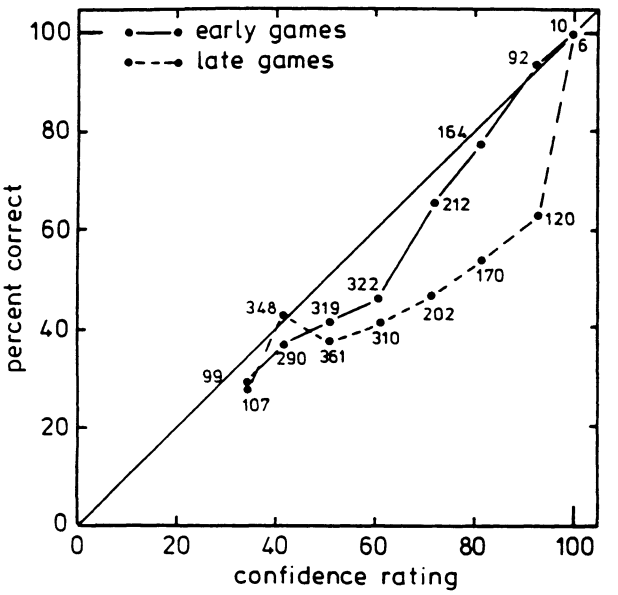

Figure 1. Calibration curves for early and late games. (Numbers next to each point indicate number of observations.)

(after collapsing those responses into 8 categories: 0.33–0.39, 0.40–0.49, 0.50–0.59, 0.60–0.69, 0.70–0.79, 0.80–0.89, 0.90–0.99, and 1.00). Although the curve for the early games shows a slight overconfidence, this is much more magnified with later games. The overall trend to be overconfident or underconfident can be measured by the mean signed difference between mean response and the corresponding proportion correct (e.g., Lichtenstein & Fischhoff, 1977). The corresponding values using this measure were 0.0768 and 0.1534 for early and late games, respectively. When this measure was computed for each subject separately, the difference between early and late games was highly significant [$t(28) = 6.43$, $p < .001$].

An alternative analysis to evaluate the two curves in Figure 1 is the calibration measure (e.g., Lichtenstein & Fischhoff, 1977; Lichtenstein, Fischhoff, & Phillips, 1982). It is the mean squared distance between each point on a curve and the identity line representing perfect calibration, weighted by the number of responses summarized in each point. Ideally it should be 0, a value indicating perfect calibration. Using this measure, the values for early and late games were 0.0080 and 0.0348, respectively. Computing it for each subject separately yielded a significantly larger value for late games [$t(28) = 3.51$, $p < .001$], suggesting poorer calibration in this condition. The calibration statistic was also computed separately for each of the three possible outcomes: the obtained values were 0.0036, 0.0159, and 0.1660 for home wins, visiting wins, and draws, respectively.

As expected, the surprise values highly reflected the percentage of correct predictions. There was a high negative correlation between mean surprise value for each game and the corresponding number of subjects who correctly predicted the outcome ($r = -0.93$). Table 1

Table 1
Mean Surprise Ratings as a Function of Mean Confidence Rating and Mean Percentage Correct

	Percent Correct	
Confidence Ratings	$x > 50\%$	$x < 50\%$
High ($x > 55$)	1.40 (n=8)	6.54 (n=4)
Low ($x < 55$)	3.13 (n=2)	4.21 (n=7)

Note—n = number of games.

presents the mean surprise ratings as a function of mean confidence ratings and mean percentage correct. The most surprising games are those with high confidence ratings but low accuracy. The least surprising games are those that received high confidence ratings and were correctly predicted.

DISCUSSION

It is clear that subjects did not simply guess (e.g., by throwing a mental coin), because the mean number of correct predictions was significantly above chance level. What information could subjects use to improve their predictions? There are at least two variables that can be used to enhance predictions above chance level. One concerns knowledge of base rates of home versus outside games; as was shown, subjects' predictions almost perfectly matched the base rates. However, simply knowing the base rates and employing them in a probability matching strategy would improve performance (above chance level) by less than 4%.[1]

The second relevant variable is the teams' rankings at the time predictions were made. Other things being equal, the team with a higher ranking should have a higher probability of winning against a team with a lower ranking. The larger the difference in ranking, the larger should be the difference in winning probabilities.

In order to obtain a baseline to which subjects' performance can be compared, we constructed the following linear model. Let P_W, P_L, and P_D stand for the probabilities of the *home team* to win, to lose, and to obtain a draw, respectively. Let R_H and R_V stand for the rank of the home team and visiting team, respectively. The probabilities of the home team to win or to lose may then be defined as

$$P_W = \alpha_W + \beta_W R_H + \gamma_W R_V \quad (1)$$

and

$$P_L = \alpha_L + \beta_L R_H + \gamma_L R_V. \quad (2)$$

The probability of a draw is then defined as a residual

$$P_D = 1 - P_W - P_L. \quad (3)$$

Suppose that for any given game we predict the outcome to be the one with the highest probability. To minimize the number of wrong predictions, we used the optimization procedure of STEPIT (Chandler, 1965; see also Hook & Jeeves, 1961). Employing this procedure, we obtained the following parameter values:

$$\alpha_W = 0.321, \alpha_L = 0.376,$$

$$\beta_W = -0.004, \beta_L = 0.011,$$

$$\gamma_W = 0.022, \gamma_L = -0.024.$$

Using these parameter values, the model correctly predicts 58 games (out of a total of 108). It is instructive to note that α_W and α_L, the two parameters associated with the base rates of home teams to win or lose, have by far the most important weight. In contrast, the four parameters associated with ranking (β_W, β_L, γ_W, γ_L) play a relatively small role. Indeed, with knowledge of the base rate of home wins, one optimal strategy is to guess that the home team will win every game. Such a strategy yields 53 correct predictions. Adding the ranking variable improves the mean accuracy of predictions only slightly.

How did subjects in our sample make their predictions? None of our subjects used the strategy of employing the base rate and constantly predicting a home win. This finding is consistent with the literature on probability learning (e.g., Estes, 1964) in which subjects rarely, if ever, use the optimal strategy of consistently predicting the event with the highest probability. Supposedly subjects felt they knew more, particularly about the ranking of the teams, and they wanted to incorporate this knowledge into their predictions. We do not know the exact strategies and computational processes employed by subjects in making their predictions; it is quite likely that subjects attempted to construct a prediction schema similar to the one we presented above. Some indirect evidence for this suggestion is provided by the surprise ratings: Among the 21 games that were rated for surprise outcomes, 11 games were correctly predicted by our model, and the mean surprise value attached to these games was 2.31; in contrast, the mean surprise value for the 10 games not predicted by the model was 4.77 [this difference was significant: $t(20) = 3.10$, $p < .005$]. In any event, subjects' performance was significantly lower than that obtained by our computational schema.

Apart from accuracy of predictions, a main interest of the present study was in probabilistic assessments. More specifically, we wanted to test to what extent judgments of likelihood are affected by the temporal setting of the events being judged. Previous research on judgments has typically ignored the temporal setting of stimuli used in experimental tasks. An exception is the study by Fischhoff (1976), who found no consistent differences in subjects' likelihood judgments regarding past and future events, which differed solely in their temporal setting. In that respect, the results obtained in our study are congruent with those reported by Fischhoff: We found complete insensitivity to the temporal factor. From a normative viewpoint, other things being equal, the farther an event is located in the future, the greater should be the uncertainty attached to that event. Certainty should be discounted by the size of the temporal factor. Our subjects did not follow this normative dictum and did not discount their probability assessments accordingly.

Despite their overconfidence, our subjects were aware that their prediction ability is limited, except that they did not think it was as poor as it was. A final question we posed to our subjects after they completed the prediction task was the following:

> As you know, the prediction of soccer games depends on several factors, particularly on skill, chance, and luck. We would like you to assess the relative importance of each of these three factors. Please give your answer in percentages, such that the sum of the three components will add up to 100.

The means were 44, 27, and 29 for skill, chance, and luck, respectively. Apparently, subjects admit that factors other than skill are predominant in their predictions, yet this admission was not sufficiently reflected in their corresponding probability assessments.

REFERENCES

Chandler, J. P. (1965). *Subroutine STEPIT* (Program OCPEGG, Quantum Chemistry Exchange). Bloomington: Indiana University.

Dawes, R., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, **81**, 95-106.

Estes, W. K. (1964). Probability learning. In A. W. Melton (Ed.), *Categories of human learning*. New York: Academic Press.

Fischhoff, B. (1976). The effect of temporal setting on likelihood estimates. *Organizational Behavior & Human Performance*, **15**, 180-194.

Hogarth, R. M. (1981). Beyond discrete biases: Functional and dysfunctional aspects of judgmental heuristics. *Psychological Bulletin*, **90**, 197-217.

Hook, R. T., & Jeeves, T. A. (1961). "Direct search"—Solution of numerical and statistical problems. *Journal of the Association for Computing Machinery*, **8**, 212-229.

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 & Human Performance*, **20**, 159-183.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibra-

tion 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 University Press.

WAGENAAR, W. A., & KEREN, G. (1986). Does the expert know? The reliability of predictions and confidence ratings of experts. In D. Hollnagel, G. Mancini, & D. D. Woods (Eds.), *Intelligent decision support in process environments* (pp. 87-106). Berlin: Springer-Verlag.

WINKLER, R. L. (1971). Probabilistic prediction: Some experimental results. *Journal of the American Statistical Association*, **66**, 675-685.

NOTE

1. Given the probabilities of 0.491, 0.241, and 0.268, respectively, for home wins, visit wins, and draws, a probability matching strategy would lead to an expected value of approximately 37% correct predictions ($0.491^2 + 0.241^2 + 0.268^2$), compared with 33% on a chance level.

(Manuscript received for publication August 22, 1986.)