Bulletin of the Psychonomic Society
1988, 26 (3), 236-237

Underconfidence in predicting future events

HENNIE VREUGDENHIL and PIETER KOELE
University of Amsterdam, Amsterdam, The Netherlands

In calibration research using future event questions, the mean of the confidence ratings assigned to each of a series of predictions is compared with the percentage of correct predictions. A person is said to be well calibrated if his/her mean confidence rating matches his/her percentage correct. In this study the temporal setting of future event questions was manipulated in order to investigate its effect on subject calibration. The subjects assigned confidence ratings to answers to questions concerning events located in the near future. Half of the questions covered a period of the next 4 days (early events); the other half covered a period of the next 14 days (late events). For both types of questions, underconfidence was found; the mean confidence rating was lower than the mean percentage correct. Although no difference in calibration was found between the two types of questions, the subjects were less confident about the late events than about the early events. These results are compared with those of some other studies in this area, and discrepancies and similarities are discussed.

In experiments that investigate the calibration of subjective probabilities, subjects usually have to perform a two-stage task. First, they select the most likely outcome of an event or the most plausible answer to a multiple-choice question. Next, they assign a subjective probability (also called confidence rating) to the selected outcome or answer. This task is repeated over a large number of trials. A subject is said to be well calibrated if, in the long run, his/her percentage of correct selections matches his/her mean confidence rating. If a subject's percentage correct is lower than his/her mean confidence rating, the subject is overconfident. If the reverse is found, the subject is underconfident.

In general, subjects are overconfident when answering questions on general knowledge topics, or when predicting the outcome of past events (Lichtenstein, Fischhoff, & Phillips, 1982). Wright (1982) reported underconfidence in subjects' predictions of the outcome of future events. Keren and Wagenaar (1987), who also used future event questions, reported overconfidence. Furthermore, Keren and Wagenaar found their subjects to be insensitive to the temporal aspects of the prediction of future events. Their subjects predicted the outcomes of soccer games in the Dutch league, to be played within 12 weeks. They assigned confidence ratings to their predictions about the home teams' winning, losing, or obtaining a draw. Games to be played during the first 6 weeks were called early games, and those to be played during the last 6 weeks were called late games. Keren and Wagenaar found that the confidence ratings assigned to the outcomes of the early games were as high as those assigned to the outcomes of the late games. For both early and late games, the percentage of correct predictions was lower than the mean confidence rating. This overconfidence proved to be strongest for the late games, because for these games the percentage of correct predictions was lower.

Because of the interesting discrepancy between the results of Keren and Wagenaar's (1987) and Wright's (1982) studies, the present study was undertaken. Its purpose was to gain more insight into calibration on future event questions. Two types of questions were used: those covering a short and those covering a longer period of time.

METHOD

Subjects

The subjects were 60 first-year students majoring in psychology at the University of Amsterdam. They participated in the experiment as part of a course requirement.

Materials

The test questionnaire contained 20 two-choice items. There were two different forms. In the early events form, the questions covered a period of 4 days [e.g., "Will Rudolf Hess die in Spandau prison within the next 4 days? (a) yes; (b) no"], and in the late events form, the questions covered a fortnight [e.g., "Will it snow in Amsterdam during the next 14 days? (a) yes; (b) no"].

Design and Procedure

The subjects were requested to select the most likely answer and subsequently assign it a probability between 50% and 100%. Half of the subjects answered the early events form; the other half answered the late events form. Both groups consisted of 7 male and 23 female students.

RESULTS

The mean percentage of correct predictions was 83% for the early events and 79% for the late events, both significantly above chance level, but not significantly different from each other [$t(58) = 1.23, p = .22$]. Evidently, the two forms did not differ in difficulty; in both cases, subjects selected the correct answer on roughly 80% of the items. The mean confidence rating was 79% for the early events and 72% for the late events. This difference

The authors are indebted to Gideon Keren for his helpful comments.
Address correspondence to Pieter Koele, Methodology Unit, Faculty of Psychology, University of Amsterdam, 1018 XA Amsterdam, The Netherlands.

is significant [$t(58) = 3.02$, $p = .001$], indicating that the subjects were less confident about their answers to the late event questions than about those to the early event questions.

Calibration curves for both forms are presented in Figure 1, which shows for each confidence category (50-59, 60-69, 70-79, 80-89, 90-99, 100) the mean confidence rating and the corresponding mean percentage correct. Both curves show underconfidence.

Two calibration measures were used. Yates (1982) introduced the distinction between *calibration in the large* and *calibration in the small*. Calibration in the small takes into account the difference between confidence ratings and percentages correct within each confidence category, whereas calibration in the large compares the overall confidence rating with the overall percentage correct. The calibration-in-the-large measure is the signed difference between the mean confidence rating and percentage correct (Lichtenstein & Fischhoff, 1977), calculated for each subject separately. The calibration-in-the-small measure (also calculated for each subject separately) is the weighted average of the mean squared differences between the proportion correct in each confidence category and the mean confidence rating in that category. In the case of perfect calibration, both measures have the value 0.

The mean value of the in-the-large measure is −5.93 for the early events and −9.06 for the late events. Using the t test for independent samples, this difference does not prove to be significant [$t(58) = 1.50$, $p = 0.14$]. The mean values of the calibration-in-the-small measure for the early and late events are 0.008 and 0.011, respectively. This difference is not significant [$t(58) = 1.19$, $p = 0.24$].

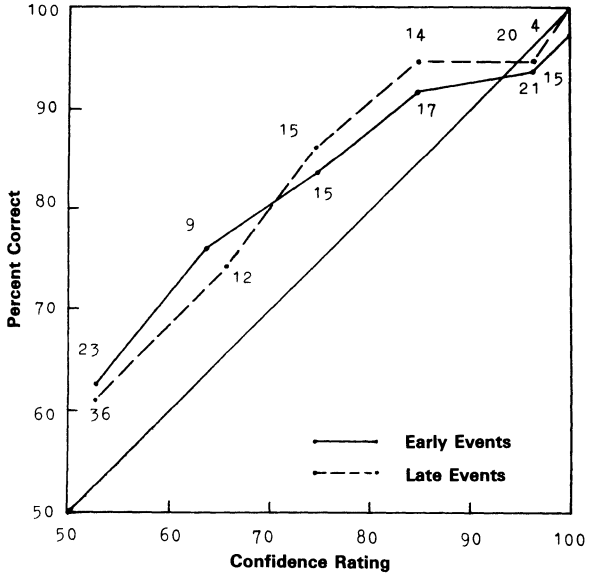

Figure 1. Calibration curves for the early and late events. (Numbers next to each point indicate percentage of observations.)

Both calibration measures indicate that the subjects were slightly miscalibrated on both forms. The calibration curves and the in-the-large measure show that this miscalibration was due to underconfidence. The only difference between the two forms lies in the mean confidence rating: subjects were less confident about the outcomes of late events than about the outcomes of early events.

DISCUSSION

Our subjects were slightly underconfident when answering future event questions. Additionally, they were less confident about late events than about early events. These results contradict the conclusion drawn by Keren and Wagenaar (1987) but are in accordance with that of Wright (1982). The subjects were not unaware of the temporal setting of questions, and they were capable of translating this awareness into lower confidence ratings for predictions covering longer periods of time.

But how could Keren and Wagenaar (1987) find overconfidence? First, their subjects were a nonrandom group, selected on the basis of their self-proclaimed knowledge of Dutch soccer. It is possible that, in order to justify their selection, the subjects felt an obligation to display an unwarranted confidence in their predictions (see also Bradley, 1981). Second, because of their knowledge, Keren and Wagenaar's subjects may have treated the questions as a kind of ability test, ignoring the time factor altogether. In our experiment, we used questions concerning catastrophes, politics, and weather. These questions are explicitly related to the future, ensuring that subjects cannot fail to acknowledge the time aspect.

These two points sufficiently explain Keren and Wagenaar's (1987) results. They may also explain the frequently found overconfidence exhibited by experts, such as physicians (Christensen-Szalanski & Busyhead, 1981). When asked to express their certainty concerning a diagnosis, physicians had unjustified confidence in their judgments. Whether this confidence was intended to reassure their clients or themselves is a question deserving further attention.

REFERENCES

BRADLEY, J. V. (1981). Overconfidence in ignorant experts. *Bulletin of the Psychonomic Society*, **17**, 82-84.

CHRISTENSEN-SZALANSKI, J., & BUSYHEAD, J. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception & Performance*, **7**, 928-935.

KEREN, G., & WAGENAAR, W. A. (1987). Temporal aspects of probabilistic predictions. *Bulletin of the Psychonomic Society*, **25**, 61-64.

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). 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 University Press.

WRIGHT, G. (1982). Changes in the realism and the distribution of probability assessments as a function of question type. *Acta Psychologica*, **52**, 165-174.

YATES, F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior & Human Performance*, **30**, 132-156.

(Manuscript received for publication September 11, 1987.)